ASSESSING EARLY WORD LEARNING AND EARLY WORD KNOWLEDGE WITH TABLETS: A WEB TECHNOLOGY–BASED APPROACH

Chang Huan Lo

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy in Psychology

July 2021

Dedicated to the ones who love me and the ones whom I love.

ABSTRACT

In light of the proliferation of tablets (and apps) in young children's lives, the overarching theme of this thesis is to examine ways in which the unique affordances of such devices can contribute to young children's early language development. More specifically, this thesis takes a detailed look at young children's word learning from tablets and the potential use of tablets as a means to assess early word knowledge.

From the word learning viewpoint, the first three studies, including a pilot study, examined 2- to 3-year-olds' word learning from a tablet app through two learning modes: active selection versus passive reception. Results from Study 1A suggest a passive advantage in terms of recognition accuracy among 30- and 40-month-olds but no such advantage was found among 24-month-olds. That is, giving children active control over their learning experiences did not appear to benefit children across the three age groups, but passive watching led to better performance among older children. While Study 1B replicated these results with a new group of 30-month-olds from a different cultural and linguistic background, no differences were found across both active and passive conditions using a more implicit looking time measure, suggesting that children learnt equally across both conditions, but there may be performance costs associated with active selection in tasks designed as in these studies.

From the word knowledge assessment viewpoint, Study 2 explored the viability of tablets in assessing early word comprehension among 1-year-olds by means of a two-alternative forced choice word recognition task. Preliminary results indicated that children as young as 18 months can engage meaningfully with a tablet-based assessment, with minimal verbal instruction and child–administrator interaction. The encouraging results further suggest that such assessments have scope for deriving a direct measure of early word comprehension that can supplement parent reports, such as the

MacArthur–Bates Communicative Development Inventories (CDI), thereby addressing concerns relating to the exclusive use of parent reports and allowing a more complete picture of children's early language development. In order to facilitate the assessment of early word knowledge, Study 3 sought to develop a language-general approach that produces adaptive short-form versions of CDIs with test items that are maximally informative and derives estimates of full CDI scores based on prior CDI data from language-, sex-, and age-matched children. Results from real-data simulations revealed that the approach was able to efficiently estimate full CDI scores with tests featuring fewer than 25 items—regardless of language, sex, and age—achieving correlations above .95 with full CDI administrations, with high levels of reliability.

Through the combination of web technology and tablets, this thesis also showcases the potential and value of web- and tablet-based methods for collecting data in early developmental research. To make web methods more accessible to researchers, this thesis additionally contributes a new authoring tool, e-Babylab, that allows users to create, host, run, and manage browser-based experiments—without the need for prior technical knowledge. Implications of the results and research limitations, along with possible avenues for future research are discussed.

PUBLISHED AND SUBMITTED PAPERS

This thesis incorporates material from the following papers:

(A) Ackermann, L.[1], **Lo, C. H.**[1], Mani, N., & Mayor, J. (2020). Word learning from a tablet app: Toddlers perform better in a passive context. *PLoS ONE*, *15*(12), e0240519. https://doi.org/10.1371/journal.pone.0240519

(B) Chai, J. H.[1], **Lo, C. H.**[1], & Mayor, J. (2020). A Bayesian-inspired item response theory–based framework to produce very short versions of MacArthur–Bates Communicative Development Inventories. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3488–3500. https://doi.org/10.1044/2020_JSLHR-20-00361[2]

(C) **Lo, C. H.**[1], Rosslund, A.[1], Chai, J. H., Mayor, J., & Kartushina, N. (2021). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy*. Advance online publication. https://doi.org/10.1111/infa.12401

(D) **Lo, C. H.**, Mani, N., Kartushina, N., Mayor, J., & Hermes, J. (2021). *e-Babylab: An open-source browser-based tool for unmoderated online developmental studies.* Manuscript submitted for publication.[3]

[1]Both authors share co-first authorship.
[2]Permission to reprint has been granted by American Speech-Language-Hearing Association.
[3]The preprint is available at https://doi.org/10.31234/osf.io/u73sy.

AUTHOR CONTRIBUTIONS

Author contributions for each of the published and submitted papers are as follows:

(A) CHL programmed the web applications for all studies, conducted the study in Malaysia, coded the gaze data, conducted formal analysis, interpreted the results, wrote the original draft, and revised the manuscript. LA conducted the studies in Germany, conducted formal analysis, interpreted the results, wrote the original draft, and revised the manuscript. NM conceptualised and supervised the study, and revised the manuscript. JM conceptualised and supervised the study, and revised the manuscript.

(B) CHL wrote the R script, conducted real-data simulations, interpreted the results, wrote the original draft, and revised the manuscript. JHC wrote the R script, conducted real-data simulations, interpreted the results, wrote the original draft, and revised the manuscript. JM conceptualised and supervised the study, and revised the manuscript.

(C) CHL programmed the web application, conducted formal analysis, interpreted the results, and wrote the original draft. AR wrote the original draft and revised the manuscript. JHC conducted formal analysis. JM conceptualised and supervised the study, and revised the manuscript. NK conceptualised, supervised, and conducted the study, and revised the manuscript.

(D) CHL programmed e-Babylab, set up e-Babylab on university servers, wrote the user manual, wrote the original draft, and revised the manuscript. NM conceptualised e-Babylab, supervised the studies, conducted formal analysis, interpreted the results, wrote the original draft, and revised the manuscript. NK supervised and conducted the studies, conducted formal analysis, interpreted the results, wrote the original draft, and revised the

manuscript. JM conceptualised e-Babylab, supervised and conducted the studies, conducted formal analysis, interpreted the results, wrote the original draft, and revised the manuscript. JH conceptualised e-Babylab, supervised and conducted the studies, wrote the original draft, and revised the manuscript.

ACKNOWLEDGMENTS

This thesis owes much to the help and support from many colleagues, friends, and family. First and foremost, I am grateful to my supervisors, Dr. Julien Mayor and Dr. Steve Stewart-Williams, without whom this thesis would not have been possible. Thank you for your patient support and guidance as this thesis takes shape. This work was also supported by the Fundamental Research Grant Scheme grant (FRGS-NFHY0004) awarded to Dr. Julien Mayor.

Next, I wish to gratefully acknowledge the childcare centres, kindergartens, parents, and children who participated in the studies, not to mention the labs—WortSchatzInsel, Sunway BabyLab, and BabyLing, which facilitated the participant recruitment process and data collection.

Finally, I wish to thank my family and friends. To my friends, old and new, I truly appreciate our friendship and will always treasure the times we spent together learning, travelling, and having lunches and coffee breaks. To dad and mum, thank you for your endless love, support, and encouragement through the years and all that you have done for Xin and me. Last but not least, to Mark—danke, dass du immer für mich da bist.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

2AFC	Two-alternative forced choice
4AFC	Four-alternative forced choice
AJAX	Asynchronous JavaScript and XML
API	Application programming interface
ASD	Autism spectrum disorder
CAT	Computerised adaptive testing
CCT	Computerized Comprehension Task
CDI	MacArthur–Bates Communicative Development Inventories
CDI–WG	CDI–Words and Gestures
CDI–WS	CDI–Words and Sentences
COVID-19	Coronavirus disease 2019
CSS	Cascading Style Sheets
EVT	Expressive Vocabulary Test
GLMM	Generalised linear mixed-effects model
GUI	Graphical user interface
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
ID	Identifier
IPLP	Intermodal Preferential Looking Paradigm
IRT	Item response theory
JSON	JavaScript Object Notation
LMM	Linear mixed-effects model
LWL	Looking-while-listening
MAD	Median absolute deviation
MTurk	Amazon Mechanical Turk
PPVT	Peabody Picture Vocabulary Test

RT	Reaction time
SES	Socioeconomic status
SQL	Structured Query Language
SSL	Secure Sockets Layer
TLS	Transport Layer Security
URL	Uniform Resource Locator
UUID	Universally unique identifier
WAIS	Wechsler Adult Intelligence Scale
WMS	Wechsler Memory Scale
XML	Extensible Markup Language

CHAPTER 1. INTRODUCTION

1.1 Focus of the Research

This thesis examines young children's word learning from tablets as well as the use of tablets in assessing early word knowledge. For this purpose, e-Babylab, an authoring tool for creating browser-based experiments was developed. Study 1 looked into young children's word learning from tablet applications ("apps") through two learning modes: active selection versus passive reception. Study 2 explored the viability of a tablet-based word recognition task in assessing young children's word knowledge. In order to facilitate early word knowledge assessments, this thesis also seeks to further develop short-form versions of the MacArthur–Bates Communicative Development Inventories (CDI)—without compromising on the accuracy and precision of the full forms. Thus, in Study 3, a language-general approach that produces adaptive short-form versions of CDIs with test items that are maximally informative and derives estimates of full CDI scores based on prior CDI data from language-, sex-, and age-matched children is presented and validated against established short forms through real-data simulations.

1.2 Background of the Research

Since the debut of Apple's iPad in 2010, iPads and similar tablet devices have become ubiquitous. Ownership of such highly intuitive touchscreen devices among American families with children aged 0 to 8 years saw almost a ten-fold increase within just a few years, from 8% in 2011 to 78% in 2017 (Rideout, 2017). The increasing prevalence of tablets is also evident in British households as 89% of families with children aged 5 to 15 years reported owning a tablet in 2019, up from 5% in 2010 (Ofcom, 2012, 2020). In 2013, more than half (51%) of

British families with younger children (aged 3 to 4 years) had a tablet and six years later, this figure rose to 85% (Ofcom, 2013, 2020).

The same reports also highlighted an equally astounding increase in child tablet ownership. In 2017, 42% of American children aged 0 to 8 years owned a tablet, compared to 2011 when less than 1% of them did (Rideout, 2017). In 2019, nearly half (49%) of all British children aged 5 to 15 years owned a tablet; in 2011, only 2% did (Ofcom, 2014, 2020). Among younger British children (aged 3 to 4 years), approximately one in every four (24%) owned a tablet in 2019, an eight-fold increase since 2013 (3%; Ofcom, 2013, 2020).

Accompanied by this expanded access to tablets is the increased use among children. In 2011, only 38% of American children aged 0 to 8 years had ever used a mobile device (e.g., smartphones and tablets) and they spent on average five minutes a day consuming mobile media; in 2017, 84% had done so and average daily usage had risen by almost 10 times, to 48 minutes (Rideout, 2017). Among British children aged 3 to 4 years, the number of children who had used a tablet had more than doubled within six years, from 28% in 2013 to 64% in 2019 (Ofcom, 2013, 2020). The increase was even more substantial for older children (aged 5 to 15 years) as this figure increased from 3% in 2010 to 80% in 2019 (Ofcom, 2012, 2020). Another survey involving 2,000 British parents of children aged 0 to 5 years revealed that children spent an average of 79 minutes daily using tablets (Marsh et al., 2015).

The increasing popularity of tablets is driven by the broad content offered via apps, not only to children, but also to their parents. To date, the Apple App Store features over 200,000 apps for education (Apple Inc., 2019) and a significant proportion of apps put under the "Education" category—either available for free or for a fee—is targeted at children, with "toddlers or preschoolers" being the most popular age category (Shuler, 2012). Despite the educational claims that these apps make, they are mostly released without prior formal evaluation (Hirsh-Pasek et al., 2015) and only few apps aimed at preschoolers provide developmentally appropriate guidance and feedback (Callaghan & Reich, 2018). Contrary to the recommendation from the American

Academy of Pediatrics to only let young children consume high-quality media (i.e., age-appropriate programs or apps containing educational value; American Academy of Pediatrics, 2016), parents seem to be buying into the great promise of these apps as 80% of them reported that they have downloaded apps for their children aged 2 to 4 years (Rideout, 2017). It remains questionable whether children, at such a young age, are capable of learning from touchscreen devices since literature on young children's learning from traditional screen media (e.g., television) has consistently found that they learn better from in-person experiences rather than from on-screen experiences (R. Barr, 2010; Troseth, 2010).

In addition, the intuitive touchscreen interface, a critical feature which makes tablets so easy to use, also adds to their appeal, especially among young children. In Marsh et al. (2015), more than half (54%) of the children aged 0 to 2 years could swipe the screen unassisted by an adult (e.g., to turn the pages of electronic books), while three-quarters of those aged 3 to 5 years were able to swipe the screen (76%), open apps (75%), and trace shapes with their fingers on the screen (75%). Abdul Aziz et al. (2014) found that out of the seven gestures typically found in iPad apps designed for children (i.e., tap, drag/slide, drag and drop, pinch, flick, spread, and free rotate), children as young as age 2 had already mastered the first two gestures and by age 3, they could perform all but the spread gesture. This presents an exciting opportunity for data collection among young children, especially in assessing early word knowledge, which is typically done via parent reports (e.g., CDIs). Coupled with recent efforts in the development of short-form versions of parent reports, the use of tablet-based tasks may open up new possibilities for directly assessing young children in an engaging and interactive manner. The mobility of tablets also means that experiments do not necessarily need to be conducted in the laboratory but can instead be conducted, for instance, at kindergartens or at children's homes, thus enabling children to be tested in their natural environment.

Considering how tablets and apps are becoming increasingly common among young children, research on the ways in which the unique affordances of

tablets and apps can contribute to young children's early language development is thus particularly relevant to parents and their children, educators, researchers working with young children, as well as app developers/publishers.

1.3 Overview of Chapters

In the next chapter, a critical analysis of existing literature on young children's learning from screens is presented, highlighting the need for research on the educational potential of tablets during early childhood. By analysing the literature on the different measures in assessing early word knowledge, including parent reports and direct language measures, the potential of tablets to be used as a means to assess early word knowledge is also considered. Finally, this chapter looks into web technology–based experimentation as a methodology for data acquisition.

Chapter 3 presents the features and technical details of e-Babylab, a new authoring tool developed as a part of this thesis, to allow users to create, host, run, and manage browser-based experiments for online testing—without the need for prior technical knowledge.

In Chapter 4, a series of studies designed to assess young children's word learning with tablets is reported. The studies include a pilot study, evaluating the feasibility of the study design, followed by two studies, conducted among young German- and Malay-speaking children respectively.

Two studies relevant to early word knowledge assessment are reported in Chapter 5. The first examines the viability of tablets in assessing early word knowledge by means of a word recognition task, while the second presents a language-general approach to producing short-form versions of CDIs and validates the approach through real-data simulations.

Finally, Chapter 6 provides an overview of the findings from both word learning– and word knowledge assessment–related studies. The implications of the findings and research limitations are also discussed, along with possible avenues for future research.

CHAPTER 2. LITERATURE REVIEW

This chapter reviews the literature relevant to three central topics: young children's learning from screens, early word knowledge assessment, and data acquisition with web technology. First, the "video deficit effect" pertaining to young children's reduced ability in learning from screens is outlined. The effects of contingency on the video deficit as well as the effects of self-direction on young children's learning are then considered, highlighting the need for gaps in the current evidence base to be bridged. Next, two general types of methods appropriate for assessing young children's early word knowledge, namely parent report and direct language measure, are discussed while providing an overview of their strengths and limitations. The potential of tablets as a data collection modality, followed by the literature on web-based methods are then considered to provide a rationale for the methodology used in the present research. This chapter incorporates material from the following papers:

> Ackermann, L.[4], **Lo, C. H.**[4], Mani, N., & Mayor, J. (2020). Word learning from a tablet app: Toddlers perform better in a passive context. *PLoS ONE*, *15*(12), e0240519. https://doi.org/10.1371/journal.pone.0240519
>
> Chai, J. H.[4], **Lo, C. H.**[4], & Mayor, J. (2020). A Bayesian-inspired item response theory–based framework to produce very short versions of MacArthur–Bates Communicative Development Inventories. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3488–3500. https://doi.org/10.1044/2020_JSLHR-20-00361[5]

[4]Both authors share co-first authorship.

[5]Permission to reprint has been granted by American Speech-Language-Hearing Association.

2.1 Young Children's Learning From Screens

2.1.1 Video Deficit Effect

The starting point for research on young children's learning from screens is the suggestion that children exhibit little learning from passive video viewing and benefit more from equivalent live experiences, an effect referred to as the "video deficit effect" (D. R. Anderson & Pempek, 2005). This effect is not task-specific and has been exhibited in various domains, including (but not limited to) action imitation (R. Barr & Hayne, 1999; R. Barr et al., 2007; Deocampo & Hudson, 2005; Dickerson et al., 2013; Hayne et al., 2003; Hudson & Sheffield, 1999; Strouse & Troseth, 2008), object retrieval (Schmitt & Anderson, 2002; Troseth & DeLoache, 1998), emotion processing (Diener et al., 2008; Mumme & Fernald, 2003), self-recognition (Suddendorf et al., 2007), and language learning (Krcmar et al., 2007; Roseberry et al., 2009; Troseth et al., 2018). In general, the literature suggests that the effect peaks around 15 to 24 months of age and then diminishes until approximately 36 months (R. Barr, 2010; DeLoache et al., 2010; Dickerson et al., 2013; Troseth, 2010), although, depending on task difficulty and measure sensitivity, the effect may persist beyond 36 months (Flynn & Whiten, 2008; McGuigan et al., 2007; Reiß et al., 2019; Roseberry et al., 2009; Strouse & Samson, 2021).

To account for the video deficit effect, researchers have put forward several non-mutually exclusive hypotheses. According to the *dual representation* hypothesis, the video deficit effect stems from infants and toddlers' immature pictorial competence, or in other words, their poor understanding of the dual nature of symbolic artefacts (DeLoache, 1987, 1991; DeLoache et al., 2003; Troseth et al., 2019; Troseth et al., 2004). Specifically, young children may not be able to grasp the fact that a symbolic object, such as a television, is in itself an object and at the same time representational of another object that it depicts (R. Barr & Hayne, 1999; Troseth, 2010; Troseth et al., 2004). Thus, young children fail to relate, and hence apply information communicated through the symbolic object to the real world.

Another hypothesis focuses on the nature of 2D inputs being *perceptually impoverished* relative to 3D inputs. That is, 2D inputs lack perceptual cues, such as motion parallax, depth perception, and texture. And because fewer details are encoded from 2D inputs, there is a higher chance of a mismatch of cues and/or specific cues needed may be missing at the time of retrieval; consequently, retrieval is impaired (R. Barr & Hayne, 1999; R. Barr et al., 2007; Schmitt & Anderson, 2002; Suddendorf, 2003). In addition, the processing of such perceptually degraded information may consume more cognitive resources and require more working memory (R. Barr et al., 2016; Choi et al., 2018; Kirkorian, Lavigne, et al., 2016). In contrast, live demonstrations which are abundant in perceptual information lead to better encoding (i.e., more detailed memories), since less cognitive resources are needed in information processing. In support of the view that young children process 2D and 3D inputs differently, studies using event-related potentials (Carver et al., 2006) and eye-tracking data (Kirkorian, Lavigne, et al., 2016) have respectively found that 18- and 24-month-olds take a longer amount of time to process 2D images than 3D objects.

While limitations in both perceptual and symbolic processing account for young children's reduced ability in learning from screen media relative to live demonstrations, neither fully accounts for all the current findings (R. Barr, 2008). For instance, M. E. Schmidt et al. (2007) found that 2-year-olds continued to perform poorly in the video condition in an object retrieval task, even when the perceptual problem related to the mapping from 2D to 3D was eliminated (i.e., by using a 2D search space of the same size as the screen on which the information was presented). In their second experiment, the need for dual representation was further removed by having the experimenter tell children explicitly where the toy was hidden, either in-person or through closed-circuit video. Yet, children still performed worse in the video condition than in the unmediated condition, suggesting that the video deficit did not result solely from perceptual or dual representation issues and that other factors could also be at play.

An additional account for the video deficit effect concerns the fact that screen media are *socially impoverished* relative to in-person experiences. Specifically, screen media lack many social cues (e.g., contingent responses, eye gaze, and name referral) which young children readily use to guide their learning about the world in social situations (see Baldwin, 2000; Baldwin & Moses, 2001, for reviews). For instance, Baldwin et al. (1996) found that infants aged 18 to 20 months rely on referential social cues (e.g., eye gaze, body posture) to direct the establishment of new word–referent associations and resist establishing associations when such cues are missing. Relatedly, due to the lack of social cues in (non-contingent) screen media as well as young children's limited experience with live video (where there is a two-way exchange of information), young children may discount televised information as irrelevant to reality (Jing & Kirkorian, 2020; Troseth, 2010; Troseth & DeLoache, 1998), and subsequently fail to treat screen models as someone who provides meaningful information about the real world (Strouse et al., 2018). Indeed, young children are more likely to succeed in their use of televised information after having experienced live video where the screen model provides socially relevant information (e.g., referring to the child's name) or responds contingently to the child's behaviour (Myers et al., 2017; Nielsen et al., 2008; Roseberry et al., 2014; Troseth, 2003; Troseth et al., 2006), although several studies using more challenging tasks have found null effects and that the presence of a co-viewer best supports young children's learning from screen media (Strouse et al., 2018; Troseth et al., 2018).

In sum, these findings suggest that the video deficit effect results from several converging factors and can be attenuated by providing young children with some form of social support (i.e., social contingency) to help them link video experiences with reality and to meet the extra cognitive or working memory demands for reconciling mismatches between video and real-world contexts.

2.1.2 Pseudo-Social Contingency

Since young children require scaffolding in learning from screens (as noted in the previous section), researchers have speculated about whether

pseudo-social contingency, such as on-screen interactive features, can support or hinder learning. On the one hand, pseudo-social contingency may pose an impediment for dual representation because the possibility to manipulate what is displayed on-screen may lead children to treat the screen as an object in itself rather than as a symbolic medium that can represent another object (Sheehan & Uttal, 2016). On the other hand, the contingent responsiveness may serve to promote engagement or direct children's attention towards relevant information presented on-screen, thereby supporting learning (Kirkorian, 2018).

Through an object retrieval task, Lauricella et al. (2010) investigated whether pseudo-social contingent computer interactions (i.e., where children interacted with a game and could steer the course of the actions presented in the game by providing user input, such as pressing particular buttons) would mitigate the video deficit for 30- and 36-month-olds. It was found that even without social interactions, the interactive computer game providing contingent responses to children's keyboard presses facilitated their learning. In fact, their performance was similar to those who were given a live demonstration and significantly better than those who only observed the game being played on the computer monitor. It is worth noting that in Lauricella et al., a keyboard cover had to be used to prevent children from continuously pressing on irrelevant keys. Thus, the relatively complex computer interface may not be suitable for younger children who are more likely to exhibit the deficit (Kirkorian, Pempek, et al., 2016).

With the advent of touchscreen devices with their highly intuitive interfaces, and consequently the surge in interactive touchscreen media use among young children (Ofcom, 2020; Rideout, 2017), researchers are now exploring the efficacy of touchscreen interactivity in supporting young children's learning from screens. Using the pseudo-social contingency afforded by touchscreens (i.e., the screen responding instantly to physical touches), Choi and Kirkorian (2016) investigated the effects of different types of contingency on 2-year-olds' performance in a tablet-based object retrieval task. In this study, children were shown either a *non-contingent*, *general-contingent*, or

specific-contingent video on a tablet. The three conditions differed in that, in the non-contingent condition, children were to passively watch a video of a cartoon teddy bear hide, while in the general-contingent and specific-contingent conditions, children were instructed to tap anywhere on the screen and tap on the teddy bear respectively to watch it hide. The results suggested that specific-contingency improved object retrieval in the younger age group ($M_{\text{age}} = 25.15$ months) but hindered performance in the older age group ($M_{\text{age}} = 33.94$ months).

Kirkorian, Choi, et al. (2016) reported similar results in the word learning domain using the same conditions as Choi and Kirkorian (2016): (a) non-contingent, which involved children passively watching a novel object being removed from a box and then labelled; (b) general-contingent, which required children to tap anywhere on the screen before a novel object was shown and labelled; and (c) specific-contingent, which required children to tap on a box to reveal a novel object and to hear its label. In particular, specific contingency supported learning in the younger age group (23.5–27.5 months) but not in the older age group (27.5–32.0 months), who instead benefited more from passive video watching. To account for these findings, Kirkorian, Choi, et al. suggest that specific contingency provides younger children with the required attentional support to encode target features in complex scenes (Franchak et al., 2015; Frank et al., 2009; Kirkorian et al., 2012). Conversely, the same support may have caused older children to encode redundant features, thereby impeding their generalisation beyond the screen context (Vlach & Sandhofer, 2011).

In another tablet-based study, Russo-Johnson et al. (2017) examined the effects of different contingency situations on 2- to 4-year-olds' word learning. Specifically, children were taught the labels of novel objects in one of three conditions in which they were to: (a) *watch* the object move across the river on the cartoon backdrop, (b) *tap* on the object to watch it move across the river, or (c) *drag* the object across the river. Although all children managed to learn words within the app, no main effect of condition was found. There was however a significant main effect of age, with the 2-year-olds learning significantly fewer

words than the 3- and 4-year-olds, and equivalent word learning among the 3- and 4-year-olds.

Taken together, the results on the effects of pseudo-social contingency on learning appear to be mixed across ages and the different types of contingency situations tested.

2.1.3 Self-Directed Learning

The studies discussed in the previous section have focused on interactivity in a controlled context, in that children had no volitional control over what they were to learn. A further way to involve children in a more active learning situation is to allow them to make decisions about the information to be learnt; in other words, learning is *self-directed* (see Gureckis & Markant, 2012, for a review). Among adults, self-directed learning has been shown to be superior to learning via passive observation (e.g., Castro et al., 2008; Markant & Gureckis, 2014). Such findings have been extended and replicated in studies involving children (e.g., Partridge et al., 2015; Ruggeri et al., 2016; Sim et al., 2015). For instance, in the category-learning task in Sim et al. (2015), 7-year-olds who could select which information they wished to acquire performed better than those who merely observed information presented in a random manner. Similarly, Ruggeri et al. (2016) found that giving 6- to 8-year-olds active control over stimuli presentation in a simple memory game enhanced their recognition memory and the advantage persisted in the follow-up test held a week later. In Begus et al. (2014), letting 16-month-olds decide on what information to receive by appropriately responding to their pointing facilitated their learning in an imitation task.

Another study examined the effect of self-direction on 3- to 5-year-olds' learning outcomes in a tablet-based word learning task (Partridge et al., 2015). Children in the *choice* condition were given control over the order in which 15 toys were labelled, whereas those in the *no-choice* condition could only tap on a button in the centre of the screen to hear the labels (in a pre-specified order). The test phase, consisting of tests on children's recognition of 1, 2, 4, and 8 toys

in separate blocks, revealed that self-direction improved information retention in children. However, since the improvement was observed only in the earlier blocks, which tested fewer word–referent associations, it is unclear whether the effect of self-direction only occurred early in learning or whether the complexity of the blocks involving more objects overshadowed the reported effect. Moreover, children could not select the *kind* of information they could learn in this task (i.e., which of a selection of objects they would rather hear the label for). They could only determine the *order* in which objects were labelled.

Recently, Zettersten and Saffran (2019) provided 4- to 8-year olds with the choice of which objects they could choose to be given more information about and examined the influence of such choice on learning. They presented children with either fully ambiguous or disambiguated word–referent mapping situations. In cases where the relative ambiguity of the presented objects was increased, children showed some evidence of preferentially selecting the objects that would resolve the ambiguity. This suggests that children actively choose objects that can reduce their information gap, at least at the older ages tested in the study (see also Sim et al., 2015).

In sum, research on young children's self-directed learning and specifically, the effects of self-direction on their learning from tablets remains extremely limited. Further work is thus needed to understand better self-direction in the context of tablet-based learning, to maximise young children's learning outcomes.

2.2 Early Word Knowledge Assessment

2.2.1 Parent Report

The MacArthur–Bates Communicative Development Inventories (CDI) are one of the most widely used set of parent-report instruments for assessing young children's early language and communicative development (Fenson et al., 2007). Originally developed in American English (Fenson et al., 1993), CDIs have since been adapted into nearly 100 languages (see CDI Advisory Board, 2015, for a list of available adaptations), such as Danish (Bleses et al., 2008a), Mandarin

(Tardif et al., 2008), and Italian (M. C. Caselli & Casadio, 1995). Adaptations have also been developed in a number of sign languages, including American Sign Language (D. Anderson & Reilly, 2002; N. K. Caselli et al., 2020), British Sign Language (Woolfe et al., 2010), and Turkish Sign Language (Sumer et al., 2017).

CDIs typically consist of three forms, namely, CDI–Words and Gestures (CDI–WG), CDI–Words and Sentences (CDI–WS), and CDI–III, each designed for use with children from different age groups. Originally targeting children 8 to 16 months of age and now extended to 18 months, CDI–WG assesses both comprehension and production of early vocabulary, as well as production of communicative gestures. CDI–WS targets children 16 to 30 months of age and assesses both productive vocabulary and morphosyntactic skills, including utterance length and sentence complexity. Finally, CDI–III is a short-form measure targeting children 30 to 37 months of age and assesses productive vocabulary, syntactic maturity, as well as language use (Dale et al., 1998; Fenson et al., 2007).

Compared to brief interactions in laboratory or clinical settings, CDIs systematically utilise parents' knowledge about their child's language and therefore allow for a more comprehensive and representative picture of children's early language development (Fenson, Pethick, et al., 2000). Beyond being cost-effective, CDIs are also reliable and valid, not only with typically developing children (Fenson et al., 1993; Fenson et al., 2007; Law & Roy, 2008; Pan et al., 2004; Rescorla et al., 2005), but also with children with developmental disabilities (Galeote et al., 2016; Luyster et al., 2007; Mayne et al., 1999; Mayne et al., 1998; Thal et al., 2007; Yoder et al., 1997).

Through the application of CDIs in various languages, similarities have been observed in lexical development trajectories in children speaking different languages (Bleses et al., 2008b; Braginsky et al., 2019; Frank et al., 2021). For instance, despite the presence of large individual differences, children typically begin to produce their first words between 12 and 20 months of age (Bleses et al., 2008b; Devescovi et al., 2005; Fernald et al., 1998). After 18 months of age, their vocabulary acquisition rate increases rapidly (E. Bates & Goodman,

1997; Fernald et al., 2006; Fernald et al., 1998). Using CDIs, E. Bates and Goodman (1997) identified two important leaps in children's vocabulary development, with the first occurring between 16 and 20 months and the second, between 24 and 30 months—although some may not experience these leaps at the same ages (Reznick & Goldfield, 1994). Other CDI-based studies (e.g., E. Bates & Goodman, 1997; M. C. Caselli et al., 1999; Conboy & Thal, 2006; Devescovi et al., 2005; Marjanovič-Umek et al., 2013; Stolt et al., 2009) highlighted a strong relationship between lexical and grammatical development. For classifying children as late talkers or late language learners, a common criterion has been total expressive vocabulary at or below the 10th percentile on CDI–WS (Dale et al., 2003; Desmarais et al., 2008; Ellis Weismer, 2007; Ellis Weismer & Evans, 2002; Rescorla & Dale, 2013).

Furthermore, studies have consistently demonstrated that children's vocabulary in their second year of life, as assessed by CDIs, is predictive of later language skills (Duff, Reen, et al., 2015; Henrichs et al., 2011; Kemp et al., 2017; Lee, 2011; Marchman & Fernald, 2008; Reilly et al., 2010), reading achievement (Bleses et al., 2016; Harlaar et al., 2008; Morgan et al., 2015), kindergarten readiness (Duff, Reen, et al., 2015; Forget-Dubois et al., 2009; Friend et al., 2018; Morgan et al., 2015), social-emotional functioning (Irwin et al., 2002), cognition (Marchman & Fernald, 2008), mathematics achievement (Bleses et al., 2016; Morgan et al., 2015), as well as behavioural functioning (Morgan et al., 2015).

Despite the many advantages and widespread applications of CDIs, the completion of the forms requires a significant amount of time and the parent to be literate. The American English CDI–WS, for instance, includes a vocabulary checklist of 680 words, organised into 22 semantic categories (e.g., vehicles, toys, people, action words, descriptive words, and question words). Under circumstances when a rapid assessment is desirable (whether in a battery of tests or in multilingual environments) or when parents have low literacy skills, the applicability of CDIs becomes limited. To address these drawbacks, Fenson, Pethick, et al. (2000) developed the first short-form versions of CDI–WG and CDI–WS with items drawn from the full forms. The former consists of an

89-item checklist, while the latter consists of two 100-item checklists to allow for repeated administrations. As with the full CDIs, these short forms have demonstrated high validity and reliability and are at the same time highly correlated with the full forms, thus making them a useful alternative when time or parental literacy is limited (Fenson, Pethick, et al., 2000). Nevertheless, due to their brevity, these short forms may not be as precise as the full forms and may fail to account for individual differences in children and in the parents completing the forms. The short-form CDI–WS, for instance, suffers from a ceiling effect after 27 to 28 months and even more so when children have a large vocabulary. Furthermore, it takes much time and effort to develop such forms for each language in order to ensure a good balance of items from different semantic categories, as well as items with varying levels of difficulty.

With the objective to develop a short-form version of CDI–WS that is tailored to each child, while maintaining the accuracy and precision of the full CDI, Makransky et al. (2016) employed item response theory (IRT; Embretson & Reise, 2000) in calibrating an item bank from which items are selected and administered through computerised adaptive testing (CAT; van der Linden & Glas, 2010). In their approach (hereafter referred to as *CDI–CAT*), item parameters (i.e., difficulty and discrimination) are first estimated, followed by the assessment of item fit for the two-parameter logistic IRT model. The testing process begins by administering 10 initial items with maximal item information sampled at random from the full CDI. The CAT algorithm then selects subsequent items based on the estimation of the child's ability at each point (i.e., item) in the test as well as the item parameters. Using the American English CDI–WS normative sample which consisted of 1,461 children between 16 and 30 months of age, real-data simulations were conducted with tests consisting of 5, 10, 25, 50, 100, 200, 400, and all 680 items sampled from the full CDI. The results revealed that CDI–CAT performed well at 50 items and above, with correlations above .95 with the full CDI, average *SE*s below .20, and reliability coefficients above .96 (above what the authors described as a minimal threshold for test acceptability). Nevertheless, as pointed out by Makransky et al., some

reduction in performance can be expected with novel empirical data due to systematic or random error. As a result of the semantically unstructured ordering of the test items in CDI–CAT (as opposed to the semantic grouping adopted in full CDIs), parents may also respond differently to the same item in the full and short forms. Furthermore, interpretation of the model-derived scores (i.e., latent ability) clearly suffers, since these cannot be directly mapped back to the scores most typically used for CDIs (i.e., raw vocabulary sums or percentiles).

Recently, Mayor and Mani (2019) presented a language-general approach that capitalises on the richness of Wordbank (Frank et al., 2017), an open repository for cross-linguistic CDI data from over 75,000 children across 29 languages. Their approach derives estimates of full CDI scores by combining a subset of items randomly drawn from the full forms with (prior) CDI data sampled from language-, sex-, and age-matched children on Wordbank. Real-data simulations conducted using the American English (Fenson et al., 2007), German (Szagun et al., 2014), and Norwegian (Simonsen et al., 2014) CDI–WS data revealed that at 50 items, correlations reached .97, with average *SE*s of .05 and reliability coefficients of .99, suggesting that their approach, which takes into account children's age and sex, outperforms CDI–CAT. Empirical validation with 25- and 50-item checklists administered to parents of German-speaking children further demonstrated good performance, with correlations of .96, average *SE*s of .14, and a reliability coefficient of .98, above Makransky et al.'s (2016) recommended thresholds, even when parents showed inconsistencies (about 10–15% of responses) in responding in the full and short forms. However, to capture the full extent of the large variations in vocabulary acquisition (e.g., within- and between-age variations, sex differences; Fenson et al., 2007), Mayor and Mani's approach requires a considerably large sample size on Wordbank. The German CDI–WS data set, for instance, being the smallest data set used in Mayor and Mani's work, consists of over 70 children in each age group. Thus, it is unclear how their approach would perform with smaller sample sizes (e.g., for languages having fewer computerised forms on Wordbank). Another obvious limitation to this approach follows from the *random* sampling of test items,

which may potentially lead to samples of items that are minimally informative. In other words, randomly sampled items may either be too easy (e.g., "car" is produced by 99% of 30-month-olds in American English) or too difficult (e.g., "sofa" is produced by just 1% of 16-month-olds), and may hence inform little about a child's language ability (Fenson et al., 2007; Frank et al., 2017).

While studies have evinced predictive relationships between early vocabulary as measured by parent report (e.g., CDIs) and later outcomes, concerns have been expressed over its exclusive use from an applied perspective (e.g., in justifying clinical decision-making)—particularly before 2 years of age (Duff, Nation, et al., 2015; Feldman et al., 2005; Feldman et al., 2000; Fenson, Bates, et al., 2000; Tomasello & Mervis, 1994), in light of the fact that (a) parent-reported vocabulary accounts for only a small, if not modest, proportion of the variance in outcome measures at the group level (Bleses et al., 2016; Dale et al., 2003; Duff, Reen, et al., 2015; Ghassabian et al., 2014; Henrichs et al., 2011; Morgan et al., 2015); and (b) the predictive power of parent-reported vocabulary is insufficient at the individual level (Dale et al., 2003; Duff, Reen, et al., 2015; Feldman et al., 2005; Law & Roy, 2008; Westerlund et al., 2006; Zambrana et al., 2014).

More precisely, parent-reported comprehension has been argued to be less reliable than production, since reports on production are based on observable instances, whereas reports on comprehension are more subjective in that parents can only infer comprehension based on children's non-verbal responses to language (Feldman et al., 2000; Houston-Price et al., 2007; Tomasello and Mervis, 1994; but see Styles and Plunkett, 2008). For instance, when parents report that their child "understands" or "comprehends" the word "milk", it is difficult to establish whether the child truly understands the word as the referent to a glass of milk. It may well be that parents assume word comprehension (a) when the child produces a response that is indicative of either recognition of the sound of a word, or familiarity with the referred object or event; and/or (b) when the child's response is cued by the rich context in which a word is heard (Houston-Price et al., 2007; Tomasello & Mervis, 1994). Additionally,

given children's vocabulary spurt during the second year, parent reports of comprehension may also be inconsistent over time on an item-by-item basis (Yoder et al., 1997). Thus, the use of supplemental measures is recommended, especially in clinical settings (Dale et al., 2003; Fenson et al., 1993).

2.2.2 Direct Language Measure

Revisiting the earlier concerns that parent-reported comprehension may be subject to interference, context-dependent, and unstable over time, a direct language measure can serve both as a convergent and a supplemental measure of parent reports, by tapping children's strong, rather than weak, word–referent associations (Friend et al., 2019). Such associations, also referred to as decontextualised associations, are stable associations which can be recognised without the supporting context in which the associations were formed (Friend et al., 2018; Friend et al., 2019). However, the challenges inherent in directly assessing very young children's language skills, including the difficulty in maintaining children's interest and attention (Friend & Keplinger, 2003) as well as behavioural non-compliance (Kaler & Kopp, 1990), have impeded research in this area. At the time of this writing, only a few direct measures have been developed to assess language comprehension among children below 2 years of age, such as the Intermodal Preferential Looking Paradigm (IPLP; Golinkoff et al., 1987; Hirsh-Pasek & Golinkoff, 1996) and its offshoot, the looking-while-listening procedure (LWL; Fernald et al., 2006; Fernald et al., 1998), as well as the Computerized Comprehension Task (CCT; Friend & Keplinger, 2003, 2008). Among these, only the CCT focuses on assessing children's decontextualised vocabulary size, whereas IPLP and LWL focus on assessing lexical comprehension and processing efficiency based on children's visual fixations respectively.

Built on IPLP (Fernald et al., 1998; Hirsh-Pasek & Golinkoff, 1996) and picture-based (Ring & Fenson, 2000) approaches, the CCT (available in English, Spanish, and French) aims to facilitate direct assessments for very young children in clinical settings, while circumventing the need for labour-intensive

gaze data coding and analysis (Friend & Keplinger, 2003, 2008; Friend & Zesiger, 2011). The CCT begins with four training trials, followed by 41 test trials and 13 reliability trials, which altogether take less than 10 minutes to complete. Pairs of images are presented in a forced-choice format on a touchscreen and the experimenter prompts the child to point to or touch an image in response to the target word heard (e.g., Where's the *bus*? Touch *bus*!). Target words consisting of nouns, verbs, and adjectives vary in difficulty and are selected from both CDI–WG and CDI–WS based on norming data at 16 months of age (Dale & Fenson, 1996). As the assessment requires both lexical retrieval (i.e., retrieving word–referent associations upon hearing the target word) and hypothesis testing (i.e., deciding on an association and selecting the image that represents the referent of the target word), correct haptic responses are taken as evidence of children's decontextualised word knowledge (Friend et al., 2019).

Such measure of children's language comprehension has been found to be reliable and valid across three languages (including bilinguals), with scores correlating significantly with CDIs (Friend & Keplinger, 2003, 2008; Friend et al., 2012; Friend & Zesiger, 2011; Hendrickson et al., 2015; Poulin-Dubois et al., 2013). The CCT is also effective in maintaining children's attention as well as improving compliance and thus, can be used with children as young as 16 months and up to 24 months of age (Friend & Keplinger, 2008; Friend & Zesiger, 2011; Hendrickson et al., 2015). In terms of predictive validity, decontextualised vocabulary comprehension as measured on the CCT has been shown to predict productive lexical diversity in a language sample (Friend et al., 2012) and language skills (Friend et al., 2019; Patrucco-Nanchen et al., 2019) in the third year of life, as well as kindergarten readiness in the fourth year of life (Friend et al., 2018). Other advantages of the CCT include its portability and ease of administration.

While the CCT offers a more objective measure of children's language comprehension, as with the development of language-specific CDIs, a tremendous amount of time and effort is required to adapt the CCT to each language so that the assessment reflects linguistic, cultural, and contextual differences; for

instance, the word "tortilla" is only relevant for young Spanish speakers. To cite another example, the word "snow" is relevant for young English speakers, but not for young Malay speakers who only acquire the word much later (Łuniewska et al., 2019). Moreover, the CCT takes a one-size-fits-all approach, in that children's language comprehension is assessed based on a fixed selection of words, which may fail to account for individual differences in children. For these reasons, a more generalisable and effective approach to developing direct assessments that are tailored to each individual child is desirable.

2.2.3 Tablet-Based Assessments

Technology-based assessments, in which microchip-based devices (e.g., computers) are used in collecting, analysing and/or reporting data, are not new and have been shown to facilitate administration (Bunderson et al., 1989; Green, 1988; Hambleton, Zaal, et al., 1991; Wise & Plake, 1989) and scoring (Bugbee Jr. & Bernt, 1990; Kyllonen, 1991). Compared to standard paper-and-pencil assessments, technology-based assessments allow a more standardised experience in many ways, which is an advantage when assessments are to be (a) administered in different locations and/or by different administrators, or (b) adapted to multiple languages. First, technology-based assessments enable precise control over the presentation of test items, including timing and order. Second, verbal instructions can be kept to a minimum, since tasks are demonstrated on-screen and practice trials can be repeated as many times as needed. Relatedly, experimenter effects can also be minimised, especially in developmental assessments which typically require experimenters to interact with children. Beyond response data (i.e., the correctness or incorrectness of a response), technology-based assessments offer the opportunity to gather process data (e.g., response latencies, sequence in which items are answered) that provide new insights into the behavioural processes underlying the course towards a response (Goldhammer et al., 2014; Han et al., 2019).

Following the advent of tablets, technology-based assessments have now extended to tablets, with tablet-based versions of standard paper-and-pencil

assessments increasingly being administered—not only to adults (e.g., the Wechsler Adult Intelligence Scale [WAIS; Wechsler, 2008], the Wechsler Memory Scale [WMS; Wechsler, 2009]), but also to children (e.g., the Peabody Picture Vocabulary Test [PPVT; Dunn, 2018], the Expressive Vocabulary Test [EVT; Williams, 2018]). For very young children, however, research on tablets has primarily focused on the educational potential of tablets (see Hirsh-Pasek et al., 2015; Reich et al., 2016; Troseth et al., 2016) rather than on their use as a data collection modality.

In developmental psychology, preferential looking paradigms are typically employed with very young children who are unsuited for standard psychophysical paradigms (e.g., object manipulation, pointing, and requests for action), since they cannot yet reliably produce manual responses to stimuli (Gurteen et al., 2011). Yet, due to the passive nature of looking-based tasks, children quickly get bored when the same paradigm is repeatedly presented (over multiple trials). Consequently, only very few items can be assessed in any single session, making such tasks unsuitable for assessing early language comprehension (Friend & Keplinger, 2003).

On the other hand, tablet-based experimental paradigms can potentially solve these issues. Owing to the absence of the need for manipulating additional input devices that may require more refined motor skills and eye-hand coordination (e.g., mouse and keyboard; Donker & Reitsma, 2007; Kucirkova, 2014), tablets are easy to operate, even for the youngest children. For instance, in Abdul Aziz et al. (2014), 2-year-olds could reliably perform both the tap and drag/slide gestures. Likewise, in Marsh et al. (2015), more than half of the children aged between 0 and 2 years could swipe the screen unassisted by an adult (e.g., to turn the pages of electronic books). In contrast to touchscreen paradigms employed in laboratory settings—such as the CCT (described in Section 2.2.2)—in which full arm movements are often necessary since screens are typically mounted on a wall or placed on a desk, tablet-based paradigms require only minimal motor movements and are much more portable, thanks to the small form factor of tablets. Additionally, tablet-based experimentation can

be more engaging than classical psychophysical paradigms as children aged between 17 and 26 months have been found to be more attentive and engaged when reading electronic picture books (on a tablet) than print versions with identical content (Strouse & Ganea, 2017). In Couse and Chen (2010), 3- to 6-year-olds who were learning to draw on tablets were seldom frustrated and persisted in learning, despite encountering multiple technical incidents (i.e., computer-related problems that interrupted children's drawing on the tablet).

To investigate the viability of tablets in collecting developmental data, Frank et al. (2016) compared three different data collection modalities: tablet, eye tracker, and picture book, using a word recognition task with 1- to 4-year-olds. In terms of data yield, the tablet modality produced high completion rates (86–100%) among children across all age groups, except for the 1-year-olds (44%). Despite the general advantage of the tablet modality over the eye tracker, the picture book paradigm produced the highest completion rates due to the involvement of an experimenter who could pace the task accordingly. Nevertheless, Frank et al.'s results indicate that tablets can be reliably used to collect reaction time (RT) and accuracy data and that the tablet modality compares favourably with both the eye-tracking and the picture book paradigms. For instance, with the tablet modality, 1-year-olds performed above chance in trials containing familiar words (regardless of the novelty/familiarity of the distractor), whereas with the eye tracker, performance was only above chance in trials with novel objects as distractors and with the picture book paradigm, trials with familiar objects as distractors. The authors also pointed out that the employment of a tablet-based experimental paradigm, due to its low cost and high accessibility, can potentially facilitate large-scale, parallel data collection (e.g., by distributing tablets to multiple participants at one time or at different locations), thus obviating the need for separate, one-on-one sittings. Hourcade et al.'s (2012) finding that the use of tablet apps encouraged pro-social behaviours (e.g., through sharing a tablet) among children with autism spectrum disorders (ASD) further suggests that the tablet modality can promote the inclusion of special populations.

Continuing in the same vein as Frank et al. (2016) to examine the viability of tablets in developmental cognitive research, Semmelmann et al. (2016) compared 1- to 10-year-olds' results with adults' gathered from six studies mediated through tablets, including two two-alternative forced choice (2AFC) sorting and recalling tasks with different levels of difficulty, an extended version of the aforementioned 2AFC sorting and recalling task with a perception task added, a visual search task, an extinction learning paradigm, as well as a simple visuo-spatial paradigm. The aim was to establish—for children across different age groups—potential limits relating to the motor requirements, sophistication, and length of experimental paradigms when these are ported to tablets. Overall, the findings suggest that children from age 2 onwards have the necessary motor skills to interact with tablets (e.g., by providing tap, drag and drop responses) and are able to produce reliable and robust results in terms of RT and accuracy with high completion rates (about 84%), as long as the experimental task design is age-appropriate. One-year-olds, on the other hand, had lower completion rates (about 64%)—in line with Frank et al. After age 5, children's RTs did not differ from adult values. With regard to accuracy, 70% of adult level was achieved starting at age 1 and accuracy increases with age. Finally, based on the finding that 9- and 10-year-olds became bored after about 15 minutes in the 25-minute long "2AFC Sort Recall Perception" task, the authors recommended that tasks be kept below 15 minutes.

While both Frank et al. (2016) and Semmelmann et al.'s (2016) findings do not seem to lend support to the employment of tablet-based experimental paradigms among 1-year-olds, due to low completion rates (as a result of their inexperience with tablets as well as the lack of motivation to engage in the tasks, or the inability to understand the task demands), the success of the touchscreen-based CCT—designed for 1-year-olds—across languages seems to suggest otherwise (Friend & Keplinger, 2003, 2008; Friend & Zesiger, 2011). Specifically, compared to the Comprehension Book (i.e., a picture book paradigm; Ring & Fenson, 2000), the CCT yielded more data, with children being more attentive, attempting more trials, and responding with higher

accuracies (Friend & Keplinger, 2003, 2008). Furthermore, 1-year-olds' consistent above-chance performance across test and retest indicate that they are able to produce reliable responses in a touchscreen-based paradigm (Friend & Keplinger, 2003, 2008; Friend et al., 2012). It is noteworthy though, that both pointing and touching are taken as valid responses in the CCT (as well as the Comprehension Book), thus compensating for children's inexperience with touchscreens. To obtain optimal performance, Friend and Keplinger (2008) also recommended that children be made aware of the context of the task.

Yet, one may still argue that the CCT does not yield complete data sets. However, Friend and Keplinger's (2008) finding that children made fewer attempts in difficult trials (i.e., trials containing later-appearing words in the lexicon) than easy trials (i.e., trials containing early-appearing words), which in part led to lowered data completeness, implies that the absence of volitional response reflects word knowledge—that children are unable to distinguish the target from the distractor—rather than behavioural non-compliance or the lack of motivation. In support of this view, Hendrickson et al. (2015), using both looking and touching measures in a CCT-based assessment, found that children were significantly faster at processing the target word (measured by the latency to shift gaze from distractor to target) in trials in which they provided a touch response (regardless of correctness) than trials in which they did not. In other words, when children provided a response, it either signified robust (associated with a correct response) or partial word knowledge (associated with an incorrect response). Conversely, when children did not provide a response, it represented children's true inability to map the target word to its referent.

In sum, coupled with recent advances in the approach to developing short-form CDIs (e.g., Makransky et al., 2016; Mayor & Mani, 2019), the employment of a tablet-based word recognition paradigm can potentially facilitate the administration of CDIs to young children, thus providing a performance-based assessment to supplement parent reports. The only requirements are that (a) the context of the task be clarified to children, (b) the duration of the task be kept below 15 minutes, and (c) that the task be

interactive. In order to extend the accessibility of the assessment to 1-year-olds, care should also be taken to familiarise them with the gestures required to provide a response (e.g., tapping) by including a training phase prior to the test phase.

2.3 Employing Web Technology in Data Acquisition

2.3.1 Web-Based Methods in Psychological Research

Web-based research methods can be divided into four categories: *non-reactive web-based methods*, *web surveys*, *web-based tests*, and *web experiments* (Reips, 2006).

Briefly, non-reactive web-based methods involve the use and analysis of data collected online, in an unobtrusive or non-invasive manner (i.e., people under investigation are unaware of the data-recording process; thus, their behaviours are measured in a "natural state"; Janetzko, 2017). Some examples of studies that employed non-reactive web-based methods include Jones et al. (2016), who investigated post-trauma word usage by analysing Twitter data and Stieger and Reips (2010), who collected data on participants' mouse cursor positions, clicks, and key presses during an online questionnaire to gain further insight into their answering processes.

Web surveys have, for a long time, been used to facilitate data collection from large and diverse samples (W. C. Schmidt, 1997). Mindell et al. (2010), for instance, investigated cross-cultural differences in young children's sleep patterns and sleep problems through a web survey conducted among 29,287 parents across 17 countries/regions. In Reimers (2007)—one of the largest studies to date—over 250,000 responses were collected in a web survey on human sex differences within just three months. Despite being the most commonly used web-based research method (due to the ease of creation and administration), web surveys were initially met with scepticism (see Gosling et al., 2004, for an evaluation of the six common preconceptions about questionnaire data collected on the internet). In short, Gosling et al. argued that four out of the six preconceptions (i.e., internet

participants are socially inept and unmotivated; and internet findings are inconsistent across different presentation formats and differ from findings based on traditional methods) are unfounded and that internet samples—despite not being representative of the general population—are at least as representative as traditional samples. While Gosling et al. confirmed that the integrity of data collected on the internet may be compromised by the anonymity provided to participants (e.g., multiple submissions from repeat responders), this can be eliminated by taking precautionary steps (Birnbaum, 2004).

Web-based tests refer to web-versions of psychological tests and are a subtype of web surveys. Web-based psychological tests have consistently been reported to be qualitatively (psychometrically) similar to standard paper-and-pencil tests, though there are also reports of instances where quantitative differences (e.g., equality of means, equality of variances) have been found (see Buchanan, 2007, for a review). Nevertheless, as Meyerson and Tryon (2003) suggest, quantitative equivalence can be established either by adding or subtracting a constant or by using an equipercentile transformation.

The earliest web experiments (in which one or more independent variables are manipulated, as contrasted with web surveys) in psychology can be traced back to 1997 when Krantz et al. conducted experiments on the determinants of perceived attractiveness of females over the web and in the laboratory, and compared the results from both methods. Despite the differences in the environmental settings and experimental procedures (e.g., in the laboratory, the pace of the experiments was controlled by the experimenter, whereas over the web, participants could respond at their own pace), correlational and regression analyses revealed high validity of the results obtained via the web-based method. In the same year, Smith and Leigh (1997) also collected data simultaneously on the web and in the laboratory by replicating Ellis and Symons' (1990) study on sex differences in sexual fantasies. Overall, the results were congruent with those reported by Ellis and Symons and no significant differences were found between data collected online and in the laboratory. Following these examples, more researchers have shown that web experiments can yield results similar to those

conducted in the laboratory across a wide range of designs (e.g., Crump et al., 2013; Germine et al., 2012; Hilbig, 2016; Semmelmann and Weigelt, 2017; see also Krantz and Dalal, 2000, for a review on earlier web experiments).

2.3.2 Advantages of Web-Based Methods

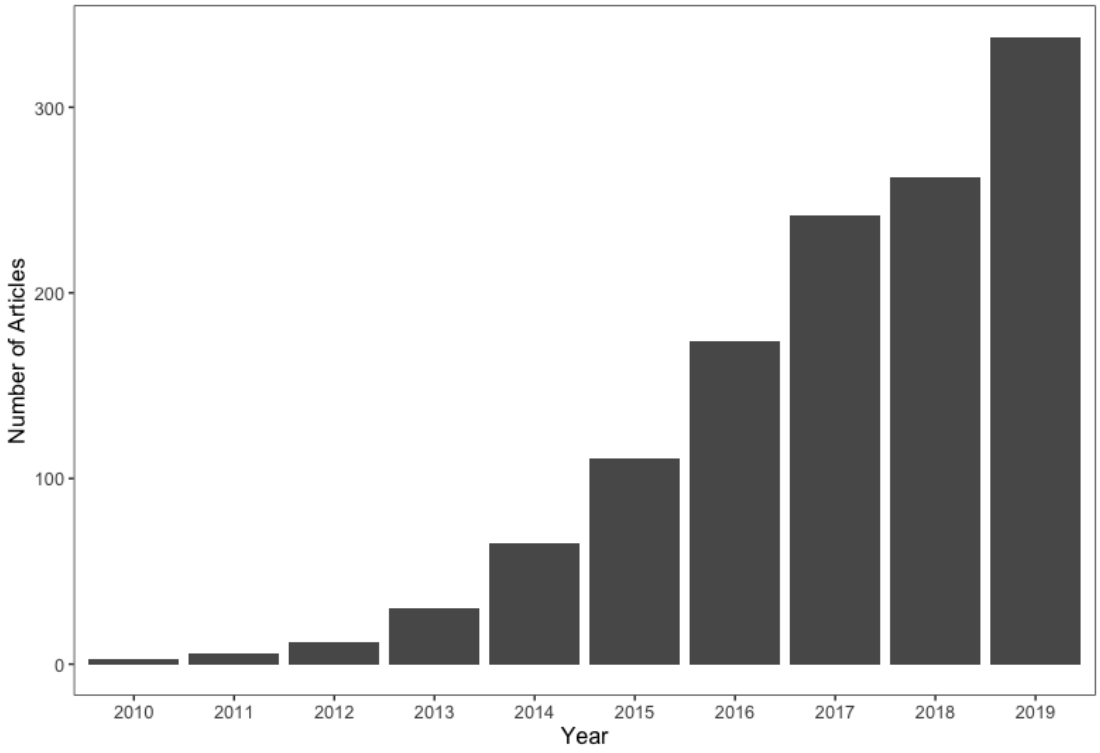
On 11 March 2020, the World Health Organization declared Coronavirus disease 2019 (COVID-19) a pandemic (World Health Organization, 2020). In the midst of the global pandemic—with countries on lockdown (e.g., "Coronavirus: UK lockdown extended for 'at least' three weeks", 2020; Hassan, 2020; Klesty & Fouche, 2020) and scientific research laboratories shuttered—researchers in the fields of cognitive, behavioural, and even developmental psychology are ramping up web experiments, particularly those in which in-person testing is of paramount importance, in order to continue data collection. Even before the pandemic hit, web-based research methods have been steadily gaining popularity in different areas of psychological research in the past decade due to their many associated advantages (Musch and Reips, 2000; Reips, 2007; Reips and Lengler, 2005; see also Figure 2.1).

First, web-based methods allow participants to easily take part in experiments from the comfort (and safety, especially in times of the current pandemic) of their own homes, or from anywhere, as long as they are connected to the internet; in other words, experiments are "brought to the participants" instead of the other way round. This advantage, combined with increased anonymity, is also especially helpful for reaching special populations—for instance, Ogston et al. (2011) assessed hope and worry in mothers of children with an ASD or Down syndrome through an online questionnaire, which allowed mothers to openly express their worries.

Beyond special populations, web-based methods can potentially reach more diverse and representative samples, in keeping with the criticism on sampling WEIRD (Western, educated, and from industrialised, rich, and democratic countries) populations in the literature (Arnett, 2008; Henrich et al., 2010; Sheskin et al., 2020). For instance, the aforementioned web survey by

Figure 2.1

Number of Psychology Articles Using Web-Based Methods by Publication Year Found on Web of Science



Note. Numbers are based on a search conducted on 29 June 2020, using the search term "Mechanical Turk or MTurk" (Amazon's crowdsourcing marketplace) within the "psychology" research area on Web of Science.

Mindell et al. (2010) allowed cross-cultural differences in young children's sleep patterns and sleep problems to be uncovered.

In contrast to laboratory-based methods, web-based methods allow large amounts of data to be collected within a short amount of time and/or in parallel, and in a relatively inexpensive way (e.g., in terms of labour and administrative costs). Through the use of crowdsourcing marketplaces, such as Amazon Mechanical Turk (MTurk)[6] and Prolific (Palan & Schitter, 2018), participants can be recruited rapidly and at lower hourly rates (Buhrmester et al., 2011). Although concerns have been expressed on ethical grounds regarding the low median wage of $2.00/hour on MTurk (Hara et al., 2018; Semuels, 2018), a more recent study suggests that this has risen to $5.70/hour (Litman et al., 2020). Furthermore, there is no need for testing rooms, laboratory equipment (including expensive software licenses), bureaucracy relating to scheduling, insurance, and so on.

Another important advantage relates to the openness (i.e., transparency and accessibility) of the research process. In response to the replication and reproducibility crises in science, the "open science" movement, which encompasses practices such as enabling open access to published research output, the methodology of studies, along with any data, code, and results, has been introduced (Crüwell et al., 2018; van der Zee & Reich, 2018). By putting experiments online, the materials and procedures involved can be made accessible to other researchers, thus permitting the open archival and sharing of experiments, as well as the possibility of collaboration across laboratories working in the same research area. Moreover, most, if not all, of the advantages of technology-based assessments (e.g., a more standardised way of presenting experiments, the possibility to collect process data; see Section 2.2.3) apply.

2.3.3 Concerns Regarding Web Experiments

While the validity and reliability of web-based methods have consistently been demonstrated (Buchanan, 2007; Germine et al., 2012; Gosling et al., 2004;

[6]https://www.mturk.com/

Krantz & Dalal, 2000; Ramsey et al., 2016), timing has been a major concern regarding web experiments—especially those that are stimulus-controlled and/or use time-based performance measures (Plant, 2016). In particular, the timing issue can be divided into the timing of stimulus presentation and the timing of response recording.

With regard to stimulus presentation, prior work has shown that stimuli may not be presented for the exact duration or at the exact time intended in web experiments (Barnhoorn et al., 2015; Garaizar et al., 2014; Pronk et al., 2020; Reimers & Stewart, 2015; W. C. Schmidt, 2001). Screens refresh at a constant rate (typically at 60 Hz, i.e., approximately 16.67 ms for each frame). If stimulus presentation is not synchronised with screen refreshes or if the presentation duration is shorter than the refresh interval, the number of frames realised may be different from the number of frames intended (what is termed *missed frames* in Garaizar et al., 2014). This may pose a problem for tasks requiring very precise or very brief stimulus presentation durations. For instance, while Crump et al. (2013) successfully replicated several classic RT and attention tasks online, including Stroop, Task-switching, Flanker, Simon, Posner cueing, and attentional blink tasks, they did not manage to fully replicate the masked priming task—when presentation durations of 64 ms or less were required. These findings were broadly mirrored by Semmelmann and Weigelt (2017) who also successfully replicated the Stroop, Flanker, Posner cueing, and attentional blink tasks and partially replicated the masked priming task, across three different settings (i.e., *lab*, *web-in-lab*, and *web*; but see Barnhoorn et al., 2015).

With regard to response timing, the use of web technology has been found to overestimate RTs and such overestimations vary with the use of different hardware (e.g., keyboards, CPUs; Neath et al., 2011; Pronk et al., 2020; Reimers & Stewart, 2015) and software (e.g., operating systems, browsers; Plant & Quinlan, 2013; Pronk et al., 2020; Reimers & Stewart, 2015; Semmelmann & Weigelt, 2017). In studies that simulate a human participant with known RTs (e.g., by using external hardware, such as a microcontroller, to detect stimuli and subsequently trigger a solenoid to generate screen touches or key presses), clear

additive lags in RT measurements have been observed when using different software packages (ScriptingRT, E-Prime, DMDX, Inquisit, and SuperLab; 56–98 ms; Schubert et al., 2013), different implementations (Flash, HTML5) on different computer systems (30–100 ms; Reimers & Stewart, 2015), different implementations (Flash, JavaScript, and Java) with different types of keyboards and CPUs (34–74 ms; Neath et al., 2011), as well as different browsers (Chrome, Firefox, and Safari) running on different devices (Android, iOS, MacOS, and Windows; 57–133 ms; Pronk et al., 2020).

Other studies comparing human participants' RTs found that JavaScript, relative to Psychtoolbox (a standard laboratory software), overestimated RTs by 25 ms (de Leeuw & Motz, 2016), 37 ms when tested in laboratory settings, and 87 ms when tested online (i.e., outside of the laboratory; Semmelmann & Weigelt, 2017). Nevertheless, RT overestimations generally appeared to vary little within any single configuration used (with standard deviations typically falling within the range of 5–10 ms) and can be compensated for (e.g., by using a within-subjects design or when using a between-subjects design, recruiting about 10% more participants; Pronk et al., 2020; Reimers & Stewart, 2015). The aforementioned solid replications of classic RT effects (Barnhoorn et al., 2015; Crump et al., 2013; Semmelmann & Weigelt, 2017)—despite the presence of additive lags—further indicate that such lags are offset when taking a difference between two or more conditions.

In addition to timing, there is generally a lack of control on environmental factors when conducting studies online, as opposed to conducting studies in highly standardised laboratory settings. A participant may be less committed or more easily distracted when taking part in a study from home—in the absence of a proctor or other participants. Indeed, online participants have self-reported that they were often engaged in other tasks, such as watching television and listening to music, while completing studies (Chandler et al., 2014). They also self-reported higher degrees of distraction from mobile phone use, talking to another person, and internet surfing relative to those who participated in the laboratory (Clifford & Jerit, 2014). However, the same study found no difference

in four of five attention checks between online and laboratory participants; and in the only case where a difference was found, online participants had a higher pass rate than laboratory participants, suggesting that decreased attention does not necessarily pervade online samples (Clifford & Jerit, 2014). Several more studies have compared data quality between web- and lab-based studies and all of these pointed to encouraging results (e.g. Casler et al., 2013; de Leeuw & Motz, 2016; Hilbig, 2016; Kim et al., 2019; Miller et al., 2018).

Taken together, these findings lend support to the notion that web technology can be suitably used for acquiring data in common psychophysical research, even in poorly standardised domestic settings (i.e., at home; Miller et al., 2018), as long as precise stimulus timing is not strictly required and relative, rather than absolute, RTs are the focus of interest. Furthermore, as continuous improvements are made in web (e.g., browser, HTML5, JavaScript) and hardware technology, concerns regarding the timing accuracy may even soon become obsolete.

2.3.4 Overcoming the Technical Barrier

Despite their many potential advantages, the potential of web experiments has yet to be realised to its full extent: as conducting web experiments requires specialised knowledge (of technological particularities), that includes, but is not limited to, constructing web pages that present stimuli, capture and transmit participants' responses, configuring servers to host experiments, as well as programming databases to store experiment data, the adoption of web experiments has generally been limited to those with the resources to overcome this technical barrier.

Recently, a growing number of tools that streamline the process of conducting web-based studies have become available, ranging from experiment builders, to study management systems, participant recruitment services, and even platforms providing holistic integrated services, thereby allowing the technical barrier to be, at least, partially alleviated (see Table A.1 in the

appendix for an overview of tools that are actively maintained, i.e., with updates in 2019).

As detailed in Table A.1, some experiment builders attempt to simplify programming for researchers by providing libraries containing pre-programmed components that are commonly used in psychological experiments (e.g., jsPsych; de Leeuw, 2015). Others eliminate the need for programming by providing a graphical user interface (GUI) with which experiments can be built with just mouse clicks—and to extend this to more complex designs, libraries containing templates for common experimental paradigms (e.g., Tatool Web, PsychoPy with PsychoJS, OpenSesame Web; Mathôt et al., 2012; Peirce et al., 2019; von Bastian et al., 2013). The option to program custom scripts is typically offered as well, to maximise the versatility of the tools.

To take experiments online, the experiment code, stimuli, and any dependencies (e.g., libraries) will need to be hosted on a server. This typically requires knowledge of and familiarity with server technologies. Fortunately, study management systems, such as Open Lab and Pavlovia, exist to take care of setting up a web server and a database, managing access permissions, etc. These tools also provide a GUI instead of the commonly used command line interface, to facilitate access and management of experiments and data. JATOS (Lange et al., 2015) offers yet another perspective by providing researchers the option to set up their own servers, in addition to the option to host experiments on its own server.

As noted previously, the two main benefits of conducting experiments online are the ability to reach wider populations and the efficiency in collecting large amounts of data. Once an experiment is published online, anyone with the link to the experiment is able to access it. One caveat is that it may be difficult to determine whether participants are who they say they are (i.e., whether they meet the inclusion criteria of an experiment). Furthermore, due to the large number of participants involved, it may also be cumbersome to handle participant compensation manually. Participant recruitment platforms, such as MTurk, Prime Panels, and Prolific (Palan & Schitter, 2018), offer solutions to

these by automating participant compensation and by employing pre-screening methods or demographic filters to ensure that only target participants are recruited. With Prolific, it is even possible to retarget participants for follow-up or longitudinal studies. The availability of an active, readily accessible pool of participants further adds to the appeal of such platforms.

The different types of tools described so far cater to specific parts of the process of conducting web-based studies (i.e., building an experiment, hosting an experiment, and recruiting participants). Thus, it may still be demanding to stitch together different types of tools to form an ecosystem: for instance, an experiment created with OpenSesame can be hosted on JATOS but is incompatible with Pavlovia; and once the experiment is set up on JATOS, additional steps need to be taken again, to set the experiment up on a participant recruitment platform (e.g., Prolific). To minimise the hassle of having to navigate among different tools, platforms that provide holistic integrated services have been developed (e.g., Gorilla, LabVanced; Anwyl-Irvine, Massonnié, et al., 2020; Finger et al., 2017). Like other experiment builders, these platforms typically feature a GUI-based experiment builder, while others attempt to simplify programming by means of a dedicated scripting language that is simpler to read and write relative to HTML and JavaScript (e.g., Inquisit, PsyToolkit; Stoet, 2017). Testable (Rezlescu et al., 2020) takes an even simpler approach which allows experiments to be created using spreadsheets. Once created, experiments need not be exported to an external study management system as these are hosted by the platform itself and can be managed within the same platform. Some platforms (e.g., Gorilla) even provide seamless integration with participant recruitment services (although, note that there is still a need to set the experiment up on the chosen participant recruitment platform).

While GUI-based tools substantially reduce the amount of effort required to create experiments, programming-based tools offer much more flexibility (in terms of the complexity of an experiment design that is achievable). There are, nevertheless, GUI-based tools that allow complex designs—but there is still the trade-off between ease of use and versatility. Gorilla, for instance, allows

"complex, counterbalanced, randomized, between-subjects designs with multiday delays and email reminders, with absolutely no programming needed" (Anwyl-Irvine, Massonnié, et al., 2020, p. 392), but to be able to do so, one must first go through a large learning curve to master its *three* complex GUIs (see Figure 2.2). In this regard, Testable appears to be superior to Gorilla as similarly complex experiments can be created in a relatively straightforward manner (i.e., working with a spreadsheet). On the downside, however, Testable, in its present state, does not offer as many features as other tools (e.g., support for mobile devices, video and audio recording).

Ultimately, the decision on which tool(s) to use boils down to both the researcher's preference and need. When selecting a tool for creating timing-sensitive experiments, extra consideration should also be given with regard to the timing performance of the tool as the timing issue inherent in web experiments described earlier still applies here. In particular, two recent studies comparing the timing performance of different experiment builders (e.g., Gorilla, lab.js, PsychoPy with PsychoJS, Testable, jsPsych) attest to the fact that these tools do not necessarily perform the same across different browsers and operating systems (Anwyl-Irvine, Dalmaijer, et al., 2020; Bridges et al., 2020). Encouragingly, within any single configuration, variability has typically been found to be under 5 ms for stimulus presentation and 10 ms for response timing (Bridges et al., 2020). These tools also provide reasonably accurate and precise timing both in terms of visual stimulus presentation (when presentation duration lasts longer than two frames, i.e., approximately 33 ms) and response recording (when RT is above 100 ms; Anwyl-Irvine, Dalmaijer, et al., 2020), thus once again, bolstering the support for the use of web technology in data acquisition.

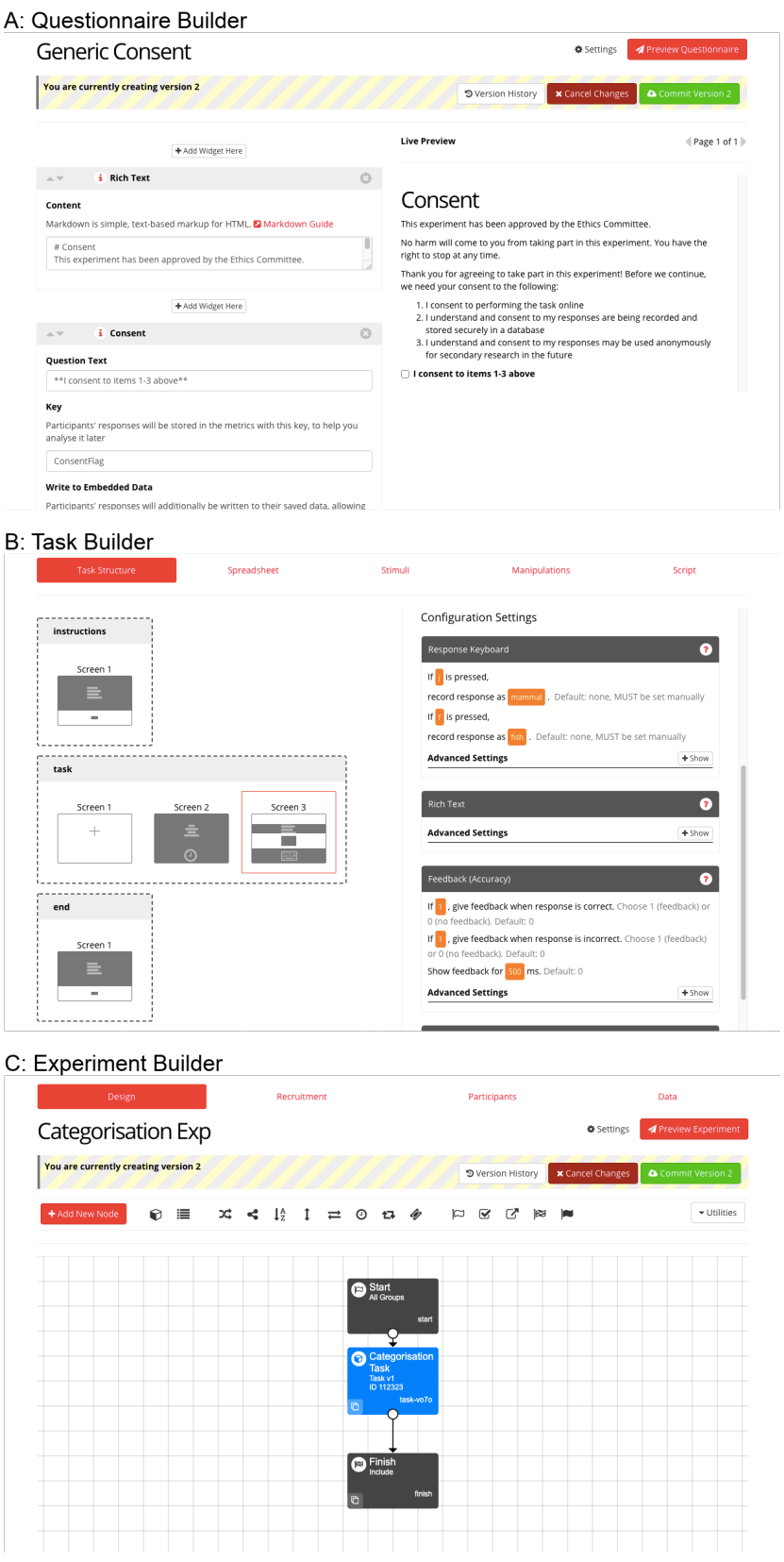
2.4 Current Research Aims and Contributions

2.4.1 Early Word Learning From Tablet Apps

In light of the proliferation of tablets in homes with young children and apps that profess to be educational, the present research aims to examine the

Figure 2.2

Graphical User Interfaces in Gorilla



effects of pseudo-social contingency on young children's word learning from tablets. By giving children active control over their learning experiences, the present research also aims to examine whether self-directed learning (in which children are allowed to make decisions to guide their learning) yields better learning outcomes.

2.4.2 Direct Language Measure via Tablets

Another aim of the present research is to leverage tablets for collecting data among young children. In particular, the present research explores the viability of a tablet-based word recognition task in assessing early word knowledge to potentially serve both as a convergent and a supplemental measure of parent reports. In order to facilitate such assessments, an efficient approach to selecting test items that are effective is desirable. Thus, the present research aims to also further develop short-form versions of CDIs that can reliably estimate children's full CDI scores—either via parent reports or direct assessments—without compromising on the accuracy and precision of the full forms.

2.4.3 Authoring Tool for Online Experiments

A further unique contribution of the present research is e-Babylab, an authoring tool that provides an easy-to-use interface for creating, hosting, running, and managing online browser-based experiments—without the need for prior technical knowledge. The purpose of creating e-Babylab is to add to the arsenal of tools that streamline the process of conducting web-based studies, thus further increasing the accessibility of web-based methods to researchers who are keen to benefit from online experimentation.

2.5 Research Questions

This thesis, therefore, seeks to answer the following research questions through the use of web technology (and e-Babylab):

1. Can young children learn words using tablets?
2. What are the factors that may affect young children's learning from tablets?
3. How can young children's word knowledge be assessed using tablets?
4. How can short-form versions of CDIs be further developed to more efficiently estimate early word knowledge?

2.6 Summary

This chapter provided a review of the literature organised around three central topics that form the basis of this thesis: young children's learning from screens, early word knowledge assessment, and data acquisition with web technology. With regard to young children's learning from screens, studies have suggested that young children learn better through real-life experiences than from passive video viewing. However, it remains unclear whether this deficit is due to reduced social interaction or the fact that children did not get to actively shape their learning situation in such studies. With regard to early word knowledge assessment, the review of the literature has highlighted the need for a direct language measure to supplement parent reports (e.g., CDIs). Short-form CDIs administered in the form of a tablet-based word recognition task may potentially facilitate the development of such measure but there is still room for improvement in established short forms. Finally, web technology–based experimentation was considered as a methodology for data acquisition in the present research and the research aims and questions were presented. The next chapter presents e-Babylab, a new authoring tool for creating, hosting, running, and managing browser-based experiments for online testing.

CHAPTER 3. E-BABYLAB: AN AUTHORING TOOL FOR CREATING ONLINE BROWSER-BASED EXPERIMENTS

This chapter introduces e-Babylab, a new authoring tool developed as a part of this thesis to facilitate the creation, hosting, running, and management of online browser-based experiments. An overview of e-Babylab, along with the typical flow of an experiment created with it, is first provided. The features of e-Babylab are then detailed with accompanying screenshots, followed by the technological aspects involved in its implementation. This chapter incorporates material from the following paper:

> **Lo, C. H.**, Mani, N., Kartushina, N., Mayor, J., & Hermes, J. (2021). *e-Babylab: An open-source browser-based tool for unmoderated online developmental studies.* Manuscript submitted for publication.[7]

3.1 Overview

3.1.1 e-Babylab

e-Babylab is an online platform that allows users or researchers to easily create, host, run, and manage online experiments—without the need for prior experience in programming. Both the authoring interface as well as the experiments are browser-based web applications. Using e-Babylab, experiments can be configured to use any combinations of image, audio, and/or video contents as stimuli and accept key presses, clicks, and touches (on touchscreens) as responses. Other types of explicit responses (e.g., pointing gestures, verbal responses) as well as implicit responses (e.g., eye movement, vocal emotion) can additionally be captured via audio or video recordings. All participant data and

[7]The preprint is available at https://doi.org/10.31234/osf.io/u73sy.

results obtained from experiments are stored in a secure database and can only be accessed via e-Babylab. As e-Babylab is open-source, users are free to download, use, and modify the source code, for instance, to extend the built-in functionality, implement custom features, or even host the tool on their own local or web servers. The e-Babylab source code and user manual are available at https://github.com/lochhh/e-Babylab and https://github.com/lochhh/e-Babylab/wiki respectively.

3.1.2 Experiment Flow

Figure 3.1 shows the flow of an experiment created with e-Babylab. An experiment is accessed via a Uniform Resource Locator (URL) and begins with a welcome page, which also functions as the participant information sheet. This is followed by an automatic browser compatibility check as only Google Chrome and Mozilla Firefox for Android and desktop[8] are currently supported; these browsers make up about 82% of the Android and desktop/laptop browser market share worldwide in 2020 (NetMarketShare, 2021). In the next steps, the consent form and participant form are provided. If the experiment involves audio or video recording, a microphone and/or webcam setup step is included. Otherwise, the setup step is omitted. Here, the browser first requests the participant's permission to access their microphone and/or webcam. When access is given, a 3-second test audio (or test video) is recorded to ensure that both recording and uploading work and that the participant can be properly heard and/or seen in the recorded media. This procedure can be repeated, if necessary. Upon successful completion of this step, the participant is redirected to the start page of the experimental task, where they are prompted to enter full-screen mode to begin the task. Throughout the task, a small exit button is shown at the bottom right corner of the screen, allowing the participant to quit the experiment at any time. If the experiment is configured to allow pauses, the participant, upon clicking the exit button, will be redirected to the pause page where they are

[8]Desktop here refers to desktop and laptop computers running on Microsoft Windows, macOS, or Linux.

given the option to resume or terminate the experiment. The end page informs the participant that they have completed the experiment.

3.2 Features

3.2.1 Experiment Wizard

At the core of e-Babylab is the Experiment Wizard with which an experiment is created (see Figure 3.2). The Experiment Wizard consists of five parts: general settings, HTML templates, consent form, participant form, and crucially, the experimental task, which comprises four layers: *lists*, *outer-blocks*, *inner-blocks*, and *trials*.

3.2.1.1 General Settings

In general settings, the basic information related to an experiment (e.g., name, date and time of creation) is specified. In addition, the access settings, list selection strategy, and recording mode of an experiment are configured here. Specifically, an experiment—including its participants and results—can be made accessible to: *owner only* (private), *everyone* (all users), or *group members only* (group-based access control will be detailed in Section 3.2.7). As an experiment can have multiple lists (i.e., versions), three list selection strategies allow users to control how the lists (or versions) are distributed among participants: (a) *least played*, in which the list having the least number of participants is always selected; (b) *sequential*, in which lists are selected according to the order they are added to an experiment; and (c) *random* (i.e., random with replacement). By selecting a recording mode, an experiment can be configured to capture: (a) *key presses or clicks only*, (b) *audio and key presses or clicks*, or (c) *video and key presses or clicks*. Note that clicks may represent mouse clicks (when a mouse is used) or touches (when a touchscreen is used). These are recorded as coordinates relative to the browser window, allowing the exact locations of clicks or touches as well as the orientation of the screen to be determined. In some cases (e.g.,

Figure 3.1

Experiment Flow

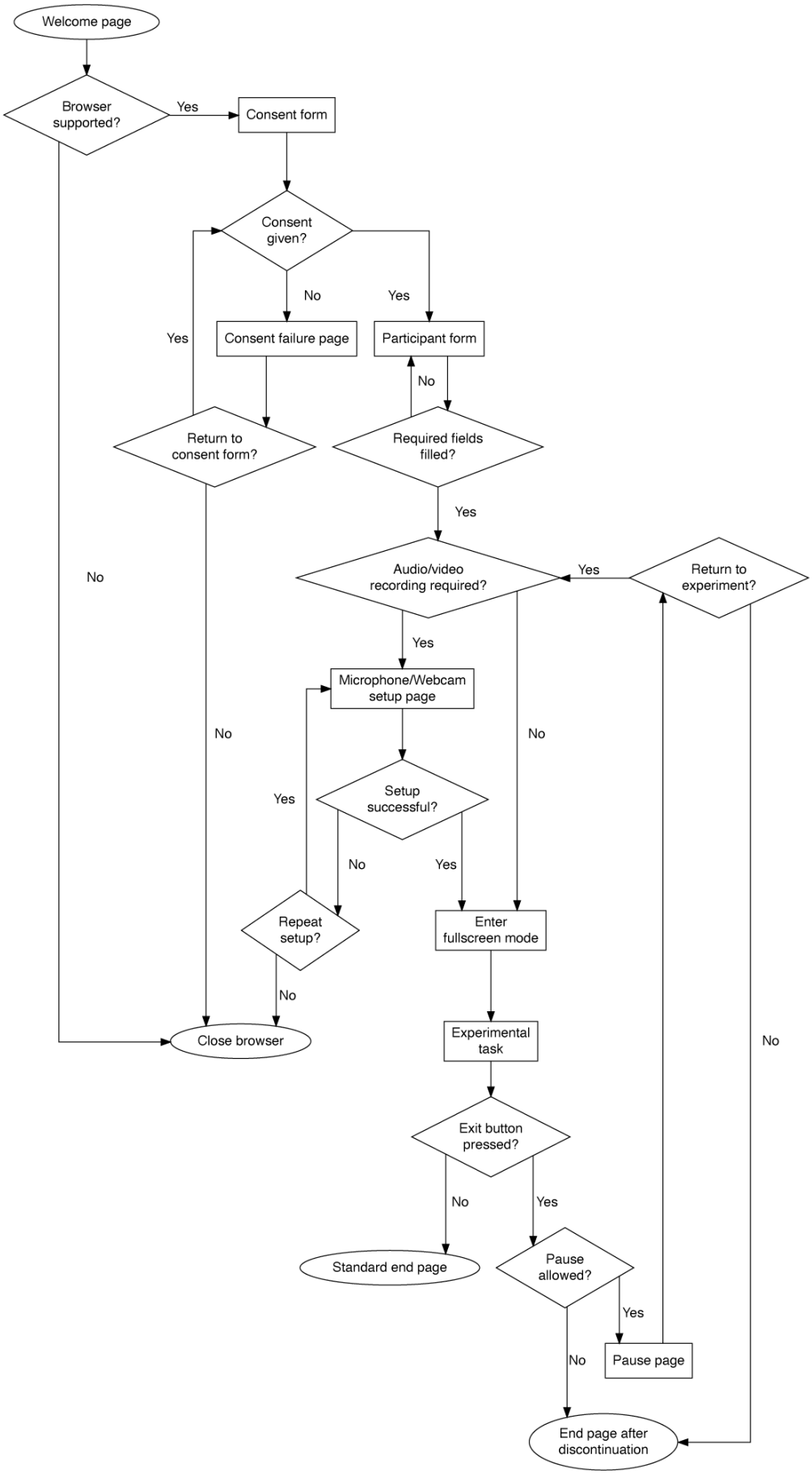
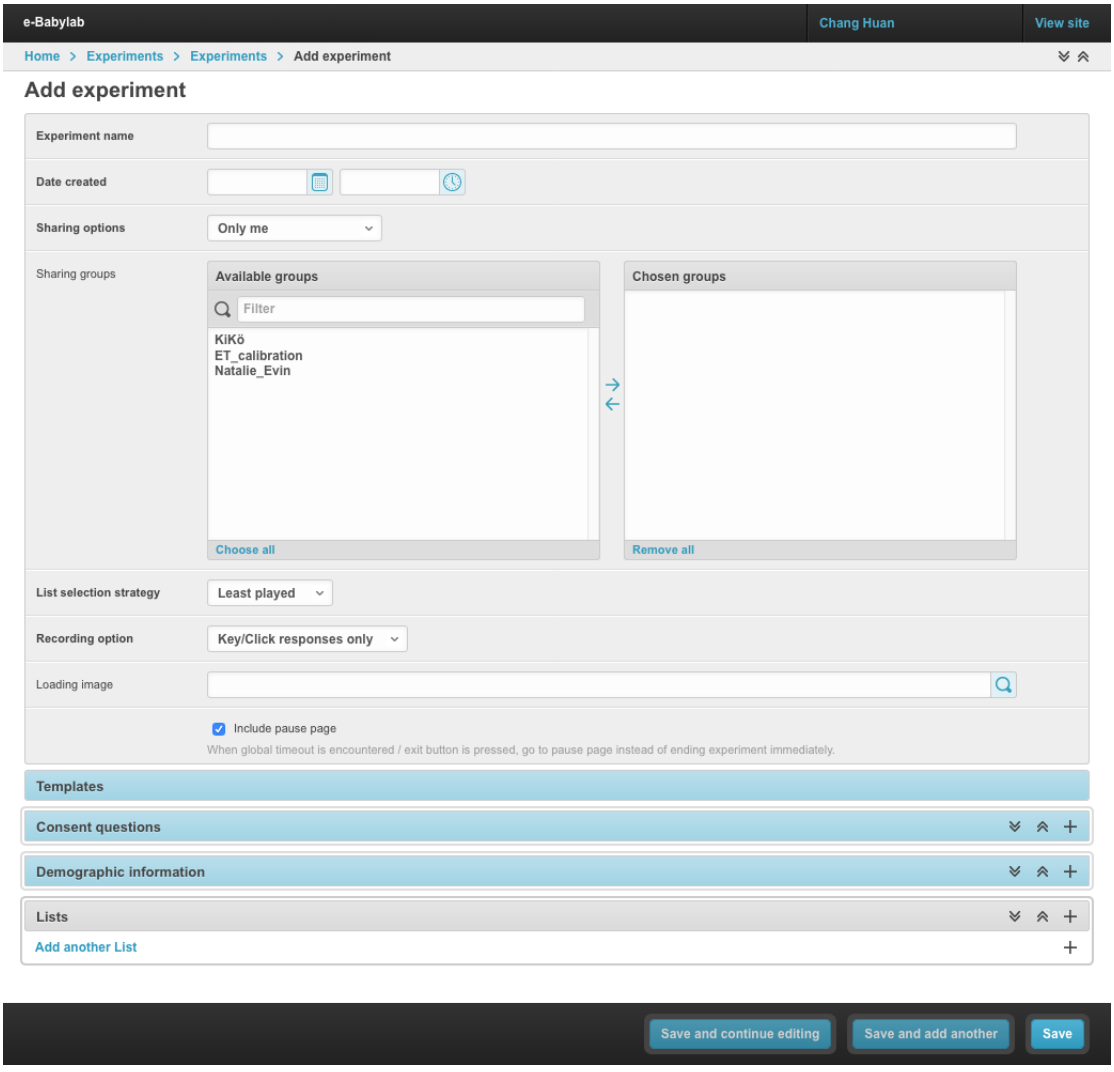


Figure 3.2

Experiment Wizard



lengthy experiments), pauses may be acceptable. For this reason, the Experiment Wizard also provides the option to include a pause page. In the event that a participant fails to complete an experiment within a given time, or when the exit button is clicked during an experiment, rather than ending the experiment immediately, the participant will be redirected to the pause page, thus giving the participant an opportunity to resume the experiment. Any "pause" events will be recorded in the results.

3.2.1.2 HTML Templates

HTML templates allow for the customisation of the looks and text (e.g., language) of all experiment webpages, including the welcome page, the consent and participant forms, the microphone and/or webcam setup pages, the experimental task page, the pause page, the error pages, as well as the end pages. A default set of HTML templates for all experiment webpages are provided for users who do not want to further customise their experiment look (see Appendix B for a sample). Alternatively, users can modify the defaults and provide their own HTML templates as well as Cascading Style Sheets (CSS) files. Customising these templates also allows for translating the entire experiment to another language.

3.2.1.3 Consent Form

This part of the Experiment Wizard allows users to specify consent questions. These will appear on the consent form as mandatory yes–no questions. Since experiments are conducted online and the experimenter may not be physically present to ensure that consent is obtained, e-Babylab automates this by checking that all consent questions are responded with "yes". In other words, a participant is only allowed to proceed with an experiment when full consent is obtained. Otherwise, the participant will be redirected to the "Failed to obtain consent" page, which by default provides an explanation as to why they are unable to proceed with the experiment as well as the option to

return to the consent form to change their responses if the responses were provided erroneously or need to be revised.

3.2.1.4 Participant Form

Different types of form fields or questions, including text fields, radio buttons, drop-down lists, checkboxes, and number fields can be included in the participant form. By setting fields as "required" or "optional", users can also control which of the form items must be answered before the form can be submitted.

3.2.1.5 Experimental Task

In general, experiments have lists, lists have outer-blocks, outer-blocks have inner-blocks, and inner-blocks have trials.

3.2.1.5.1 Lists

Lists, being the outermost layer of an experimental task, may represent different versions of the experimental task or different conditions of a between-subjects experiment. As each experiment has its own unique URL, an added benefit of having multiple lists instead of having multiple experiments is that only a single URL needs to be sent to all participants and e-Babylab automatically distributes participants across the different versions or conditions of an experiment based on the list selection strategy defined in general settings. Optionally, a list can be temporarily "deactivated", to exclude the list from being selected and distributed to future participants; this can be particularly useful when a list has had enough participants and future participants are to be distributed to other lists.

3.2.1.5.2 Outer-Blocks and Inner-Blocks

Outer-blocks and inner-blocks make up the second and third layers of an experimental task design respectively. During an experimental task, outer-blocks

are presented in a fixed order, whereas inner-blocks can either be presented in a fixed or random order, thereby increasing the flexibility in experimental task designs. For instance, when two visual stimuli are to be presented in succession within a single trial, this trial may be represented by an inner-block consisting of two trials, each presenting a visual stimulus, in either a fixed or a random order. This flexibility in presentation of stimuli in inner-blocks would not be possible without the outer–inner block structure, where stimuli can only be presented in either a fixed or a random order, but not both. Such flexibility is desirable in many experiments where introductory trials (e.g., training, familiarisation) typically precede test trials, while test trials, on the other hand, are typically randomised.

3.2.1.5.3 Trials

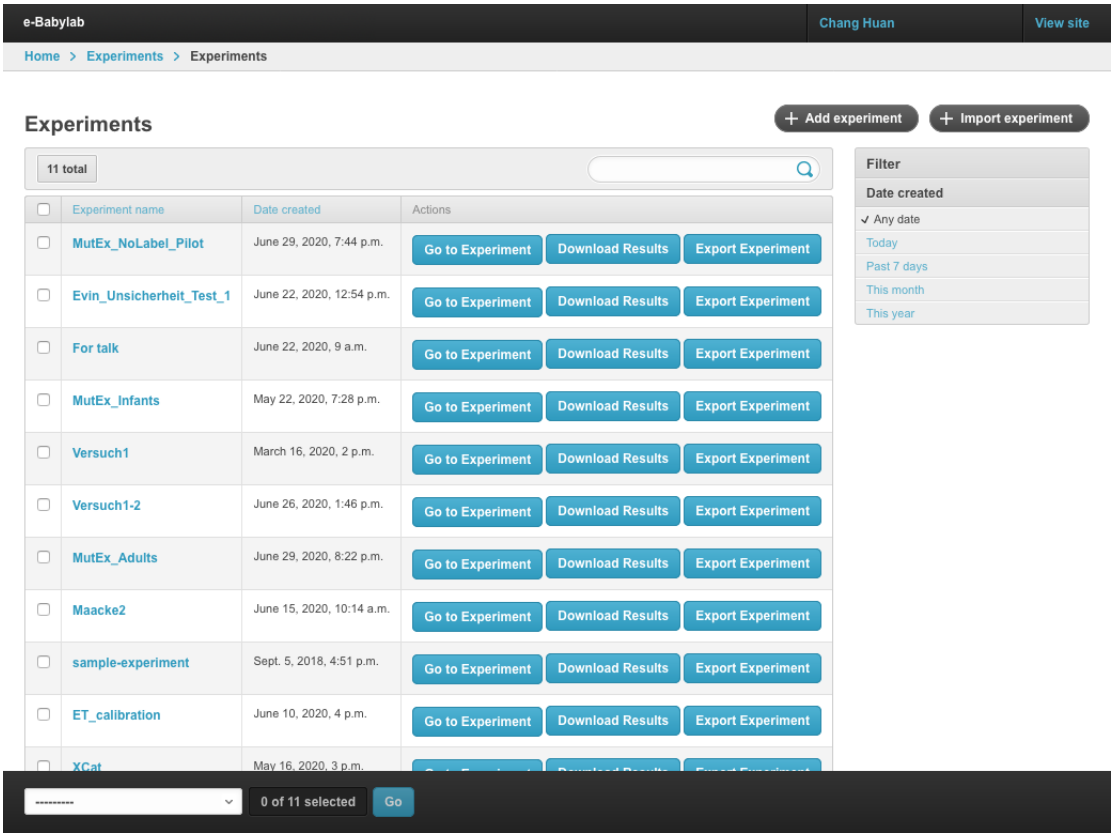
Trials are the innermost and most crucial layer of an experimental task design. As with inner-blocks, trials can either be presented in a fixed or random order. To allow a more granular control over trial setup, the specific responses that are accepted (e.g., clicks, left arrow key, space bar) as well as the maximum duration of a trial are defined on a trial-level. In addition to a visual stimulus (this can be an image or a video), an audio stimulus can also be used. Stimuli presentation can be timed by setting the visual and audio onsets (in ms). By default, these values are set to 0 so that the stimuli are presented as soon as a trial begins.

3.2.2 Experiment Management

Experiments are managed through the Experiment Administration interface (see Figure 3.3), which presents a list of experiments a user has access to. Through this interface, an experiment *setup* can be imported and exported. This enables the sharing of experiment setups, which in turn allows experiments to be reused and adapted (e.g., for replications) with minimal effort. The results of an experiment can be downloaded from here as well (detailed later in Section 3.2.4).

Figure 3.3

Experiment Administration

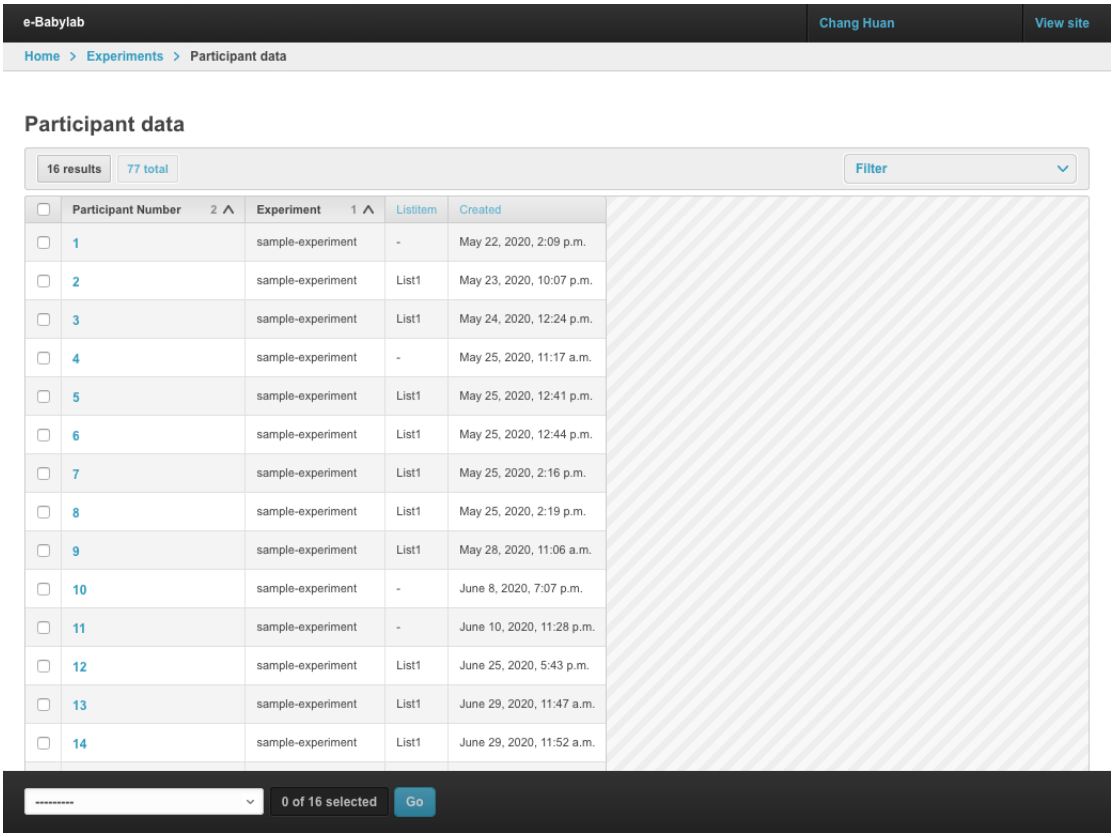


3.2.3 Participant Data Management

Participant data is managed through Participant Data Administration in which a list of participants in all experiments a user has access to is shown (see Figure 3.4). By clicking on a participant, users can view the participant's data, which includes the information provided in the participant form, their screen resolution, participant number, universally unique identifier (UUID; automatically assigned to distinguish participants from different experiments having the same participant number), participation date, experiment participated in, as well as list assigned. Deleting a participant removes all of their data and results.

Figure 3.4

Participant Data Administration



3.2.4 Results Output

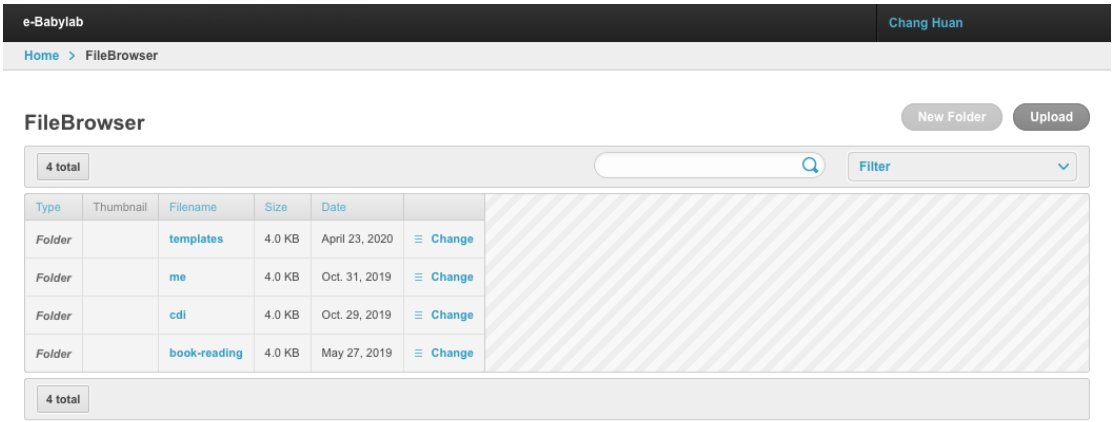
Results are downloaded as a ZIP archive containing an Excel (.xlsx) file for each participant and the media recordings (in .webm format, if any). Each Excel file contains two worksheets. The first contains the participant's information provided in the participant form, consent form responses, as well as aspect ratio and resolution of their screen. The second contains the information of all trials (e.g., stimuli presented, maximum duration allowed), the reaction times, responses given (e.g., keys pressed, mouse click coordinates), and the file names of any media recordings.

3.2.5 File Management

e-Babylab also features a file browser which allows users to create folders, upload, and manage their own experiment files, such as audio and visual stimuli, custom HTML templates and CSS files (see Figure 3.5). The supported file extensions for each of the allowed file types can be found in the user manual.

Figure 3.5

File Browser

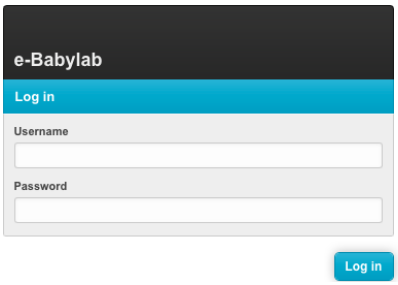


3.2.6 Authentication and Authorisation

Access to e-Babylab and its data is secured by authentication and authorisation. Essentially, authentication verifies the identity of a user and authorisation determines the operations an authenticated user can perform on a system (i.e., access rights). Figure 3.6 shows the e-Babylab login page used to authenticate users. Two types of user accounts are offered: *normal user* and *administrator*. By default, an administrator has all permissions needed to perform particular functions within e-Babylab (e.g., adding a user, changing an experiment, assigning permissions) without explicitly assigning them. A normal user, on the other hand, does not have any permissions, but instead requires permissions to be assigned by another user who has the permission to do so (e.g., an administrator).

Figure 3.6

Login Page



3.2.7 Group-Based Access Control

An experiment, including its participant data and results, can be made accessible to other users through *groups*. For instance, a group can be created for a particular research group or laboratory and an experiment can be shared among all users belonging to this group. As permissions can be assigned on a group-level, groups can also be used to more efficiently manage access rights by

assigning users to groups. In other words, a user need not be directly assigned permissions, but rather acquire them through their assigned group(s).

3.3 Technologies

This section is intended for readers interested in the technical underpinnings of e-Babylab and may require some technical knowledge. Others may decide to skip this section (without losing key information from the perspective of a user) and proceed to Section 3.4.

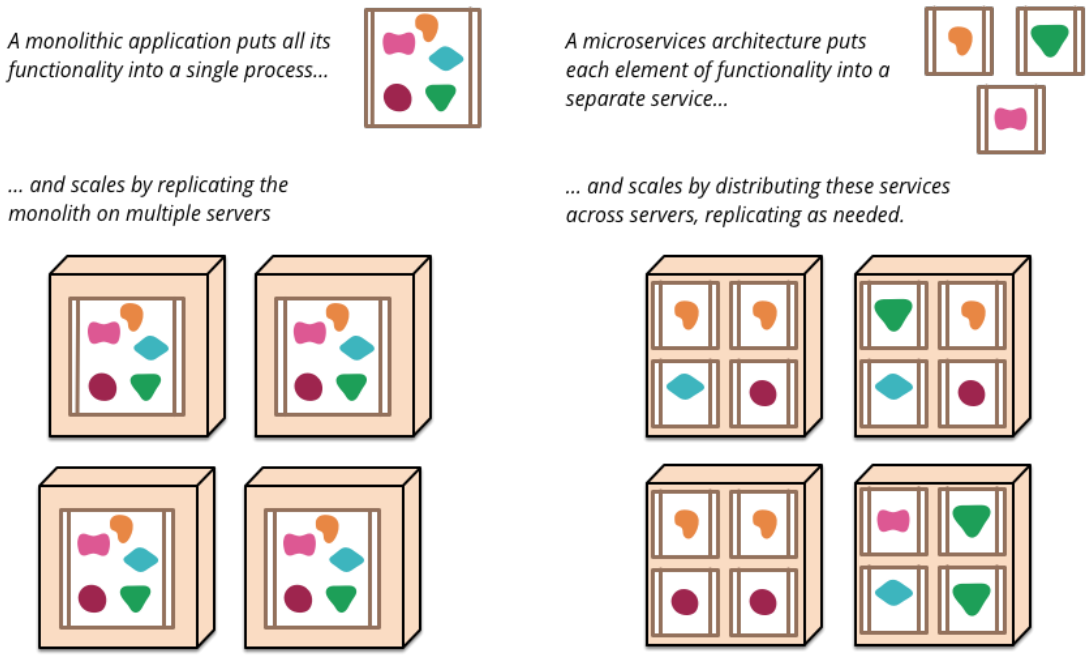
3.3.1 Microservices and Docker

e-Babylab is developed using a microservices architecture. Contrary to the commonly used monolithic architecture where all the components of an application are developed as a single entity and run in the same process, the microservices architecture centres on developing an application as a set of lightweight and loosely coupled services (or small applications), each running in its own process and serving a specific purpose (see Figure 3.7; Lewis & Fowler, 2014). As a result, services of the same application can be developed, deployed, and maintained independently—and rapidly. The independence of services also means that the failure of a single service will not affect other services (i.e., the rest of the application remains functional). Moreover, services can be reused and applied to other applications, thus reducing development costs.

Microservices lend themselves well to operating system–level virtualisation (also known as containerisation), which involves bundling the application code with all its libraries, system tools, configuration files, and dependencies (with the exception of the operating system) so that the application will always run the same, regardless of the computing environment (IBM Cloud Education, 2019). Such bundles, referred to as *containers*, are lightweight in that they share the host machine's operating system kernel, effectively eliminating the overhead of running multiple operating systems. This further translates into faster start-up

Figure 3.7

Monoliths and Microservices



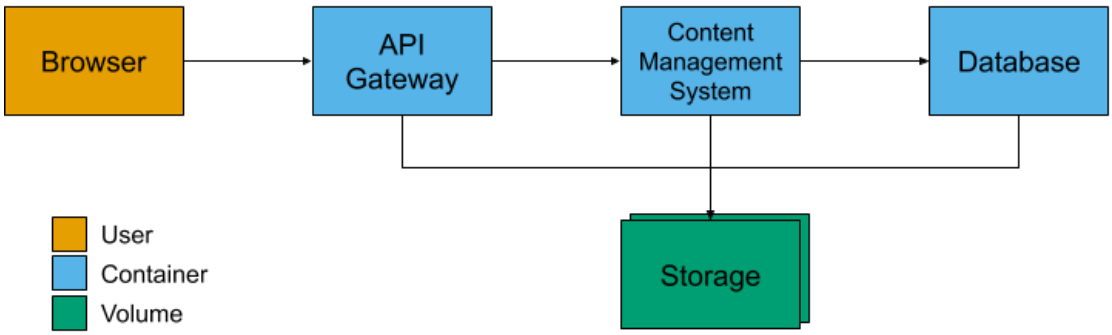
Note. From Lewis and Fowler (2014).

times and smaller memory footprints. For these reasons, Docker,[9] an open-source, lightweight container virtualisation platform that runs on Mac, Windows, and Linux is chosen to deploy the e-Babylab services.

As shown in Figure 3.8, e-Babylab is built of three services—the application programming interface (API) gateway, the content management system, and the database—each encapsulated in a container. The arrows represent dependencies between services, which are started in dependency order. In other words, the database is started before the content management system and lastly, the API gateway. As containers are ephemeral, such that they can be stopped, destroyed, rebuilt, and replaced as needed, data generated or used by containers does not persist when the containers are destroyed. Thus, data that needs to be persisted is stored in volumes managed by Docker on the host machine. In doing so, containers can easily be replaced (e.g., in upgrading a service) without any loss of data.

Figure 3.8

Components of e-Babylab



Apart from containers, the Docker architecture includes two other major components, namely *images* and *registries*. Briefly, containers are created from images which serve as blueprints. Each image is defined by a Dockerfile that contains the instructions to create a given image. During a build process, the instructions in a Dockerfile are executed and stored as an image. For ease of distribution and sharing, images can be pushed to registries where images are

[9]https://www.docker.com/

stored. The Docker-Compose file specifies whether images are to be pulled (i.e., downloaded) from a registry or built locally (using a Dockerfile). The API gateway and the database images of e-Babylab are pulled from Docker Hub (i.e., Docker's public registry) as they can be used as is. On the other hand, as the content management system is heavily customised, the image is built locally.

To orchestrate these services (i.e., to automatically configure, coordinate, and manage them) and start up e-Babylab, Docker Compose is used. By running `docker-compose up`, Docker Compose pulls the images for the API gateway and the database, builds an image for the content management system, and finally starts and runs the e-Babylab services as defined in the Docker-Compose file.

3.3.2 API Gateway

The API gateway is implemented using the open-source version of NGINX,[10] a multipurpose web server which also acts as a reverse proxy and Transport Layer Security (TLS) terminator. The API gateway acts as the entry point into e-Babylab and forwards a client's (e.g., browser) requests to the content management system and database services. With the addition of a TLS certificate, this entry point is protected by TLS, the successor to Secure Sockets Layer (SSL), which takes care of securing end-to-end communications (e.g., data transfer) between two systems. Put simply, e-Babylab is served over Hypertext Transfer Protocol Secure (HTTPS). Additionally, NGINX is configured to redirect any unsecured Hypertext Transfer Protocol (HTTP) requests to HTTPS.

3.3.3 Content Management System

3.3.3.1 Django

The content management system which provides the administrative interface to manage experiments as well as participant data and results is implemented with Django,[11] an open-source Python-based web framework. With

[10] Official NGINX image on Docker Hub: https://hub.docker.com/_/nginx
[11] Official Django image on Docker Hub: https://hub.docker.com/_/django

its aim to encourage rapid development, Django provides a complete set of ready-made components needed in most web development tasks, such as the authentication system and the dynamic administrative interface described in Section 3.2. On top of the aforementioned TLS/HTTPS protection, Django provides an extra layer of security by preventing most common security vulnerabilities in web applications, such as cross-site scripting, cross-site request forgery, Structured Query Language (SQL) injection, and clickjacking (see The OWASP Foundation, 2017, for more details on common security vulnerabilities). Thus, focus can be placed on developing the parts of a project that are unique, which in the case of e-Babylab, are the experiments as well as participant data and results.

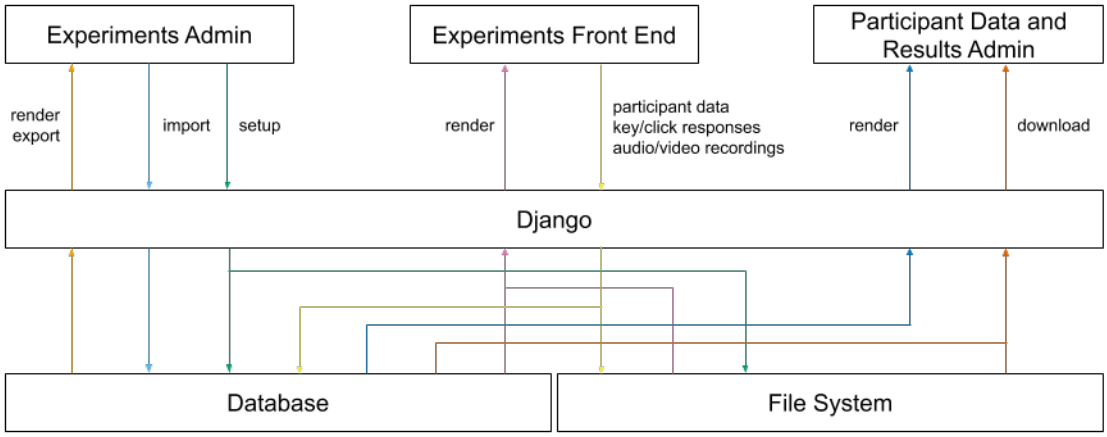
In order to generate HTML dynamically, both for e-Babylab and the experiment front end, Django's own template system, namely the Django template language is used. Typically, a template contains both static (non-editable) and dynamic (editable) parts of the desired HTML output, allowing the same design to be reused while the content changes dynamically. As shown in Figure 3.9, Django retrieves data from the database and the file system—where template files, stimuli, and media recordings are stored—and renders (i.e., interpolates) the templates with these data to dynamically display contents on the user-facing administration system and the participant-facing experiment front end. The figure also shows the flow of data in setting up, importing, and exporting experiments; in recording participant data and responses during an experiment; as well as in downloading participant data and results.

3.3.3.2 Import and Export of an Experiment Setup

The import and export functions are realised using JavaScript Object Notation (JSON), a lightweight, human-readable, text-only data interchange format used in storing and transporting data. For exporting an experiment setup, all parts of the experiment setup, from the general settings until the trials, are first retrieved from the database and serialised into JSON objects, which are

Figure 3.9

Data Flow in the Content Management System



then downloaded as a single JSON file. Likewise, for importing an experiment setup, a user-uploaded JSON file containing JSON objects making up the parts of an experiment setup is simply deserialised and a new experiment is created, set up, and stored in the database.

3.3.3.3 Media Recording

An important feature offered in experiments created with e-Babylab is the capability of recording audio and video. This is enabled by the MediaStream Recording API.[12] As the API is only available in Google Chrome and Mozilla Firefox for Android and desktop, experiments programmed with e-Babylab will not run on current iOS devices, such as iPhones and iPads. For this reason, a browser compatibility check is included as part of every experiment (as mentioned in Section 3.1.2).

Media recording involves both the front end and the back end. On the front end, the `getUserMedia()` function of the MediaDevices interface asks for permission to use the participant's media input devices (e.g., microphone and/or webcam) and produces a `MediaStream` object containing audio or video tracks, depending on the type that is requested. This `MediaStream` object is then passed to a `MediaRecorder` object which is configured to record media as 1-second chunks to be uploaded via Asynchronous JavaScript and XML (AJAX) to the Django back end. Media is recorded per trial. When the final chunk of a trial is received on the back end, individual chunks are merged as a single media file which is then stored on the file system and referenced in the database. To account for low bandwidth environments, videos are recorded in 640×480 pixels.

3.3.4 Database

The database where experiments as well as participant data and results are stored is a relational database created using the open-source relational

[12] https://developer.mozilla.org/en-US/docs/Web/API/MediaStream_Recording_API

database management system PostgreSQL.[13] In a relational database, data is stored in tabular form where rows are referred to as *records* and columns, *attributes*. Records in different tables can be linked—or related—based on a unique *key* attribute. With this key, data from multiple tables can be retrieved with a single query. For instance, downloading participant data and results of an experiment requires data to be retrieved from the participant data table, the experiments table, the lists table, the outer-blocks table, and so on. This can be easily achieved using the experiment identifier (ID) which serves as the key. In addition, as PostgreSQL is supported by Django, any changes made to the database schema, such as the addition of new tables, can simply be stored by running `python manage.py makemigrations` which automatically generates the SQL commands needed to modify the database schema. To execute these commands (i.e., to apply the changes) the `python manage.py migrate` command is used (see "Django documentation: Migrations", n.d., for more details on Django migrations).

3.4 Summary

This chapter presented e-Babylab, a new authoring tool that offers a means to easily create, host, run, and manage browser-based experiments without the need for prior technical knowledge and may be of interest to researchers looking to conduct experiments online. The technologies involved in the realisation of e-Babylab were also explained for those interested in the implementation details. In the next chapter, a series of studies designed to examine young children's word learning with tablets is presented.

[13] Official PostgreSQL image on Docker Hub: https://hub.docker.com/_/postgres

CHAPTER 4. ASSESSING EARLY WORD LEARNING WITH TABLETS

This chapter describes a series of three studies conducted to assess young children's word learning with tablets. In particular, it seeks to address research questions 1 and 2, that is, to examine whether 2- to 3-year-olds are capable of learning from interactive touchscreen media and how different experiences with screens affect their learning in a tablet-based word learning task. The first is a pilot study that aimed to test the feasibility of the study design and the instruments used. Based on the outcomes of the pilot study, modifications and improvements were made accordingly in the main studies (i.e., Study 1A and Study 1B). The main studies are available as

> Ackermann, L.[14], **Lo, C. H.**[14], Mani, N., & Mayor, J. (2020). Word learning from a tablet app: Toddlers perform better in a passive context. *PLoS ONE*, *15*(12), e0240519. https://doi.org/10.1371/journal.pone.0240519

This paper has been adapted to suit the style of this thesis.

4.1 Introduction

Within a few years of the iPad's debut, the popularity of touchscreen devices has skyrocketed. For example, American and British households with children have seen approximately a ten-fold increase in tablet ownership in the last years (American: 8% [2011] to 78% [2017]; British: 7% [2010] to 89% [2019]), with one in every two (49%) British children reported to own their own tablet in 2019 (Ofcom, 2012, 2020; Rideout, 2017). British children were also found to spend an average of 79 minutes daily using tablets (Marsh et al., 2015). In parallel with this surge in children's tablet access, there has been an explosive growth in apps with many of these targeting at young children and claiming to

[14]Both authors share co-first authorship.

be educational (Shuler, 2012). Yet, very little is understood about whether young children are capable of learning from interactive touchscreen media and how different experiences with screens affect their learning, given that young children experience difficulty in learning from traditional screen media (i.e., the "video deficit effect"; D. R. Anderson & Pempek, 2005).

As detailed in the literature review, the video deficit effect has been demonstrated in various tasks, including word learning (Krcmar et al., 2007; Roseberry et al., 2009; Troseth et al., 2018), in which children have been passively exposed to training stimuli on a screen (e.g., where they were given no choice in what they were being trained on). This video deficit effect can be mitigated by providing children with a more interactive learning context. For instance, the provision of socially contingent feedback on infants and toddlers' behaviour has been shown to improve performance in object retrieval (Troseth, 2003; Troseth et al., 2006), action imitation (Nielsen et al., 2008), and word learning tasks (Myers et al., 2017; Roseberry et al., 2014).

The review of the literature further suggested that this deficit may be mitigated with pseudo-social contingent computer interactions (e.g., Lauricella et al., 2010). However, the results on the effects of pseudo-social contingency on learning appear to be mixed across ages and the different types of contingency tested (Choi & Kirkorian, 2016; Kirkorian, Choi, et al., 2016; Russo-Johnson et al., 2017).

In addition, the above studies have focused on interactivity in a controlled context, in that children could not choose the *kind* of information to be learnt. As detailed in the literature review, self-direction (i.e., having active control over one's learning experiences) has shown to be beneficial to adults (e.g., Castro et al., 2008; Markant & Gureckis, 2014) and to children (e.g., Begus et al., 2014; Partridge et al., 2015; Ruggeri et al., 2016; Sim et al., 2015). Nevertheless, the benefit of having control over the *order* in which objects were labelled reported in Partridge et al. (2015) remains in doubt—specifically, whether the benefit occurred early in learning or whether the benefit is limited to simpler tasks—since improvement in performance was only observed in the early test

blocks which assessed fewer word–referent associations than the later test blocks. Furthermore, the participants could only determine the order in which objects were labelled but not the kind of information they could learn. In Zettersten and Saffran (2019), children preferentially make ambiguity-reducing selections when in control of their learning input.

In the present studies, young children were taught novel words in a yoked design, that is, either via active selection, where children could decide *which* objects they could hear the label for; or passive reception, where selections were made for them, based on the choices made by yoked age-matched children in the active condition. To control for overall exposure during the learning phase, the sequence, exposure time, and content of the learning phase were held constant across each yoked active–passive pair. Word learning was examined in the context of recognition tasks.

Prior to undertaking the two main studies, a pilot study was conducted using a convenience sample of children aged 18 to 48 months recruited at a family fair to test the feasibility of the study design and instruments, thereby allowing the early identification of actual and potential flaws. In the main studies, a wide age range of children across ages (24-, 30-, and 40-months) that have been targeted in previous studies (e.g., Choi & Kirkorian, 2016; Kirkorian, Choi, et al., 2016) suggesting differences in the influence of active learning on performance was tested. This allowed the developmental time course of the impact of active learning on word learning to be examined. Based on this previous work on the effects of interactivity in learning, the active condition was expected to improve performance in the younger age group, relative to the passive condition, while the opposite pattern was expected in the older age groups. Note that this prediction contrasts with findings of an active advantage in Partridge et al. (2015) as learning was examined in older children ($M_{\text{age}} = 47$ months, range: 36–59 months).

4.2 Pilot Study

4.2.1 Method

4.2.1.1 Participants and Design

The pilot sample consisted of 17 typically developing, primarily monolingual German-speaking children with ages ranging from 18 to 48 months, recruited at Lokolino, an annual family fair held on 4–5 February 2017 in Göttingen, Germany (see Table 4.1 for the distribution of participants by age group and condition). The study took place in the WortSchatzInsel Göttingen laboratory. Participants were paired according to their age groups and were assigned to either the active or the passive condition. In the active condition, participants could select four novel objects to be told the label of, while in the passive condition, participants were automatically given the labels for the objects chosen by their yoked active peers. Thirteen additional participants were excluded from the analysis for the following reasons: (a) failing to complete the study ($n = 4$), (b) showing a clear side preference in selection (i.e., tapping seven times consecutively on the image shown on a particular side; $n = 5$), (c) providing incorrect responses in all familiar trials ($n = 3$), and (d) being incorrectly assigned to an active peer from a different age group ($n = 1$). The study was reviewed and approved by the ethics committee of the Georg Elias Müller Institute of Psychology, University of Göttingen. Caregivers gave written consent to their child's participation in the study.

4.2.1.2 Apparatus and Materials

The study was carried out using an iPad Pro with a web application developed based on the framework provided in Frank et al. (2016). Images of eight novel objects and four familiar objects were chosen for the study (see Figure 4.1 and Figure 4.2). Vocabulary development norms suggest that over 75% of all 24-month-olds and close to 100% of all 30-month-olds already produce the four familiar words: "Apfel" [apple], "Auto" [car], "Baby" [baby], and "Ball"

Table 4.1

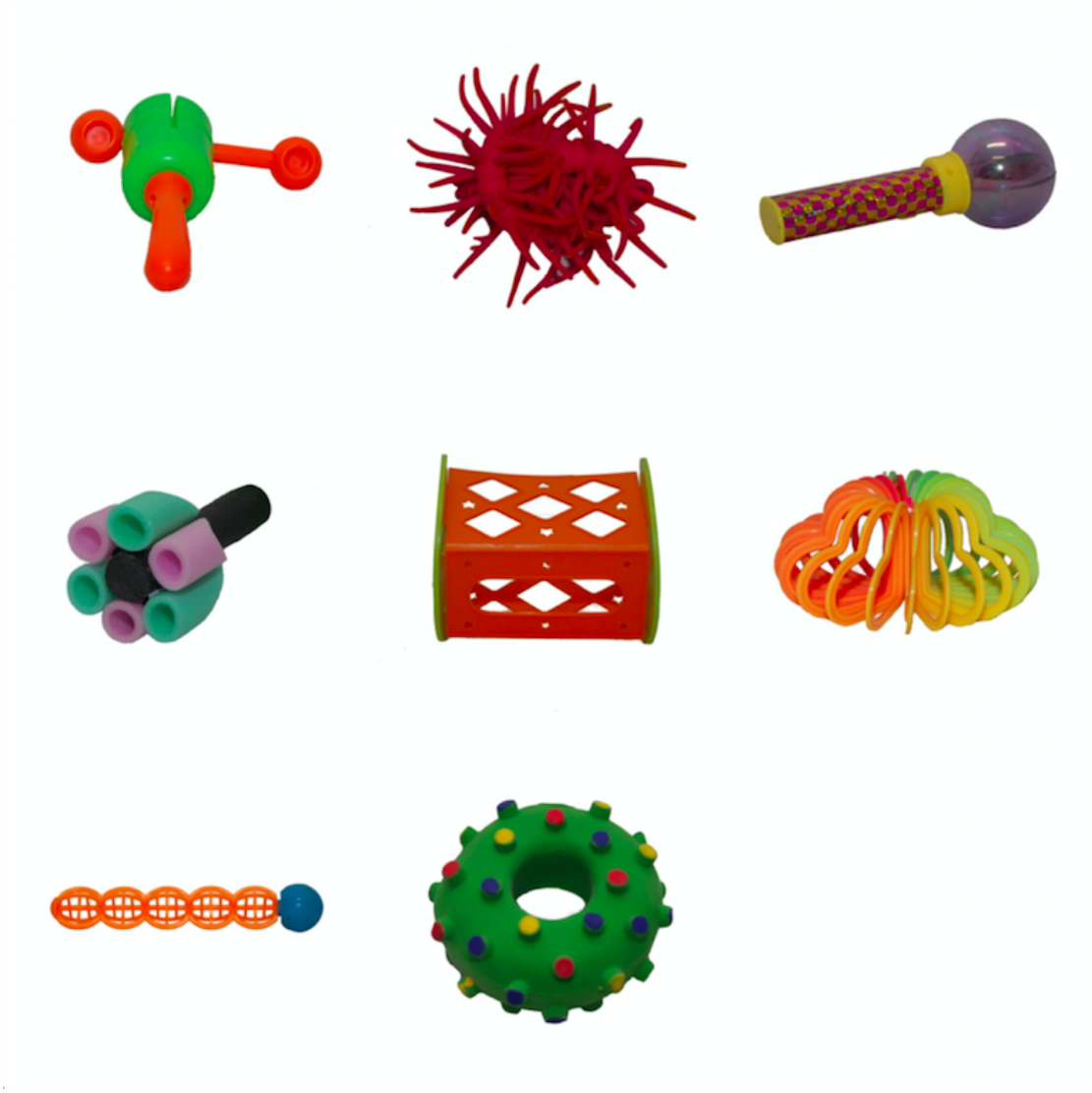
Distribution of Participants by Age Group and Condition

Age group (months)	Active	Passive
18 – <24	1	1
24 – <30	1	0
30 – <36	1	0
36 – <42	4	4
42 – <48	4	1
Total	11	6

[ball] (Braginsky, 2018; Szagun et al., 2014). Four disyllabic, novel words were selected as labels for the chosen novel objects: “Batscha”, “Foma”, “Kolat”, and “Widex”. To prevent disambiguation of novel words based on the use of determiners (e.g., “der”, “die”, “das”), the neuter article “das” was used with all novel words. These words obey the phonotactic constraints of German (see Appendix C for further details). All auditory stimuli used were recorded by a female native speaker of German in child-directed speech.

Figure 4.1

Novel Objects



Note. From Horst and Hout (2016).

Figure 4.2

Familiar Objects

4.2.1.3 Procedure

The study began with a learning phase, followed by a test phase.

4.2.1.3.1 Learning Phase

Active Condition The learning phase consisted of four trials and each trial began with a prompt asking the participant to select one of the two randomly combined images of the novel objects placed on the left and right sides of the screen respectively. In the first trial, the prompt was "Guck mal, hier sind zwei Bilder. Du kannst auf eines drücken." [Look, here are two pictures. You can tap on one.] In subsequent trials, the prompt was "Drück mal auf ein Ding, dann hörst du seinen Namen." [Tap on an object, then you'll hear its name.] Tapping was only enabled 300 ms after the prompt had ended to ensure that the tap could reliably be interpreted as a response to the presentation of stimuli. Upon tapping, a red outline was shown around the selected image while that which was not selected was hidden. The selected novel object was then labelled five times in the same trial using various carrier phrases, including: (a) "Guck mal, ein X!" [Look, a/an X!], (b) "Das ist ein X!" [This is a/an X!], (c) "Wow, da ist ein X!" [Wow, there is a/an X!], (d) "Siehst du das X?" [Do you see the X?], and (e) "Toll! Das ist ein X!" [Great! This is a/an X!], where X was the novel word. The time taken by the participant to make their selection was automatically recorded to be used to time stimuli presentation for the passive learning peer so that both participants saw the images for exactly the same amount of time. The subsequent trial began 1500 ms after the labelling had ended. In each trial, the pairs of novel objects displayed and the novel word given to each selected object were generated at random with no repeats. Thus, at the end of the learning phase, the participant was presented with four distinct novel labels for their chosen four novel objects.

Passive Condition A passive learning participant was not required to do anything but to watch and listen as they would be exposed to their active learning, age-matched peer's selections according to the exact timings of the

active peer. Throughout the learning phase, tapping was disabled. Instead of being prompted to select something, an introductory audio "Siehst du die zwei Bilder? Sind sie schön?" [Do you see the two pictures? Are they beautiful?] was played to attract the participant's attention to the images. The participant had to wait for as long as their active peer took to select between the two novel objects displayed before the selected object was outlined in red and the unselected object was hidden. As with the active peer, the participant heard the novel object labelled with the novel word five times. A 1500 ms pause followed before the subsequent trial began. The order of the learning trials was identical to that which had been given to the active peer.

4.2.1.3.2 Test Phase (All Participants)

The test phase consisted of 14 two-alternative forced choice (2AFC) trials, of which two were familiar trials to keep the participant engaged and 12 were test trials to assess the participant's recognition of the novel word–referent associations. The trials were ordered as follows: one familiar trial, six test trials, one familiar trial, and six test trials. In each familiar trial, the participant was presented with a pair of randomly generated familiar objects, whereas in the test trials, each novel word was tested three times (by pairing the target object separately with each of the three other chosen objects as distractors), in counterbalanced order. Upon presentation of the pair of objects (also placed on the left and right sides of the screen respectively), the participant was asked to tap on the object associated with the heard target word X embedded in the carrier phrase "Drück mal auf das/den X." [Tap on the X.] Tapping was disabled until the onset of the target word X in the carrier phrase. There was no time limit for the participant to respond and no feedback was given after the participant had responded, regardless of which object they tapped on. The participant's response and reaction time (RT) were then recorded. A 1500 ms pause followed before the subsequent trial began.

4.2.2 Results

4.2.2.1 Reaction Time

RT was measured in ms from the onset of the target word. As the data did not follow a normal distribution as indicated by a Shapiro-Wilk normality test ($W = 0.547, p < .001$), the data was log transformed prior to further analysis to approximate a normal distribution. To ensure that only those trials where the child was engaged in the task were included in the analysis, outliers were removed using a criterion of 2 *SD*s above the mean (4 active, 7 passive). Mean and standard deviation of RT for each condition, before and after outlier removal are detailed in Table 4.2.

Table 4.2

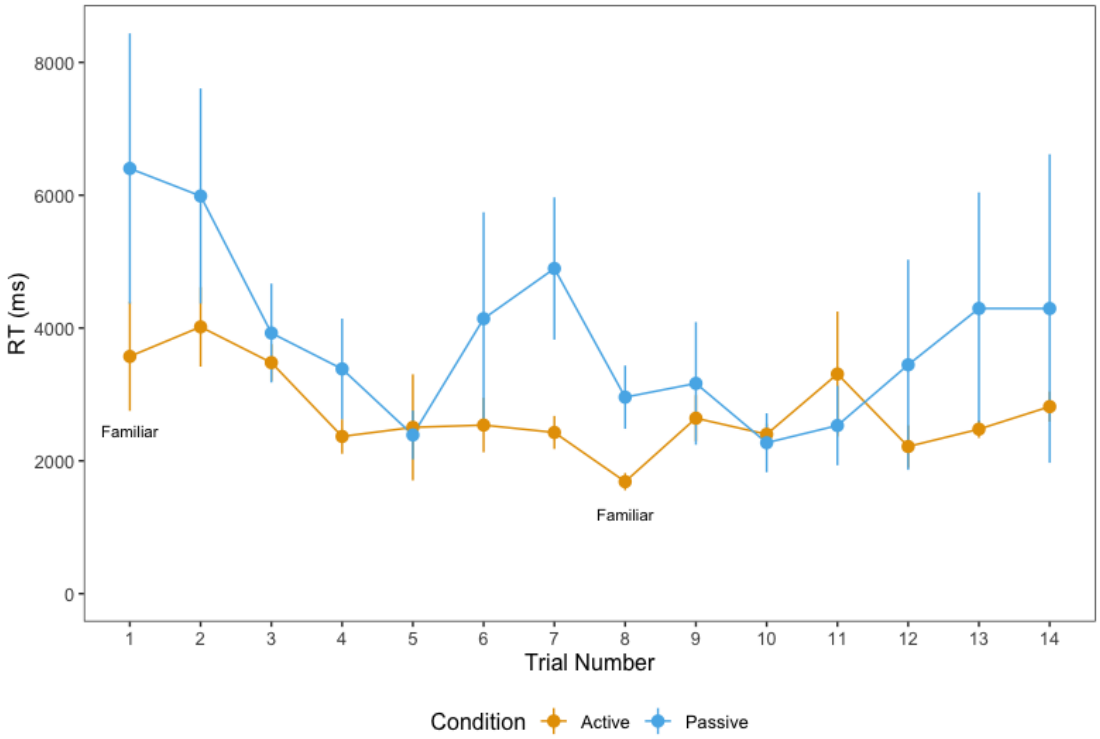
Mean and Standard Deviation of RT Before (Unadjusted) and After (Adjusted) Outlier Removal, Split by Condition

Condition	Unadjusted M_{RT} (s)	Unadjusted SD_{RT} (s)	Adjusted M_{RT} (s)	Adjusted SD_{RT} (s)
Active	2.876	2.078	2.621	1.359
Passive	5.014	5.585	3.660	2.370

Figure 4.3 and Figure 4.4 show children's trial-by-trial RT and children's RT split by trial type (i.e., familiar and test trials) respectively. Results from a Welch's *t*-test indicated that children assigned the active condition ($M_{\text{logRT}} = 7.707, SD_{\text{logRT}} = 0.649$) did not differ significantly from their passive peers ($M_{\text{logRT}} = 8.159, SD_{\text{logRT}} = 0.584$) in terms of their speed in responding in the familiar trials; $t(17.388) = -1.839, p = .083$. Likewise, in the test trials, there was no statistically significant difference between the active ($M_{\text{logRT}} = 7.836, SD_{\text{logRT}} = 0.427$) and the passive ($M_{\text{logRT}} = 8.002, SD_{\text{logRT}} = 0.640$) groups; $t(69.853) = -1.580, p = .119$.

Figure 4.3

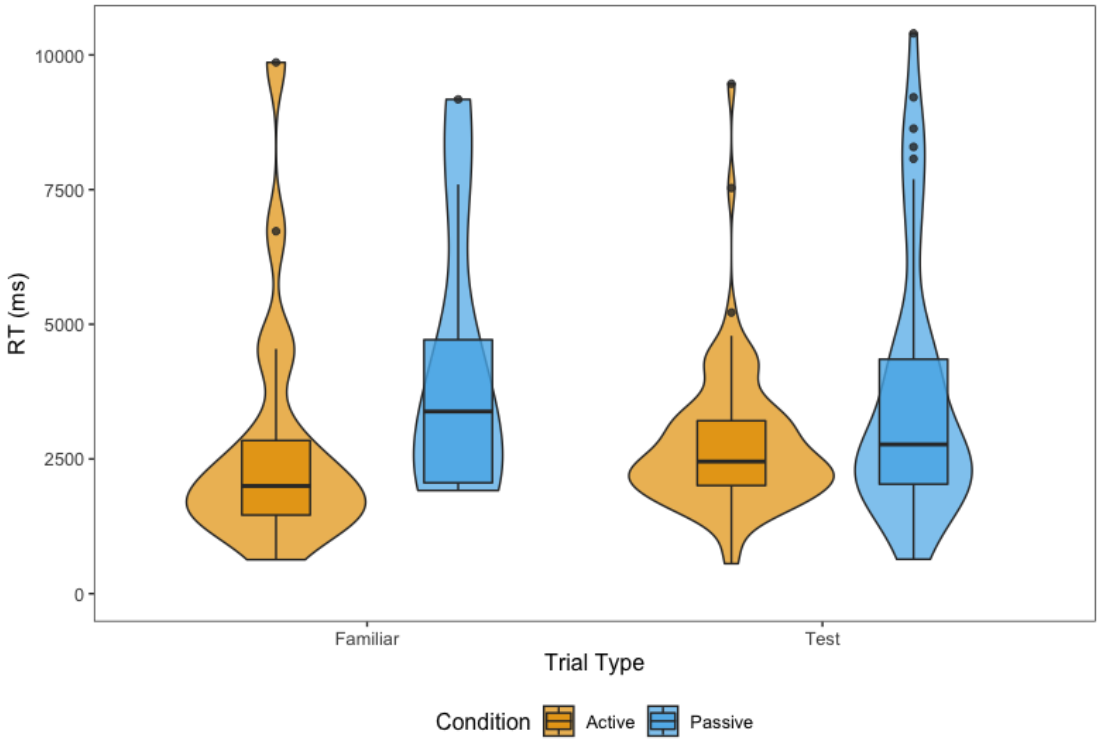
RT by Trial Number



Note. Only trials in which children responded correctly are considered.

Figure 4.4

RT by Trial Type



Note. Only trials in which children responded correctly are considered.

On the other hand, visual inspection of the trial-by-trial RT data (Figure 4.3) suggests the presence of a difference in RT between both conditions early in the test phase. Thus, in an exploratory analysis, the data was split into "early" (first seven) and "late" (last seven) trials. Indeed, Welch's t-tests revealed that children in the active condition were significantly quicker to tap on the target object ($M_{\text{logRT}} = 7.881, SD_{\text{logRT}} = 0.568$) than their passive peers ($M_{\text{logRT}} = 8.250, SD_{\text{logRT}} = 0.553$) in the early trials; $t(47.288) = -2.640, p = .011$, but not in later trials where the passive group ($M_{\text{logRT}} = 7.860, SD_{\text{logRT}} = 0.640$) caught up with the active group ($M_{\text{logRT}} = 7.749, SD_{\text{logRT}} = 0.367$); $t(43.492) = -0.894, p = .376$. This is concurrent with the experimenter's observation that a number of children in the passive condition did not know what to do at the beginning of the test phase and needed help from the experimenter to proceed with the task.

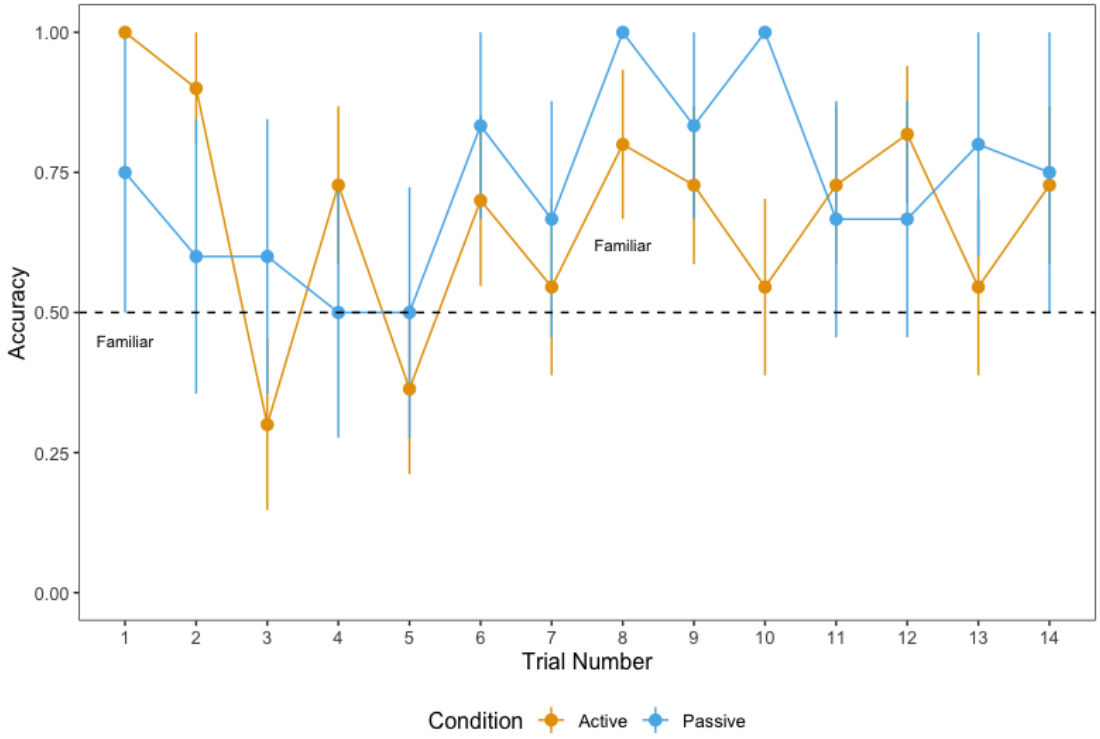
4.2.2.2 Accuracy

Figure 4.5 and Figure 4.6 show children's accuracy in identifying the labelled object in each trial of the test phase and children's accuracy split by trial type (i.e., familiar and test trials) respectively. Results from a Welch's t-test indicated that children assigned the active condition ($M = 0.905, SD = 0.301$) did not differ significantly from their passive peers ($M = 0.900, SD = 0.316$) in terms of their accuracy in the familiar trials; $t(17.005) = 0.040, p = .969$. Likewise, children's performance in the test trials did not differ significantly between the active ($M = 0.636, SD = 0.483$) and the passive ($M = 0.701, SD = 0.461$) conditions; $t(139.370) = -0.933, p = .353$.

To determine whether children were responding above chance across the test trials, one-sample t-tests were also conducted by comparing the accuracies of the active and the passive groups to 0.50. Results indicated that both groups were responding above chance; $t(66) = 3.189, p < .001$ (active); $t(66) = 3.577, p < .001$ (passive).

Figure 4.5

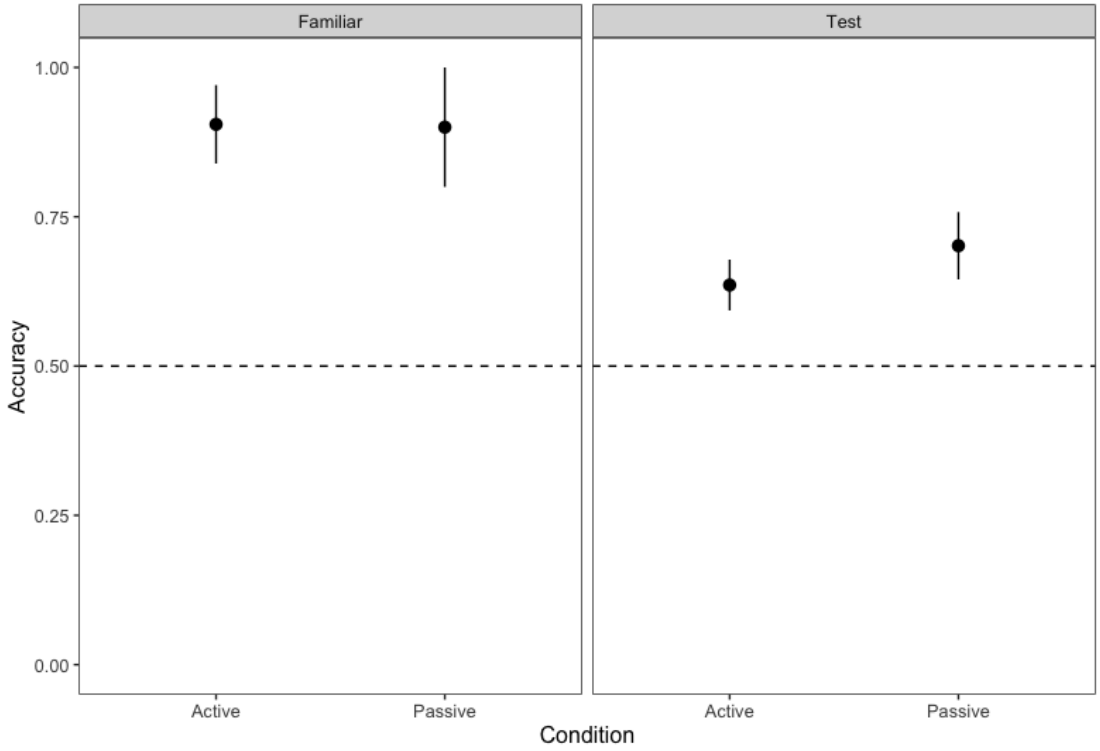
Accuracy by Trial Number



Note. Dashed line represents chance (.50).

Figure 4.6

Accuracy by Condition and Trial Type



Note. Dashed line represents chance (.50).

4.2.3 Discussion

The primary aim of the pilot study was to conduct a preliminary investigation of the feasibility of the study design and the web application used to assess young children's word learning with tablets. Overall, all children performed above chance in the novel word recognition tasks and no differences were found both in terms of RT and accuracy between children in the active condition and children in the passive condition. These results are, however, not taken as evidence of any impacts of active or passive learning on children's word learning. Instead, they provide initial support for the usability of a web application for collecting data in the tablet-based paradigm employed here. Anecdotally, it has also been observed that most children found the task interesting and some even repeated the novel words while listening to the objects being labelled during the learning phase.

Crucially, conducting the pilot study has allowed the early identification of a potential confound in the study design. Specifically, children in the active condition were quicker to tap on the target objects relative to their passive, age-matched peers during the first half of the test phase but not during the second half. This difference, concurrent with observations that children in the passive group needed guidance in the first test trials, likely arose because children in the active condition had prior experience in tapping on objects during the learning phase, while the familiarisation phase was the first point in the study where children in the passive condition were asked to tap on the screen. Thus, to deal with this potential confound, a familiarisation phase consisting of six trials was included prior to the test phase in Study 1A and Study 1B. Given children's good performance in the 2AFC test trials, a four-alternative forced choice (4AFC) test phase was added to provide a more stringent test of word learning as well.

4.3 Study 1A

4.3.1 Method

4.3.1.1 Participants and Design

A total of 130 typically developing, primarily monolingual German-speaking children were recruited from a research participant database administered by the WortSchatzInsel Göttingen laboratory, with 42 participants in the 24-month age group and 44 participants in each of the 30- and 40-month age groups. Mean age, age range, and standard deviation for each age group are detailed in Table 4.3. The study took place in the laboratory. Yoked age-matched pairs of participants (ages at date of testing within 2 months of each other) were assigned to either the active or the passive condition. As in the pilot study, participants assigned to the active condition could select four novel objects to be told the labels of, while those assigned to the passive condition were automatically given the labels for the objects chosen by their yoked active peers. An additional pair of participants in the 24-month age group had to be excluded due to missing data and an additional two pairs in the 30-month age group had to be excluded for showing a clear side preference in selection (i.e., tapping eight times consecutively on the image shown on a particular side) and inattentiveness (i.e., getting up and walking around during the study). The study was reviewed and approved by the ethics committee of the Georg Elias Müller Institute of Psychology, University of Göttingen. Caregivers gave written consent to their child's participation in the study.

Table 4.3

Age Mean, Standard Deviation, and Range

Age group	n	M_{age} (months)	SD_{age} (months)	$\text{Range}_{\text{age}}$ (months)
24-month	42	24.31	1.16	22.05–25.96
30-month	44	29.81	1.49	28.16–35.22
40-month	44	39.69	3.52	36.01–47.97

4.3.1.2 Apparatus and Materials

The study was carried out using an iPad Pro with a modified version of the web application used in the pilot study. The same novel words (i.e., "Batscha", "Foma", "Kolat", and "Widex"), images of novel and familiar objects as well as auditory stimuli used in the pilot study were used in the present study. In addition to the four familiar objects, two more familiar objects were included (see Figure 4.7). Vocabulary development norms suggest that over 70% of all 24-month-olds and close to 100% of all 30-month-olds already produce the six familiar words: "Apfel" [apple], "Auto" [car], "Baby" [baby], "Ball" [ball], "Baum" [tree], and "Schuh" [shoe] (Braginsky, 2018; Szagun et al., 2014).

Figure 4.7

Familiar Objects

4.3.1.3 Procedure

Based on results from the pilot study, two procedural changes were made, including the addition of a familiarisation phase and a 4AFC test phase. The study began with the learning phase, followed by the familiarisation phase, the 2AFC test phase, and finally the 4AFC test phase.

4.3.1.3.1 Learning Phase

The learning phase was the same for participants in both conditions as in the pilot study. The active participants were to select four novel objects and heard the objects labelled five times each, while the passive participants were given their active peers' selections according to the exact timings and order, and heard the objects labelled in the same manner as the active condition, which repeated the novel word five times.

4.3.1.3.2 Familiarisation Phase

Instead of having a single familiar trial precede the test phase and another in the middle of the test phase (as was the case in the pilot study), a familiarisation phase consisting of six familiar trials was included following the learning phase to: (a) familiarise the passive group with tapping and (b) keep all participants engaged. Trials were presented in the same manner as the familiar trials in the pilot study, where participants were presented with a pair of randomly generated familiar objects and instructed to tap on one of these objects based on a given label X embedded in the carrier phrase "Drück mal auf das/den X." [Tap on the X.] There was no time limit for the participant to respond and no feedback was given after the participant had responded, regardless of which object they tapped on. The participant's response and RT were then recorded. A 1500 ms pause followed before the subsequent trial began.

4.3.1.3.3 2AFC Test Phase

This phase consisted of the same 12 2AFC test trials in the pilot study, where each novel word was tested three times (paired separately with each of the three other chosen objects), in counterbalanced order. As with the familiar trials, there was no time limit for responding and no feedback was given after the participant had responded. The participant's response and RT were also recorded and a 1500 ms pause followed before the subsequent trial began.

4.3.1.3.4 4AFC Test Phase

This phase consisted of 8 4AFC trials where each novel word was tested twice, in counterbalanced order. In each trial, participants were shown all four novel objects which they had learnt labels for and asked to tap on the object associated with the heard novel word. The images of the novel objects were positioned randomly in a 2×2 grid on the screen. There was also no time limit for responding and no feedback was given. Again, the participant's response and RT were recorded and a 1500 ms pause followed before the subsequent trial began.

4.3.2 Results

4.3.2.1 Reaction Time

RT was measured in ms from the onset of the target word. As participants were not given a time limit to respond, RTs included outliers as high as 1015 s (in one case where the participant got up and played with something else, before returning to make their selection and continue with the task). Since the data did not follow a normal distribution as indicated by a Shapiro-Wilk normality test ($W = 0.082, p < .001$), the data was log transformed prior to further analysis to approximate a normal distribution. To ensure that only those trials where the child was engaged in the task were included in the analysis, outliers were removed using a criterion of 2 *SD*s above the mean. The number of outliers decreased with increasing age with roughly equal number of

outliers in each condition (35 active, 37 passive in the 24-month age group; 20 active, 22 passive in the 30-month age group; 14 active, 17 passive in the 40-month age group). Mean and standard deviation of RT for each age group and condition, before and after outlier removal are detailed in Table 4.4.

Table 4.4

Mean and Standard Deviation of RT Before (Unadjusted) and After (Adjusted) Outlier Removal, Split by Age Group and Condition

Age group	Condition	Unadjusted M_{RT} (s)	Unadjusted SD_{RT} (s)	Adjusted M_{RT} (s)	Adjusted SD_{RT} (s)
24-month	Active	4.856	7.519	3.352	2.742
	Passive	6.955	44.068	3.428	2.826
30-month	Active	4.570	10.302	3.262	2.370
	Passive	4.398	7.040	3.354	2.304
40-month	Active	3.264	2.850	2.985	2.051
	Passive	3.302	3.869	2.744	1.819

Figure 4.8 shows children's trial-by-trial RT across the familiarisation phase as well as the 2AFC and 4AFC test phases, whereas Figure 4.9 shows the distribution of children's RT in each of the three phases, split by age group and condition. To assess whether RTs differed across conditions (active vs. passive) in each of the three phases, linear mixed-effects models (LMMs) were fitted and analysed using the `mixed()` function from the *afex* package (Singmann et al., 2020), which relies on the *lme4* package (D. Bates et al., 2015) for model fitting. The models included condition (active, passive), age group (24-month, 30-month, 40-month), and the interaction between condition and age group as fixed effects. Both condition (-1: passive; 1: active) and age group (-1: 24-month; 1: 30-month, 40-month) were sum-coded. Additionally, to determine models with a parsimonious random effect structure (Matuschek et al., 2017), the forward "best-path" approach (D. J. Barr et al., 2013) was used to test random slopes for inclusion ($\alpha = 0.20$). The resulting models therefore also included selected object

and participant pair as random intercepts, with by-participant-pair adjustment to the slope of condition:

$$\text{RT}_{\text{log}} \sim \text{Condition} * \text{Age group} + (1 + \text{Condition}|\text{Participant pair}) + (1|\text{Object})$$

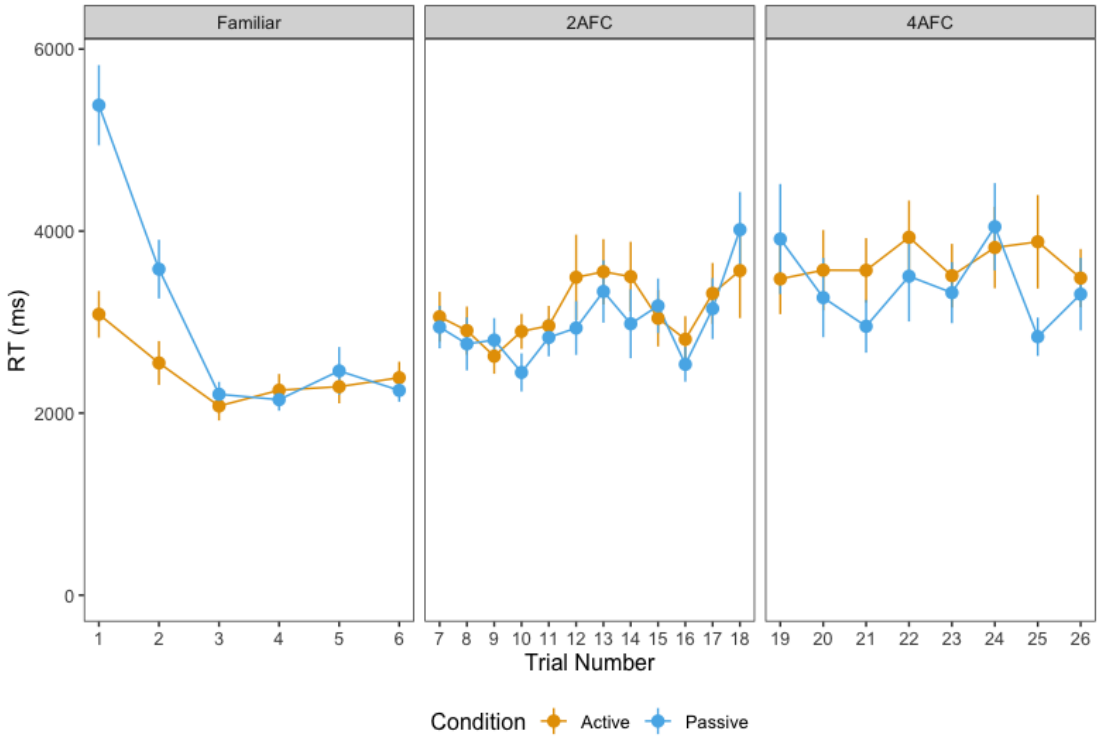
The results of the models are detailed in Tables 4.5, 4.6, and 4.7, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons, with p-values corrected using the Tukey method, were conducted using the `pairs()` function in the *emmeans* package (Lenth, 2020).

As shown in Table 4.5, there was a significant main effect of condition in the familiar trials, with children in the active condition being quicker to tap on the target object relative to children in the passive condition, potentially due to the latter being required to tap on the screen for the first time in these trials (see Figure 4.8). The analysis also revealed a significant main effect of age group as well as a significant interaction between condition and age group. Results from the follow-up tests indicated that 24-month-olds were significantly slower than 40-month-olds ($\beta = 0.307, SE = 0.064, t = 4.823, p < .001$) to tap on the target object, but not 30-month-olds ($\beta = 0.127, SE = 0.065, t = 1.963, p = .129$). Compared to 40-month-olds, 30-month-olds were also significantly slower ($\beta = 0.180, SE = 0.062, t = 2.932, p = .013$) in the familiarisation phase. The simple main effect of condition (active vs. passive) was significant in both the 24-month age group ($\beta = -0.338, SE = 0.109, t = -3.110, p = .003$) and the 30-month age group ($\beta = -0.227, SE = 0.103, t = -2.208, p = .031$), with children in the active condition being quicker than children in the passive condition in responding. No such effect was found in the 40-month age group ($\beta = 0.035, SE = 0.101, t = 0.351, p = .727$).

As indicated in Table 4.6 and Table 4.7, no significant main effects of condition and age group were found in the 2AFC and 4AFC test phases. No significant interaction between condition and age group was found either.

Figure 4.8

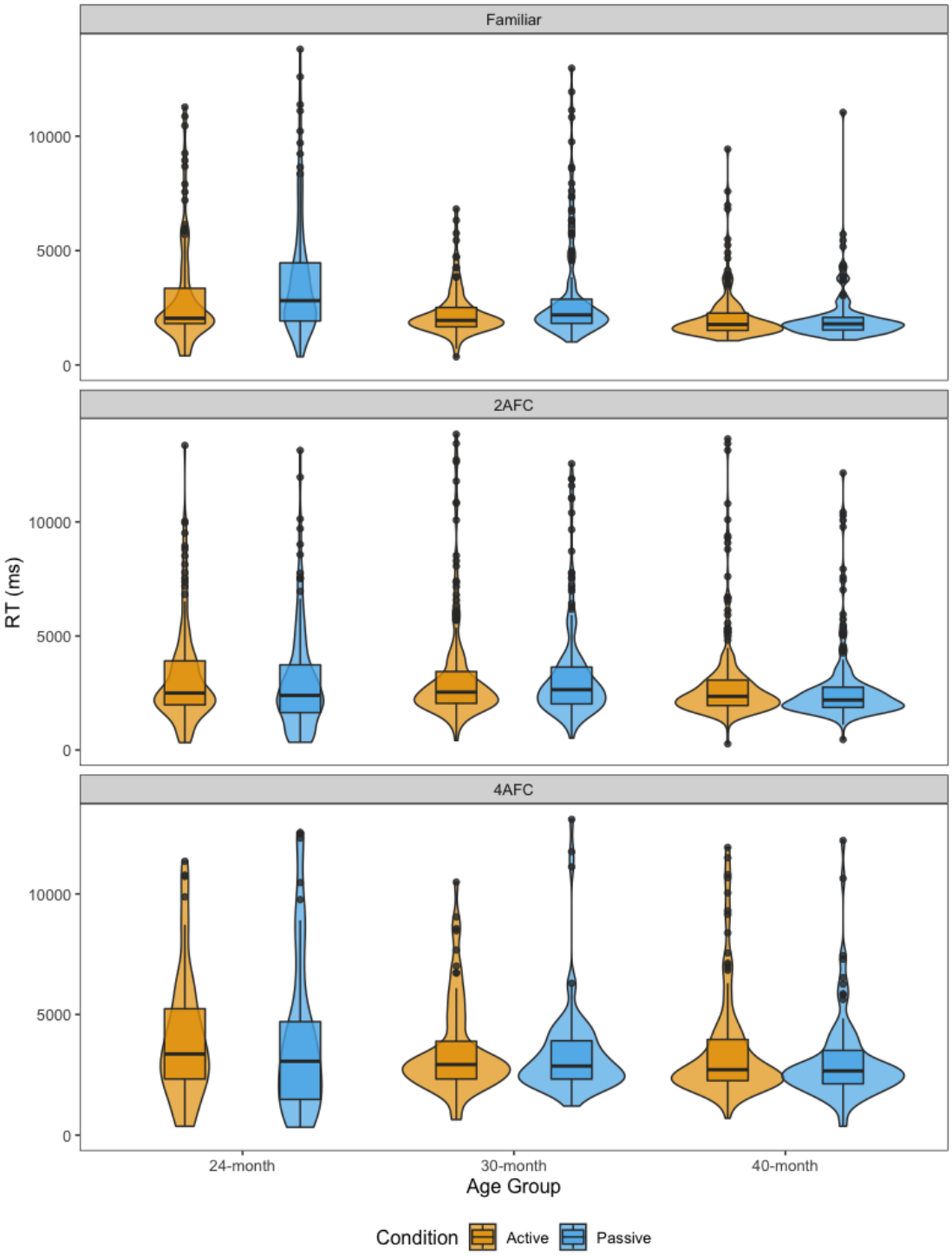
RT by Trial Number



Note. Only trials in which children responded correctly are considered. By-age plots can be found in Appendix D.

Figure 4.9

RT by Phase and Age Group



Note. Only trials in which children responded correctly are considered.

Table 4.5

LMM Results for RT in the Familiarisation Phase

	Model summary			Model comparison		
	β	*SE*	*t*	χ^2	*df*	*p*
Intercept	7.730	0.025	308.233	65.432	1	<.001***
Condition	-0.088	0.029	-3.027	8.719	1	.003**
Age group				20.298	2	<.001***
30-month	0.018	0.035	0.507			
40-month	-0.162	0.035	-4.680			
Condition:Age group				7.034	2	.030*
Condition:30-month	-0.025	0.041	-0.616			
Condition:40-month	0.106	0.041	2.612			

*$p < .05$, **$p < .01$, ***$p < .001$

Table 4.6

LMM Results for RT in the 2AFC Test Phase

	Model summary			Model comparison		
	β	*SE*	*t*	χ^2	*df*	*p*
Intercept	7.849	0.027	291.322	83.419	1	<.001***
Condition	0.042	0.032	1.339	1.768	1	.184
Age group				5.796	2	.055
30-month	0.090	0.038	2.385			
40-month	-0.018	0.037	-0.482			
Condition:Age group				1.408	2	.495
Condition:30-month	-0.048	0.044	-1.077			
Condition:40-month	0.001	0.044	0.033			

*$p < .05$, **$p < .01$, ***$p < .001$

Table 4.7

LMM Results for RT in the 4AFC Test Phase

	Model summary			Model comparison		
	β	*SE*	*t*	χ^2	*df*	*p*
Intercept	7.988	0.056	142.469	69.677	1	<.001***
Condition	0.043	0.031	1.385	1.884	1	.170
Age group				0.396	2	.820
30-month	0.042	0.070	0.596			
40-month	-0.006	0.069	-0.084			
Condition:Age group				1.053	2	.591
Condition:30-month	-0.039	0.043	-0.903			
Condition:40-month	0.034	0.042	0.809			

*$p < .05$, **$p < .01$, ***$p < .001$

4.3.2.2 Accuracy

Figure 4.10 shows children's trial-by-trial accuracy across each phase, whereas Figure 4.11 shows the distribution of children's accuracy in identifying the labelled object in each phase, split by age group and condition. Binomial generalised linear mixed-effects models (GLMMs) with a logit link function were fitted using the aforementioned `mixed()` function to analyse children's accuracy in the three phases. The models included condition (active, passive), age group (24-month, 30-month, 40-month), and the interaction between condition and age group as fixed effects, as well as selected object and participant pair as random intercepts. Both condition (-1: passive; 1: active) and age group (-1: 24-month; 1: 30-month, 40-month) were sum-coded. As none of the random slopes fell below the inclusion criterion ($\alpha = 0.20$), the random-intercepts-only models were retained:

$$\text{Accuracy} \sim \text{Condition} * \text{Age group} + (1|\text{Participant pair}) + (1|\text{Object})$$

The results of the models are detailed in Tables 4.8, 4.9, and 4.10, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons were conducted with p-values corrected using the Tukey method.

As shown in Table 4.8, there were significant main effects of condition and age group, as well as a significant interaction between condition and age group in the familiarisation phase. Results from the follow-up tests indicated that 24-month-olds were significantly less accurate than both 30-month-olds ($\beta = -2.077, SE = 0.595, z = -3.491, p = .001$) and 40-month olds ($\beta = -2.462, SE = 0.525, z = -4.688, p < .001$) in the familiar trials, but 30-month-olds' performance did not differ significantly from 40-month-olds' ($\beta = -0.384, SE = 0.719, z = -0.535, p = .854$). The simple main effect of condition (active vs. passive) was only significant in the 30-month age group ($\beta = -2.620, SE = 1.052, z = -2.492, p = .013$), with children in the passive condition being more accurate than children in the active condition. No significant difference in accuracies was found across both conditions among

24-month-olds ($\beta = -0.047, SE = 0.345, z = -0.135, p = .893$) and 40-month-olds ($\beta = -0.492, SE = 0.878, z = -0.560, p = .575$).

With regard to the 2AFC trials, there were significant main effects of condition and age group, as well as a significant interaction between condition and age group (see Table 4.9). Results from the follow-up tests indicated that 24-month-olds performed significantly worse than both 30-month-olds ($\beta = -0.944, SE = 0.221, z = -4.267, p < .001$) and 40-month-olds ($\beta = -1.319, SE = 0.226, z = -5.838, p < .001$). Performance between the 30-month-olds and the 40-month-olds did not differ significantly ($\beta = -0.376, SE = 0.231, z = -1.625, p = .235$). Crucially, the simple main effect of condition was significant in both the 30-month-olds ($\beta = -0.477, SE = 0.223, z = -2.143, p = .032$) and the 40-month-olds ($\beta = -0.558, SE = 0.239, z = -2.333, p = .020$), with children in the passive condition being more accurate than children in the active condition. No such effect was found in the youngest age group (i.e., 24-month; $\beta = 0.117, SE = 0.193, z = 0.607, p = .544$).

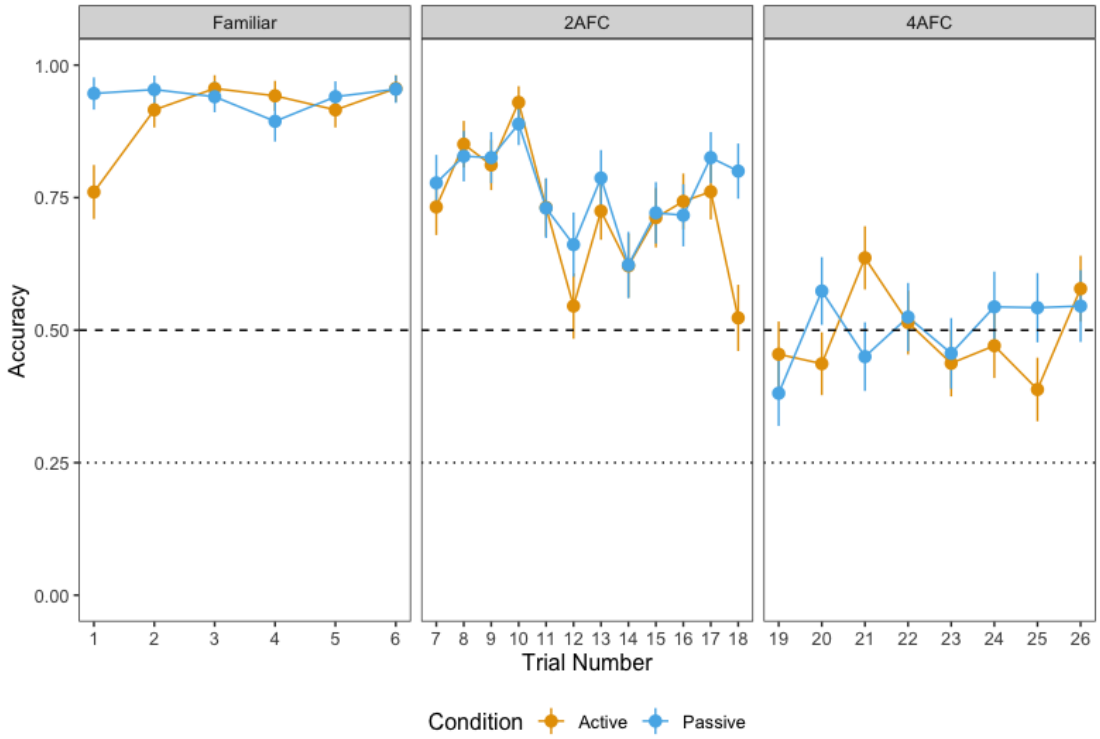
In the 4AFC test phase, there was neither a significant main effect of condition nor a significant interaction between condition and age group (see Table 4.10). Only a significant main effect of age group was found and results from the follow-up tests were similar to those obtained for the 2AFC test phase, with 24-month-olds performing significantly worse than both 30-month-olds ($\beta = -0.618, SE = 0.215, t = -2.871, p = .011$) and 40-month-olds ($\beta = -0.818, SE = 0.212, t = -3.862, p < .001$). Performance between the 30-month-olds and the 40-month-olds did not differ significantly ($\beta = -0.200, SE = 0.207, t = -0.969, p = .597$).

4.3.3 Discussion

This study set out to examine whether being given the opportunity to choose the objects that will be labelled influences 24-, 30-, and 40-month-olds' learning of these word–referent associations in a tablet-based word learning task. Children were assigned to either an active learning task, where they were allowed

Figure 4.10

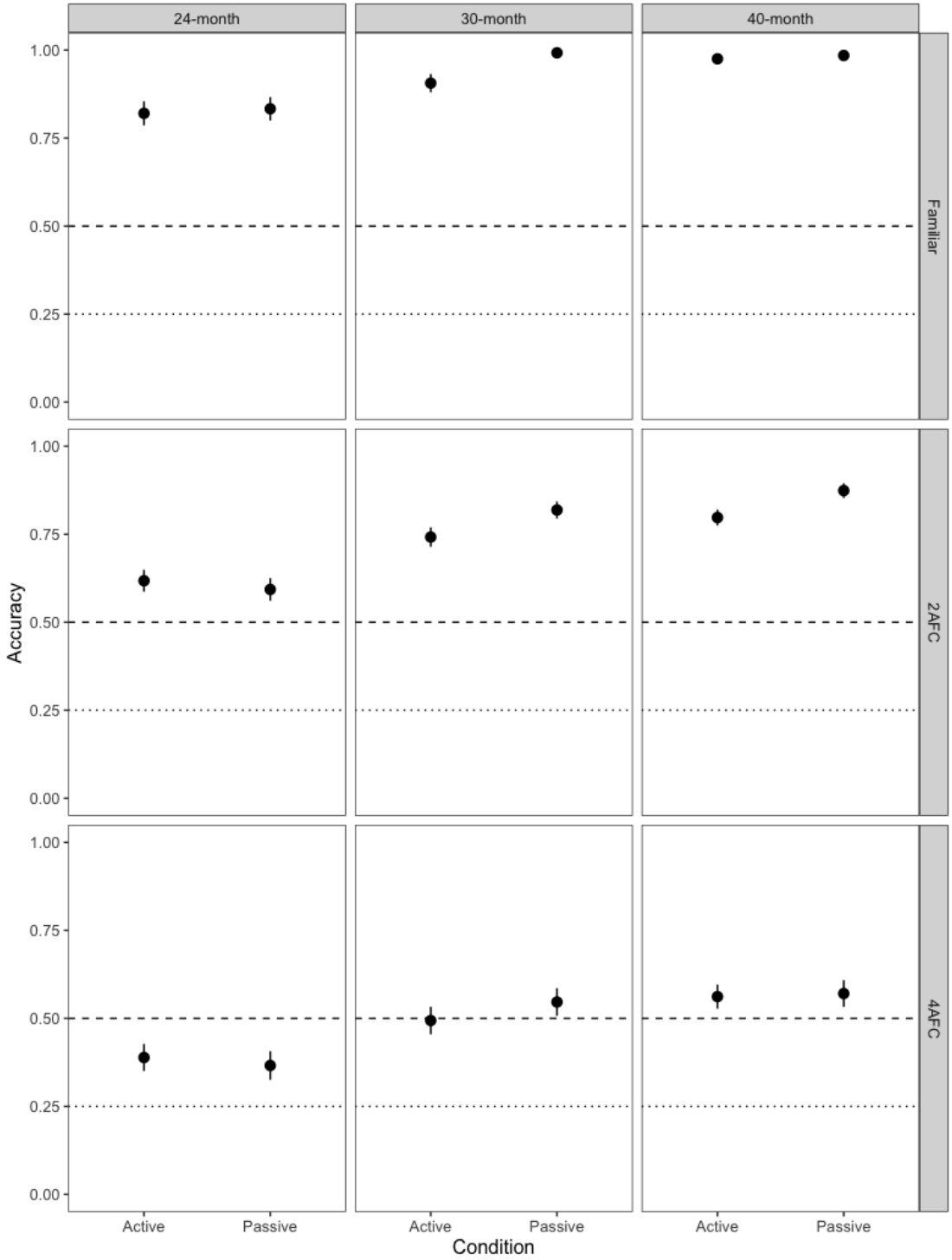
Accuracy by Trial Number



Note. Dashed line represents chance (.50) in the familiar and 2AFC test phases; dotted line represents chance (.25) in the 4AFC test phase. By-age plots can be found in Appendix D.

Figure 4.11

Accuracy by Phase and Age Group



Note. Dashed line represents chance (.50) in the familiar and 2AFC test phases; dotted line represents chance (.25) in the 4AFC test phase.

Table 4.8

GLMM Results for Accuracy in the Familiarisation Phase

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	3.184	0.278	11.461	30.526	1	<.001***
Condition	-0.527	0.236	-2.236	6.288	1	.012*
Age group				31.386	2	<.001***
30-month	0.564	0.404	1.398			
40-month	0.949	0.370	2.565			
Condition:Age group				8.707	2	.013*
Condition:30-month	-0.784	0.384	-2.039			
Condition:40-month	0.280	0.346	0.810			

*$p < .05$, **$p < .01$, ***$p < .001$

Table 4.9

GLMM Results for Accuracy in the 2AFC Test Phase

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	1.183	0.093	12.667	32.680	1	<.001***
Condition	-0.153	0.063	-2.418	5.861	1	.015*
Age group				30.554	2	<.001***
30-month	0.189	0.131	1.449			
40-month	0.565	0.133	4.236			
Condition:Age group				6.311	2	.043*
Condition:30-month	-0.086	0.090	-0.949			
Condition:40-month	-0.126	0.094	-1.346			

*$p < .05$, **$p < .01$, ***$p < .001$

Table 4.10

GLMM Results for Accuracy in the 4AFC Test Phase

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	-0.033	0.114	-0.288	0.082	1	.775
Condition	-0.018	0.067	-0.272	0.074	1	.786
Age group				14.802	2	<.001***
30-month	0.139	0.122	1.145			
40-month	0.339	0.120	2.838			
Condition:Age group				1.019	2	.601
Condition:30-month	-0.091	0.094	-0.969			
Condition:40-month	0.016	0.091	0.178			

*$p < .05$, **$p < .01$, ***$p < .001$

to choose the objects they would hear the label of or a yoked passive learning task, where they would hear the label of an object a yoked active age-matched child had chosen.

In the familiarisation phase, children were asked to tap on one of two familiar objects based on the label they were presented with. Here, 24- and 30-month-olds in the active condition were quicker to tap on the target object relative to children in the passive condition, while 40-month-olds' speeds did not differ across conditions. This finding was not unexpected as the familiarisation phase was the first point in the study where the passive group was asked to tap on the screen, while the active group had been doing so since the learning phase. In fact, the familiarisation phase was included to remove the potential confound of prior experience in tapping after finding a similar pattern in the pilot study. Thus, while there appears to be an active advantage in the recognition of familiar objects, this appears to be an artefact of the task and the experience that children in the two groups had with tapping on the screen. Across the three age groups tested, 40-month-olds were the quickest in identifying the target object, while 24- and 30-month-olds were relatively slower. With regard to the accuracy of children's responses, a passive advantage was found, with 30-month-olds responding more accurately in the passive condition, but no such passive advantage was found in both the younger and the older age groups. Even with the 30-month-olds, this appears to be limited to the first trial and not to later trials (see Figure D.2). Especially with regard to the older age groups, responding was at ceiling (see Figure 4.11). Given this pattern of responding, the differences between active and passive children in the familiarisation phase should be treated with caution.

In the 2AFC test phase, children were asked to tap on one of two novel objects based on the label they were presented with. Overall, no differences were found in terms of RT, suggesting that children in the passive condition, regardless of age, had familiarised themselves with the tapping paradigm through the familiar trials and that all children were recognising and tapping the target at similar speeds (see Figure 4.9). With regard to the accuracy measure, a

significant main effect of condition was found and this interacted with age, suggesting a developmental difference in the passive advantage across the ages tested. Specifically, older children (i.e., 30- and 40-month-olds) who were assigned the passive condition responded with increased accuracy relative to the yoked active age-matched children. No such difference in accuracy was found in the youngest children (i.e., 24-month-olds). Across the three age groups, the youngest had the lowest accuracies, while the older age groups' performance did not differ significantly. This is congruent with Russo-Johnson et al. (2017), where the youngest children learnt significantly fewer words than the older children and the older children learnt equally. Similar age effects were found in the 4AFC test phase, where children were asked to identify the target object among four novel objects. In particular, the youngest children responded with the lowest accuracies, while the older children responded with similar accuracies. No differences across conditions were found both in terms of RT and accuracy in the 4AFC test phase however.

Although the finding of a developmental difference in the observed passive advantage in terms of accuracy is in line with other studies showing improvement in performance for children assigned to a passive condition relative to conditions including pseudo-social contingency (Choi & Kirkorian, 2016; Kirkorian, Choi, et al., 2016), what remains uncertain is whether the differences across the two conditions found here relate to differences in children's performance or their competence. In other words, do children assigned to the active condition merely perform worse than their passive peers while nevertheless having learnt the words to an equal degree or do children assigned to the active condition also learn worse than their passive peers? For instance, one explanation for the poorer performance of the active children may be that they continue to choose the objects that they like (i.e., treating the test phases as the learning phase) rather than choosing the objects whose label they have been presented with, despite having learnt the novel word–referent associations. Clarification of the competence–performance distinction is therefore required before further interpretation of the results is possible.

Study 1B examined this issue in further detail using a more implicit measure of children's eye movements as they completed the word learning task. If having an active choice disrupts children's learning from tablets, a poorer performance (i.e., less accurate fixations to the target object in the test trials) would be expected in the active children, even on such an implicit measure. On the other hand, if the lower accuracies of the active children are due to their non-conformance to the demands of the task (i.e., to identify and tap on the labelled object), similar performance, as indexed by the looking time measure, would be expected across both the active and the passive conditions. Study 1B thus attempted to replicate the results of the present study, while extending this using an additional implicit looking time measure (similar to the preferential looking tasks used in laboratory studies). In addition, Malay-speaking children from Malaysia were tested to allow the examination of the extent to which the findings replicate in children from a different cultural and linguistic background.

4.4 Study 1B

4.4.1 Method

4.4.1.1 Participants and Design

Thirty-two typically developing, primarily monolingual Malay-speaking children, aged between 28 and 35 months
($M = 30.25, SD = 1.71, \text{range} = 27.59–34.76$) were recruited from nine childcare centres in Selangor, Malaysia. The study took place in a quiet room at the participants' respective childcare centres. Yoked age-matched pairs of participants (ages at date of testing within half a month of each other) were assigned to either the active or the passive condition. As in the pilot study and Study 1A, in the active condition, participants could select four novel objects to be told the labels of, while in the passive condition, participants were automatically given the labels of the objects chosen by their yoked active peers. Due to a clear side preference in selection (i.e., tapping eight times consecutively on the image shown on a particular side; $n = 3$) and inattentiveness (i.e., getting

up and walking around during the study; $n = 3$), an additional six pairs of participants had to be excluded from the analysis. The study was reviewed and approved by the Science and Engineering Research Ethics Committee of the University of Nottingham Malaysia. Caregivers gave written consent to their child's participation in the study and webcam video recording of their child during the study.

4.4.1.2 Apparatus and Materials

The study was carried out using a Microsoft Surface Pro 3 tablet with a web application[15] that captures both a participant's implicit (gaze)—with the device's built-in front-facing camera—and explicit (tapping) responses. Images of eight novel objects and six familiar objects were chosen for the study (see Study 1A). Four disyllabic, novel words were selected to be used as labels for the chosen novel objects: "banung", "ifi", "mipo", and "pafka". These words obey the phonotactic constraints of Malay (see Appendix E for further details). The six familiar words were: "epal" [apple], "kereta" [car], "bayi" [baby], "bola" [ball], "pokok" [tree], and "kasut" [shoe]. All auditory stimuli used were recorded by a female native speaker of Malay in child-directed speech.

4.4.1.3 Procedure

The procedure was identical to Study 1A with the only differences being the language in which the stimuli were presented and that webcam videos of the participants were recorded for the entire duration of the study.

4.4.1.3.1 Learning Phase

Active Condition The learning phase was set up identically to that of Study 1A, with the only difference being the language in which the prompts were produced. Thus, in the first trial, the prompt asking participants to select one of

[15]Programmed using an adapted version of e-Babylab (Chapter 3) that allows test trials to be dynamically generated based on the novel objects selected during the learning phase.

the two randomly generated images of the novel objects was “Tengok ni, sini ada dua gambar. Pilih satu.” [Look, here are two pictures. Pick one.] For subsequent trials, the prompt was “Pilih satu gambar, lepas tu kita akan dengar nama dia.” [Pick a picture and then we’ll hear its name.] Upon tapping, the selected novel object was then labelled five times in the same trial using various carrier phrases, including: (a) “Tengok, X!” [Look, a/an X!], (b) “Ini adalah X!” [This is a/an X!], (c) “Wow, itu X!” [Wow, that is a/an X!], (d) “Nampak tak X?” [Do you see the X?], and (e) “Bagus! Ini adalah X!” [Great! This is a/an X!], where X was the novel word.

Passive Condition Passive learning participants were only required to watch and listen as they would be exposed to the age-matched active learning peer’s selections according to the exact timings of the active peer. The auditory prompt presented in the first trial was “Nampak tak dua gambar tu? Cantik kan?” [Do you see the two pictures? Beautiful, right?], and in subsequent trials, “Mari kita dengar nama untuk gambar lagi.” [Let’s hear names for pictures again.] to attract participants’ attention to the images. All other details were identical to Study 1A

4.4.1.3.2 Familiarisation Phase

As in Study 1A, six familiar trials were included. In each familiar trial, participants were presented with a pair of familiar objects, followed by the instruction to tap on one of these objects based on a given label X embedded in the carrier phrase “Tunjukkan gambar X.” [Show (me) the picture of X.]

4.4.1.3.3 2AFC/4AFC Test Phase

All details of the design for the 2AFC and 4AFC tasks were identical to Study 1A, with the exception being that the auditory prompts were in Malay (see carrier phrase from the familiarisation phase).

4.4.1.4 Gaze Analysis

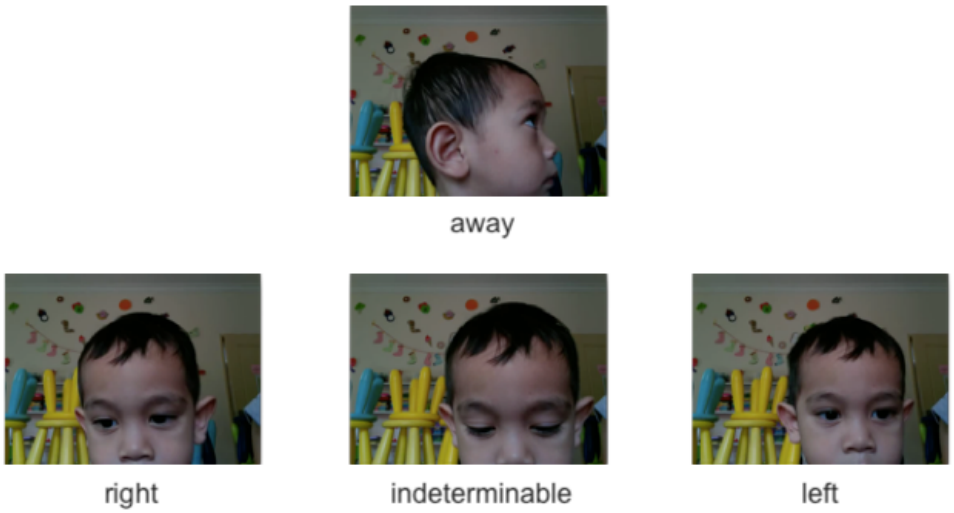
In addition to participants' explicit responses, participants' eye movements were also recorded in all trials, including trials in the learning phase. To quantify this, each video was split into 200 ms chunks, as in Semmelmann et al. (2017), on the basis that saccades take approximately 200 ms to initiate (Purves et al., 2012). These video chunks were presented in a random order to the rater who was to rate them as: (a) "left", when the participant was looking to the left side of the screen; (b) "right", when the participant was looking to the right side of the screen; (c) "away", when the participant was looking away from the screen; or (d) "indeterminable", when none of the three other options were applicable (see Figure 4.12 for examples). To avoid potential biases, rating was carried out in a blind rating situation under which the position of the target was unknown to the rater. As it was not feasible to rate 4AFC trials, only the learning trials, the familiar trials, and the 2AFC trials were rated. Participants' eye movements were rated for all four learning trials from the onset of the labelling for the selected novel object (i.e., right after a selection was made). Looking time during the learning phase was used as a predictor in the models examining learning in the 2AFC and 4AFC test phases to account for differences in attention to the labelled object during learning across conditions. In both the familiarisation phase and the 2AFC test phase, participants' eye movements were rated from the onset of the presented target word to when participants chose an object.

Ten percent of the video chunks were rated by two raters. Calculating Cohen's Kappa, a substantial agreement (McHugh, 2012) was found between the two raters overall, $\kappa = 0.705$ (79.7% agreement). Upon excluding video chunks which were rated as "indeterminable", an almost perfect agreement was found, $\kappa = 0.950$ (97.1% agreement). When only differentiating between "left" and "right", agreement rose to 99.2%, $\kappa = 0.984$. Thus, it can be inferred that both raters agreed on the side of the screen participants were looking at, when they were able to decide on one.

Following video rating, the proportion of looks to the target in each trial was computed, as is standard in the literature (e.g., Bion et al., 2013; Fernald et al., 2010; Johnson & Huettig, 2011). The target was set as the object that was labelled in both the learning trials and the test trials. This measure (i.e., proportion of target looks), together with the time course graphs and the overall statistics (presented in the next section), captures not only the duration of looks to the target but also the duration of look-aways to the distractor, since the proportion of target looks would correspondingly drop at any given time if the child was looking at the distractor rather than the target.

Figure 4.12

Video Rating Scale



Note. Each video was split into 200 ms chunks and rated as either looking at the "left" or "right" (side of the screen), "away" (from the screen) or "indeterminable". Written consent for publication of the participant's pictures was obtained from the caregiver.

4.4.2 Results

4.4.2.1 Gaze

To examine potential differences between the active and the passive participants' gaze patterns over the course of the learning trials, familiar trials, and 2AFC trials, three cluster-based permutation analyses were conducted for each of these trial types (c.f. Dautriche et al., 2018; Hahn et al., 2015; Kartushina & Mayor, 2019; Von Holzen & Mani, 2012) using the *eyetrackingR* package (Dink & Ferguson, 2018). The first compared the average proportion of looks to the target between the two conditions (active vs. passive), whereas the second and third compared the average proportion of looks to the target in each condition to the chance level (0.50; active vs. chance and passive vs. chance).

To minimise the effect of motor planning, only fixations that occurred between 200 and 2000 ms post target word onset were considered for the familiar trials, whereas for the 2AFC trials, the analysis time window was between 400 and 2200 ms post target word onset as children take longer in mapping newly learnt words than familiar words (Bion et al., 2013; Booth & Waxman, 2009). Earlier eye movements were also excluded given that the mobilisation of an eye movement in infants requires at least about 2–300 ms (Canfield et al., 1997; Haith et al., 1993). Furthermore, similar criteria have been used in word recognition studies involving the use of eye movements (e.g., Fernald & Marchman, 2012; Fernald et al., 2010; Swingley & Fernald, 2002).

Prior to the analyses, trials where more than 25% of the video chunks were rated as "indeterminable" were removed. This retained 113 of 128 trials from all 32 participants in the learning phase, 182 of 184 trials from all 32 participants in the familiarisation phase, and 311 trials from 31 participants of 372 trials from 32 participants in the 2AFC test phase. All proportions of target looks were arcsine-root transformed to better fit the assumptions of the t-test conducted at each time point to compare the proportions of target looks to chance or between the two conditions. Time points with a significant effect ($t > 2, p < .05$) were then grouped into a cluster, for which its size was obtained

from the summation of all t-values within this cluster. To test the significance of a cluster, 1000 simulations in which conditions (active vs. passive, active vs. chance, passive vs. chance) were assigned randomly for each trial were conducted. The size of the biggest cluster in each simulation was then obtained using the same procedure as before with the real data. If the probability of observing a cluster—from the randomised data—with the same size as or bigger than the cluster from the real data was smaller than 5% ($p < .05$), the cluster from the real data was considered significant; in other words, the differences (active vs. passive, active vs. chance, passive vs. chance) were significant.

4.4.2.1.1 Learning Phase

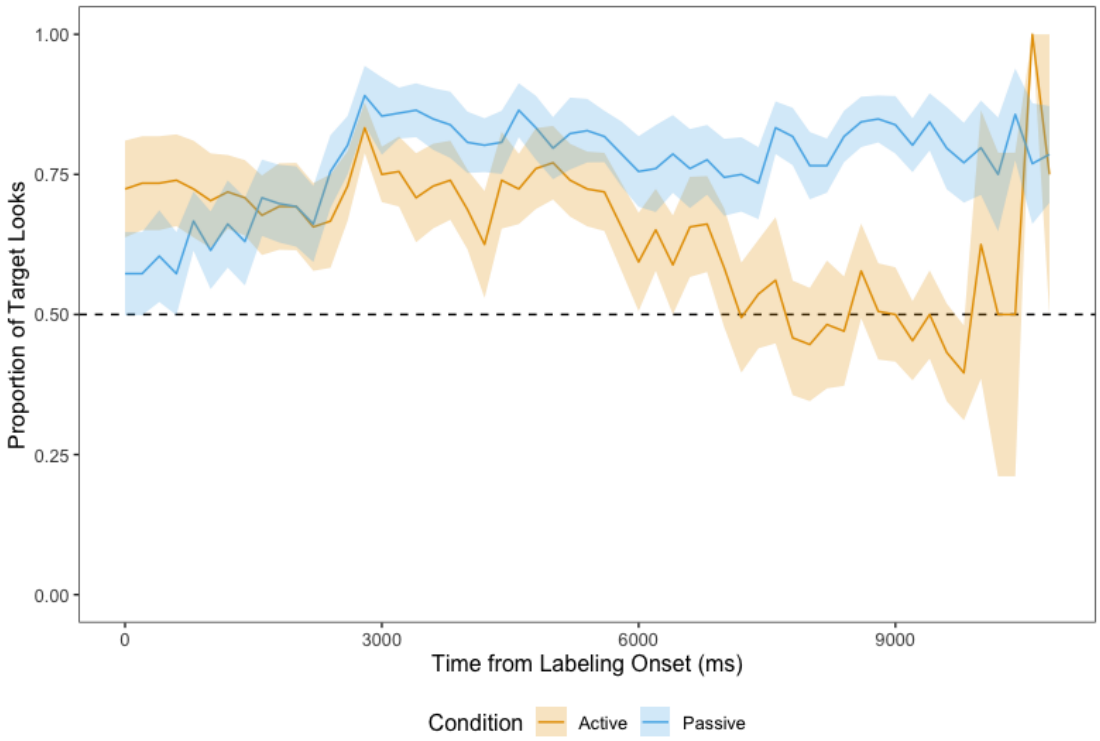
Figure 4.13 shows children's proportion of looks to the target across all four learning trials, from the onset of the labelling of the selected novel object. As the figure suggests, children in the passive condition looked more at the target than children in the active condition overall. Indeed, the cluster-based permutation analysis led to the identification of a significant difference across conditions between 7600 ms and 9800 ms following the onset of the label ($p = .001$). Children in the passive condition fixated the target significantly above chance (0.50) for most of the duration of the 10 s labelling phase (from 1600 ms to 10000 ms, $p < .001$), while their active peers fixated the target significantly above chance (0.50) during the first half of the labelling phase (from 0 ms to 2000 ms, $p = .007$; from 2600 ms to 4000 ms, $p = .006$; from 4400 ms to 5600 ms, $p = .018$).

4.4.2.1.2 Familiarisation Phase

Figure 4.14 shows children's proportion of looks to the target from the onset of the target word in the familiar trials. The cluster-based permutation analysis revealed no time points where a significant difference between the active and the passive conditions could be found. Children from both conditions fixated the target significantly above chance (0.50) shortly after the target word onset (from 800 ms to 2000 ms, $p < .001$).

Figure 4.13

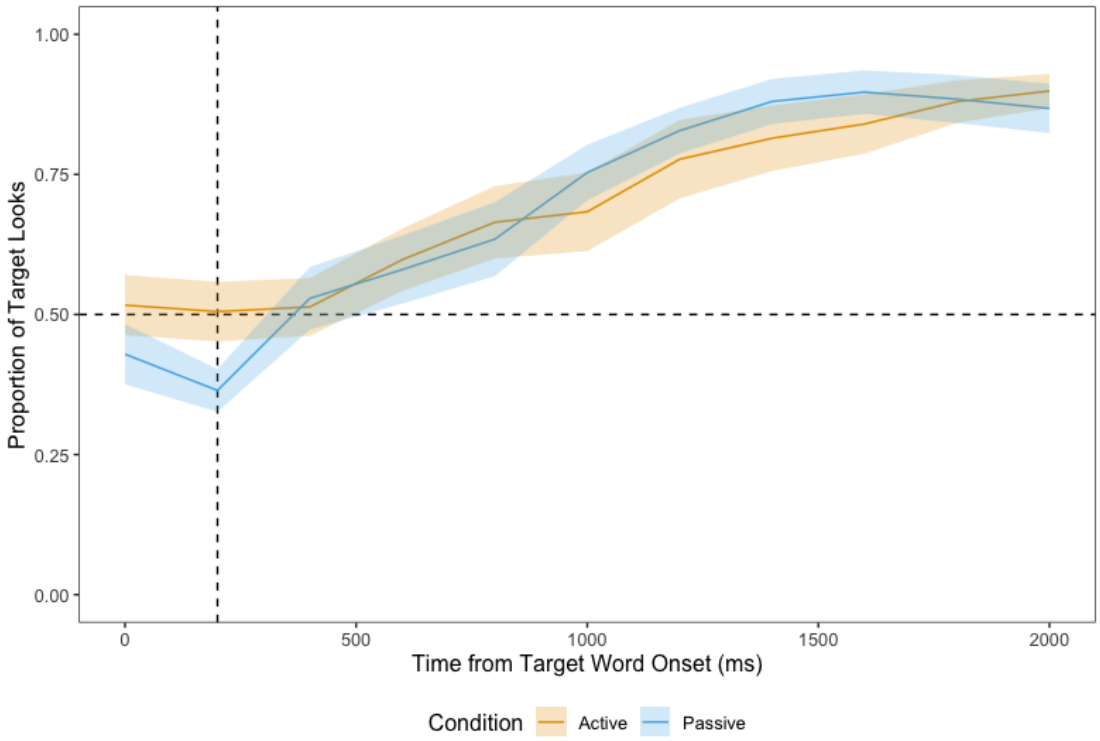
Proportion of Target Looks in the Learning Trials



Note. Proportion of target looks is time-locked to the labelling of the selected novel object. Dashed line represents chance (0.5).

Figure 4.14

Proportion of Target Looks in the Familiar Trials



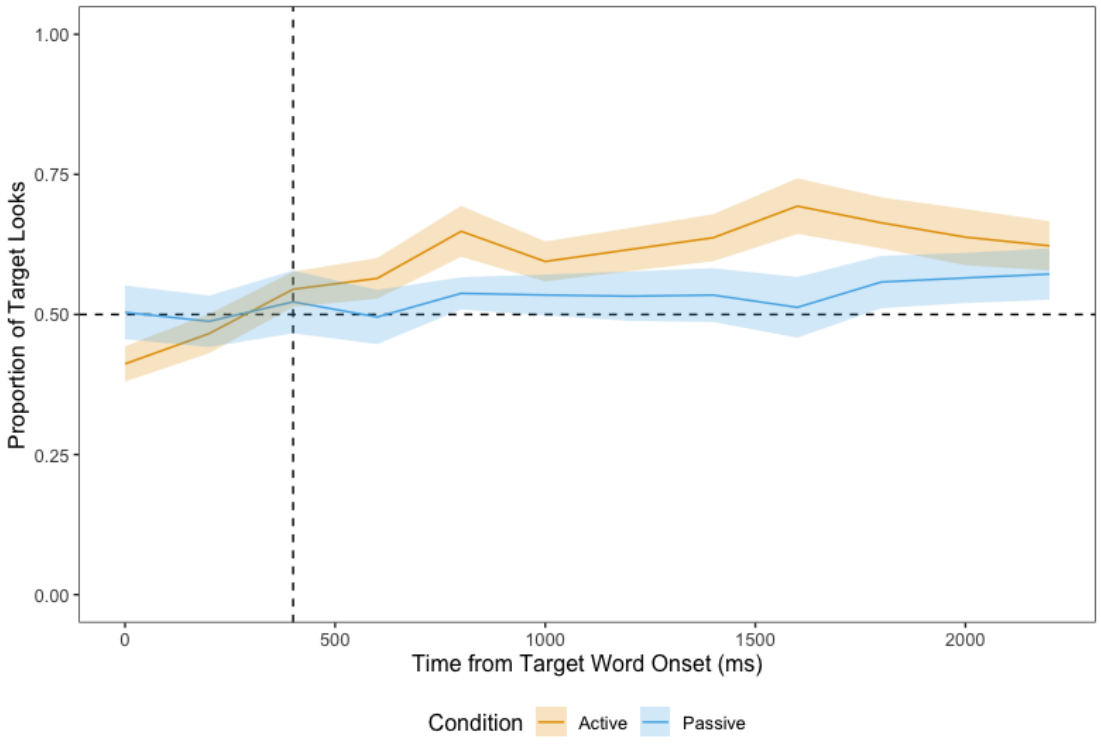
Note. Proportion of target looks is time-locked to the onset of the target word. Dashed vertical line at 200 ms marks the beginning of the analysis window; dashed horizontal line represents chance (0.5).

4.4.2.1.3 2AFC Test Phase

Figure 4.15 shows children's proportion of looks to the target from the onset of the target word in the 2AFC trials. The cluster-based permutation analysis revealed no time points where a significant difference between the active and the passive conditions could be found. Children in the active condition fixated the target significantly above chance (0.50) shortly after the target word onset (from 800 ms to 2200 ms, $p < .001$). On the other hand, no significant time point was identified for children in the passive condition, although a one-sample t-test across the entire time window indicated that they looked significantly above chance; $t(160) = 1.928, p = .028$.

Figure 4.15

Proportion of Target Looks in the 2AFC Trials



Note. Proportion of target looks is time-locked to the onset of the target word. Dashed vertical line at 400 ms marks the beginning of the analysis window; dashed horizontal line represents chance (0.5).

4.4.2.2 Reaction Time

RT was measured in ms from the onset of the target word. As participants were not given a time limit to respond, RTs included outliers as high as 114 s. Since the data did not follow a normal distribution as indicated by a Shapiro-Wilk normality test ($W = 0.495, p < .001$), the data was log transformed prior to further analysis to approximate a normal distribution. To ensure that only those trials where the child was engaged in the task were included in the analysis, outliers were removed using a criterion of 2 *SD*s above and below the mean (24 active, 11 passive). Mean and standard deviation of RT for each condition, before and after outlier removal are detailed in Table 4.11.

Table 4.11

Mean and Standard Deviation of RT Before (Unadjusted) and After (Adjusted) Outlier Removal, Split by Condition

Condition	Unadjusted M_{RT} (s)	Unadjusted SD_{RT} (s)	Adjusted M_{RT} (s)	Adjusted SD_{RT} (s)
Active	4.671	7.063	4.109	3.447
Passive	5.340	6.476	4.607	3.848

Figure 4.16 shows children's trial-by-trial RT across the familiarisation phase as well as the 2AFC and 4AFC test phases, whereas Figure 4.17 shows the distribution of children's RT split by phase. LMMs were fitted to assess whether RTs differed across conditions (active vs. passive) in each of the three phases. The model for the familiarisation phase included condition (sum-coded; -1: passive; 1: active) as a fixed effect, whereas the models for the 2AFC and 4AFC test phases included an additional fixed effect of proportion of looks to the target during the learning trials. As in Study 1A, parsimonious models were determined using the forward "best-path" approach to test random slopes for inclusion ($\alpha = 0.20$). The resulting models for the familiarisation phase and the

2AFC test phase therefore included selected object and participant pair as random intercepts, with by-participant-pair adjustment to the slope of condition:

$$\text{RT}_{\text{log}} \sim \text{Condition} + \text{Learning looks} + (1 + \text{Condition}|\text{Participant pair}) + (1|\text{Object})$$

The model for the 4AFC test phase included target word and participant as random intercepts, with by-participant-pair and by-object adjustments to the slope of condition:

$$\begin{aligned}\text{RT}_{\text{log}} \sim \text{Condition} + \text{Learning looks} \\ + (1 + \text{Condition}|\text{Participant pair}) + (1 + \text{Condition}|\text{Object})\end{aligned}$$

The results of the models are detailed in Tables 4.12, 4.13, and 4.14, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. As the tables suggest, children in both the active and passive conditions did not differ significantly in terms of speed in responding overall. While children in the passive condition were slower in the first few trials of each test phase, they quickly caught up with children in the active condition (see Figure 4.16 and Figure 4.17). A significant effect of proportion of target looks during the learning trials was found in the 4AFC test phase, with children who spent more time fixating the target during the learning phase being quicker to tap on the target object.

Table 4.12

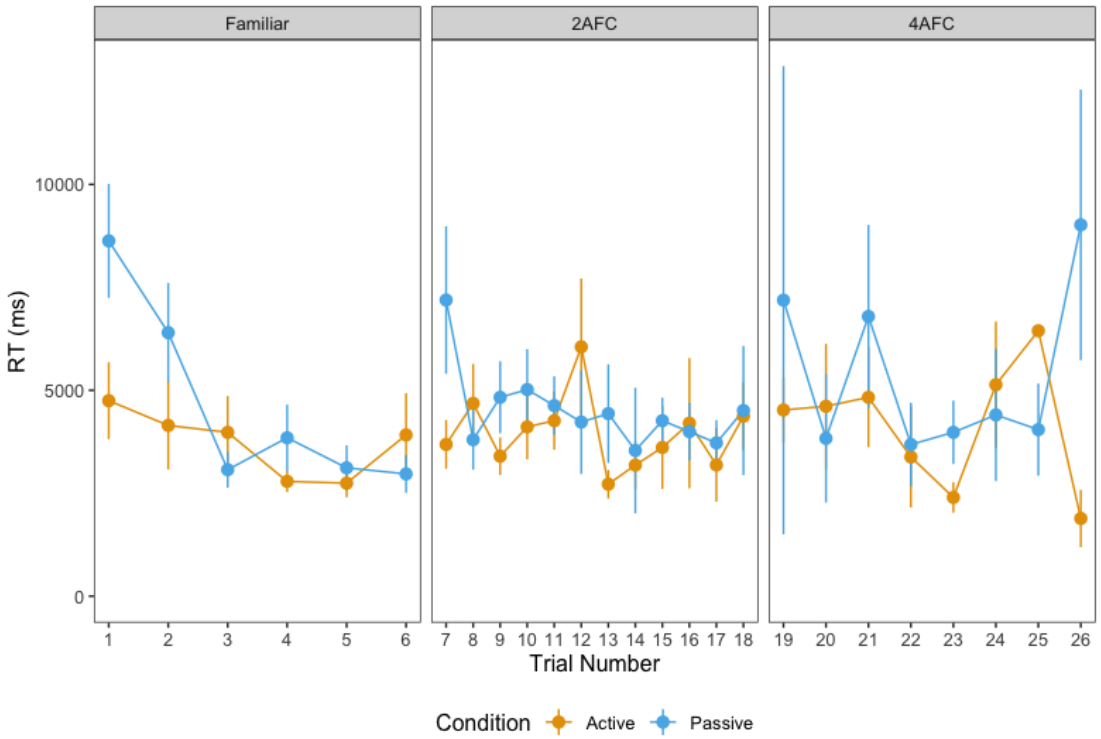
LMM Results for RT in the Familiarisation Phase

	Model summary			Model comparison		
	β	*SE*	*t*	χ^2	*df*	*p*
Intercept	8.072	0.116	69.799	41.887	1	<.001***
Condition	-0.064	0.094	-0.689	0.468	1	.494

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 4.16

RT by Trial Number



Note. Only trials in which children responded correctly are considered.

Table 4.13

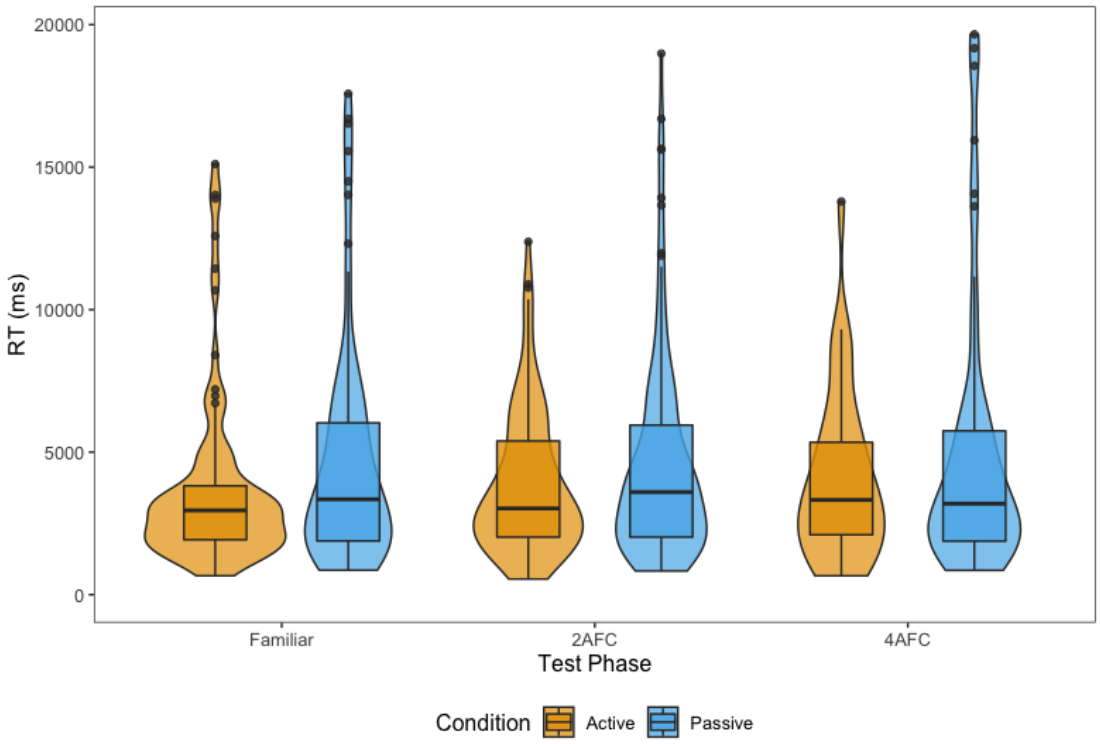
LMM Results for RT in the 2AFC Test Phase

	Model summary			Model comparison		
	β	SE	t	χ^2	df	p
Intercept	8.214	0.149	55.131	107.880	1	<.001***
Condition	-0.069	0.104	-0.664	0.436	1	.509
Learning looks	-0.144	0.190	-0.756	0.567	1	.451

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 4.17

RT by Phase



Note. Only trials in which children responded correctly are considered.

Table 4.14

LMM Results for RT in the 4AFC Test Phase

	Model summary			Model comparison		
	β	SE	t	χ^2	df	p
Intercept	9.258	0.281	32.973	59.731	1	<.001***
Condition	-0.132	0.157	-0.845	0.690	1	.406
Learning looks	-1.564	0.352	-4.440	11.872	1	<.001***

*$p < .05$, **$p < .01$, ***$p < .001$

4.4.2.3 Accuracy

Figure 4.18 shows children's trial-by-trial accuracy across each phase, whereas Figure 4.19 shows children's mean accuracy in identifying the labelled object in each phase. Binomial GLMMs with a logit link function were fitted to analyse children's accuracy in the three phases. The model for the familiarisation phase included condition (sum-coded; -1: passive; 1: active) as a fixed effect as well as selected object and participant pair as random intercepts. The models for the 2AFC and 4AFC test phases included an additional fixed effect of proportion of looks to the target during the learning trials. As none of the random slopes fell below the inclusion criterion ($\alpha = 0.20$), the random-intercepts-only models were retained:

$$\text{Accuracy} \sim \text{Condition} + \text{Learning looks} + (1|\text{Participant pair}) + (1|\text{Object})$$

The results of the models are detailed in Tables 4.15, 4.16, and 4.17, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. As Table 4.15 suggests, there was no significant main effect of condition on accuracy in the familiarisation phase. However, in both the 2AFC and 4AFC test phases, condition significantly predicted accuracy, with children in the passive condition providing more accurate responses than children in the active condition. Proportion of looks to the target during the learning trials was not a significant predictor in both critical test phases.

Table 4.15

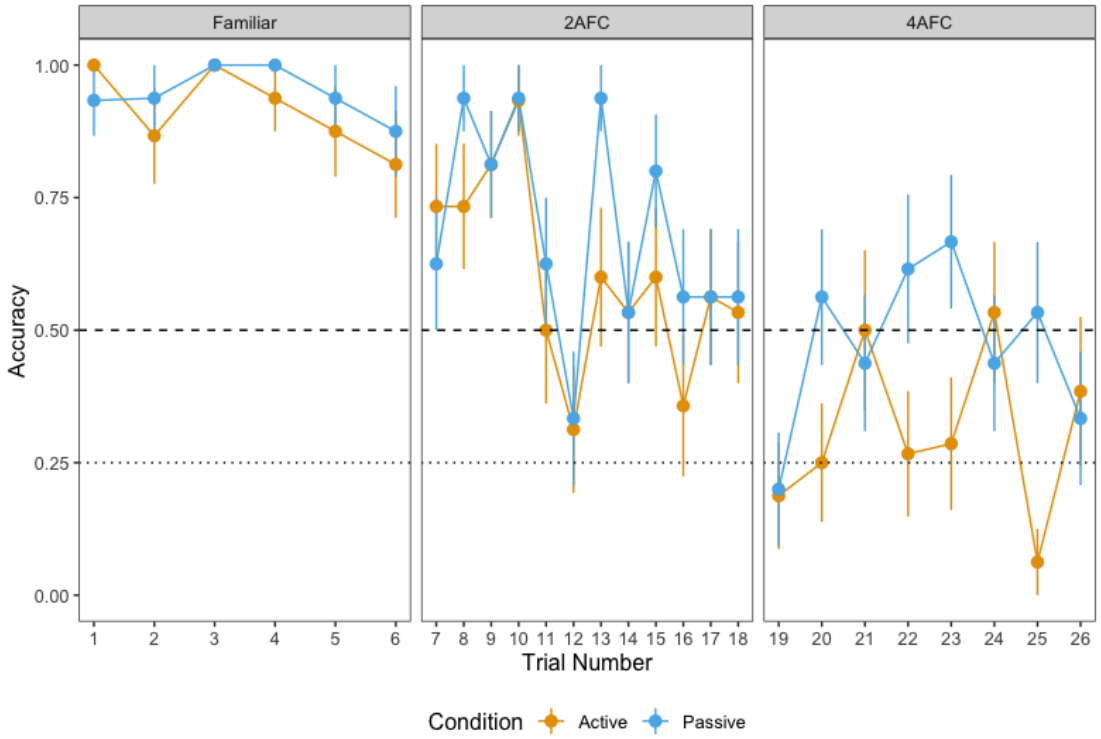
GLMM Results for Accuracy in the Familiarisation Phase

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	2.627	0.295	8.907	18.191	1	<.001***
Condition	-0.264	0.295	-0.894	0.820	1	.365

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 4.18

Accuracy by Trial Number



Note. Dashed line represents chance (.50) in the familiar and 2AFC test phases; dotted line represents chance (.25) in the 4AFC test phase.

Table 4.16

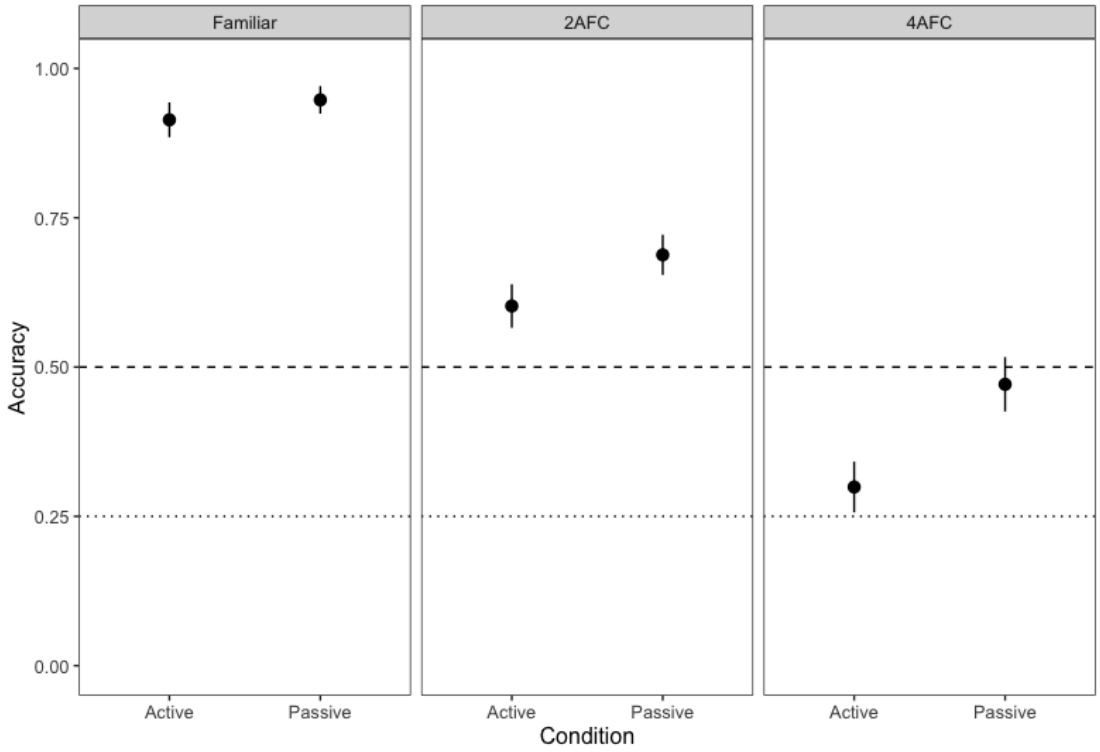
GLMM Results for Accuracy in the 2AFC Test Phase

	Model summary			Model comparison		
	β	SE	z	χ^2	df	p
Intercept	0.989	0.393	2.518	6.676	1	.010**
Condition	-0.246	0.115	-2.146	4.657	1	.031*
Learning looks	-0.549	0.531	-1.033	1.084	1	.298

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 4.19

Accuracy by Phase



Note. Dashed line represents chance (.50) in the familiar and 2AFC test phases; dotted line represents chance (.25) in the 4AFC test phase.

Table 4.17

GLMM Results for Accuracy in the 4AFC Test Phase

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	-1.220	0.500	-2.442	6.318	1	.012*
Condition	-0.290	0.142	-2.047	4.222	1	.040*
Learning looks	1.035	0.666	1.554	2.481	1	.115

*$p < .05$, **$p < .01$, ***$p < .001$

4.4.3 Discussion

Study 1B set out to replicate the findings of Study 1A with children from a different cultural background while also examining a more implicit measure of recognition performance—namely, looking time data—across the active and the passive conditions. With regard to RT and accuracy, a very similar pattern of responding was found among same-aged children from Germany and Malaysia (30-months). In particular, children were equally fast in identifying the target object across both conditions, but children in the passive condition responded with greater accuracy than children in the active condition, in the 2AFC test phase. The Malaysian children also demonstrated a passive advantage in terms of accuracy in the 4AFC test phase. With regard to their performance in the familiarisation phase, no differences were found across the two conditions.

Interestingly, the analysis of children's gaze behaviour in the learning phase revealed that children in the passive condition fixated the labelled target object significantly longer and more robustly than their active peers, suggesting that children in the passive group may be more engaged with the learning material. One possible explanation for this pattern is that the design of the learning phase set the stage for different learning experiences across conditions: active children, who are allowed to tap from the very beginning, have a more game-like experience than their passive peers, who are only allowed to tap later. Passive children might thus take the task more seriously, resulting in taking more time to encode the word–referent associations. Alternatively, it may also be that active children have already explored the object in depth before making the choice and once their choice is made, they no longer need to examine this object in further detail, while passive children may reengage with the target object once this object has been presented as the target.

Nevertheless, analysis of children's performance in both the 2AFC and 4AFC test phases revealed that gaze duration during the learning phase has no significant effect on children's accuracy in the test phases. While children in the passive condition looked longer at the target object during the learning phase and outperformed their active peers in terms of accuracy, the former did not

predict the latter. Neither did children's gaze behaviour in the 2AFC trials differ across the two conditions. This is particularly revealing given that children's accuracies differed in both test phases. Taken together, there appears to be no evidence that passive children's increased engagement with the learning material led to their improved recognition performance in terms of accuracy. There is also no evidence for a difference across conditions in children's gaze behaviour during the 2AFC test phase, suggesting that all children spent an equal proportion of time fixating the target. The implications of these results are further discussed in the next section.

4.5 General Discussion

In recent years, tablet ownership in families with children has increased drastically (Rideout, 2017) and parents have, at their fingertips, a wide selection of educational apps that claim to boost children's learning. However, as a majority of these apps have not been formally evaluated before release (Hirsh-Pasek et al., 2015), many may fall worryingly short of their pledge.

The present studies aimed to bring together recent debates on active learning and learning from interactive touchscreen media. They set out to explore how active selection of learning experiences affects word learning from a tablet-based app in 24-, 30-, and 40-month-old children. Children were assigned to either an active or a yoked passive condition. In the active condition, children were allowed to choose the object they wanted to hear the label of and then assessed on their recognition of the novel word–referent associations using both a tapping task (Study 1A and Study 1B) and implicit gaze data (Study 1B). In both studies, differences across conditions were found in terms of children's accuracy in the identification of the target object. In particular, a passive advantage was found at 30- and 40-months, with children in the passive condition showing greater accuracy in target recognition.

This apparent passive advantage may either be explained by a competence or a performance deficit with regard to the active children. The competence deficit explanation would suggest that interacting with the app by

tapping during the learning phase may take up valuable cognitive resources. Children in the passive condition, who do not have to allocate resources to tapping, have more capacity to encode and retain the information presented to them. In this case, the active children may actually learn and encode the novel word–referent associations worse than the passive children. On the other hand, the performance deficit explanation would suggest that children in the passive condition may approach the task differently relative to children in the active condition. As children in the active condition are allowed to tap on their preferred objects during the learning phase, they might treat the test phases as an extension of the learning phase and thus continue to merely indicate their preference for one of the objects during the test phases. Relatedly, the learning phase might have primed children in the active condition to tap reflexively and set the prepotent (tapping) response in motion, such that instead of paying attention to the task goal during the test phases (i.e., to identify and tap on the labelled object), children might be waiting for their next chance to tap and do so as soon as they can, regardless of instruction. In contrast, tapping might have been more reflective (requiring thoughtful attention) than reflexive for children in the passive condition, since for them, tapping was only allowed in the test phases and was always associated with the same task goal throughout the study. This interpretation would be in line with Russo-Johnson et al. (2017) who argue that engaging in prepotent tapping response may distract children from focusing on the task at hand. Here, the observed passive advantage does not reflect children's competence, but rather their performance: the difference in the design of the learning phase affects how children approach the task, which in turn influences their behaviour in the subsequent test phases.

Given the different possible reasons for the findings in Study 1A, Study 1B examined the root of this passive advantage. In other words, did active children not learn and correctly map the novel words to the objects (relative to the passive children), or did they merely not perform correctly (i.e., not tap on the target object despite knowing what the target object was)? To answer these questions, children's eye movements were recorded as they completed the task in

Study 1B. Despite finding a very similar pattern of responding as in Study 1A, no evidence for a difference in the time course of active and passive children's recognition of the target object was found, that is, children in both conditions fixated the target above chance and for the same proportion of time during the test phase. While passive children, relative to their active peers, fixated the target longer during the learning phase, the fact that active children fixated the target object, at the very least, in a similar manner to the passive children during the test phase suggest that differences found in the accuracy measure are unrelated to their competence in word learning but rather their performance in tapping.

Taken together, these results suggest caution in advocating for either a boost in learning when children are allowed to choose what they want to learn (Partridge et al., 2015) or when children are passively presented with new information (Choi & Kirkorian, 2016; Kirkorian, Choi, et al., 2016). At the very least, no differences were found in children's competence across the active and the passive conditions. Rather, the difference lies in children's performance across the two conditions, highlighting issues with the design of active learning tasks that may need to be considered in planning digital learning tools. Given that cognitive flexibility is not well developed at such a young age, children may not yet be able to reliably adapt their behaviour in response to changing task demands. For instance, when asked to sort coloured shapes, 2.5- and 3-year-olds could not reliably switch from the initial rule (e.g., sort by colour) to a new rule (e.g., sort by shape; Blakey et al., 2016). Likewise, children in the active condition may have difficulties changing course during the word learning task, moving from actively choosing what they want to learn more about to indicating what they have learnt, despite being told what they needed to do across the different phases of the study.

Nevertheless, no such passive advantage was found (at least after the first trials) in familiar trials. In other words, a reliable passive advantage was only found in trials where children were tested on their knowledge of the novel word–referent associations and not in trials where they were tested on their

recognition of highly familiar word–referent associations. Thus, it may also be that the robust word knowledge associated with the familiar objects overcomes their prepotent tapping response and conversely, the partial word knowledge associated with the novel objects is too fragile to overcome the prepotent tapping response.

While German children's accuracies did not differ across conditions in the 4AFC test phase, a passive advantage was demonstrated among Malaysian children. It is likely that the sudden increase in difficulty, as the number of distractors increased from one to three, might have had an impact on children, thus overriding the differences across some children in this task. Nevertheless, a passive advantage in performance may be expected at some ages even in such a task. Indeed, visual inspection of the data from the German 30-month-olds suggests a potential passive advantage in all but two trials (a similar pattern is observed among Malaysian children of the same age; see Figure D.2 and Figure 4.18).

Lured by the bold claims that some educational apps make, parents of young children may be tempted to download a large number of apps in hopes of fostering their children's learning in various domains. However, the present studies add to the growing body of evidence that these claims should be taken with caution, since the apps may not be adequately tapping into children's learning progress. Depending on how an educational app is structured, it places the child in the role of an active, self-guided learner. While there is evidence that children can benefit from active learning in some circumstances, the present studies paint a different picture, suggesting that an active advantage or a passive advantage is highly contingent on the task structure and taking this further, the app structure. Depending on the structure of the learning experience, an active choice may actually decrease children's performance in certain tasks, without having much impact on their learning competence. Thus, the attentional and locomotor constraints specific to touchscreen usage should be kept in mind when talking about learning from interactive touchscreen media.

4.6 Summary

This chapter described a series of studies that examined whether 2- to 3-year-olds benefit from an active choice of learning materials in a tablet-based word learning task. Children were assigned to either the active condition, where they could select the novel objects they wish to learn about, or the yoked passive condition, where they were presented with the objects chosen by their age-matched active peers. While children in the passive condition outperformed those in the active condition in terms of accuracy in both Study 1A and Study 1B, Study 1B found no differences in their recognition of the novel word–referent associations on a more implicit looking time measure. These results suggest that there may be performance costs associated with active tasks designed as in the present studies and there may not always be systematic benefits associated with active learning in touchscreen-based word learning tasks. Thus, the present studies add to the evidence that educational apps need to be evaluated before release: while children may benefit from interactive apps under certain conditions, task (and app) design and requirements need to consider factors that may detract from successful performance. In the next chapter, two studies aiming to address questions related to the assessment of early word knowledge are presented.

CHAPTER 5. ASSESSING EARLY WORD KNOWLEDGE WITH TABLETS

This chapter describes two studies conducted to address research questions 3 and 4, that is, to explore the viability of tablets in assessing young children's word knowledge (Study 2) and to further develop short-form versions of the MacArthur–Bates Communicative Development Inventories (CDI) to more efficiently estimate early word knowledge (Study 3). Study 2 is available as

> **Lo, C. H.**[16], Rosslund, A.[16], Chai, J. H., Mayor, J., & Kartushina, N. (2021). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy*. Advance online publication. https://doi.org/10.1111/infa.12401

Study 3 is available as

> Chai, J. H.[16], **Lo, C. H.**[16], & Mayor, J. (2020). A Bayesian-inspired item response theory–based framework to produce very short versions of MacArthur–Bates Communicative Development Inventories. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3488–3500. https://doi.org/10.1044/2020_JSLHR-20-00361[17]

These papers have been adapted to suit the style of this thesis.

5.1 Study 2

5.1.1 Introduction

Historically, studies of early language development involved longitudinal observations of children's spontaneous behaviours when they are interacting with their parents, an experimenter, or a clinician (e.g., Clark, 1974). Despite this method's undeniable appeal of ecological validity, the process of collecting,

[16]Both authors share co-first authorship.

[17]Permission to reprint has been granted by American Speech-Language-Hearing Association.

transcribing, and analysing spontaneous language samples is labour-intensive and time-consuming.

To go beyond these drawbacks, researchers have turned to a more indirect method, that is, parent report, that provides "quick and easy" data on children's communicative–linguistic development. As detailed in the literature review, parent reports systematically utilise parents' extensive experience with their children, and thus allow for the collection of data that is not only more extensive than what is attainable from brief laboratory or clinical sessions, but may also be more representative of children's abilities (Fenson, Pethick, et al., 2000). Furthermore, the application of parent reports (e.g., CDIs) in cross-linguistic studies has provided invaluable insight into children's early language development (e.g., Bleses et al., 2008b; Braginsky et al., 2019; Frank et al., 2021), while other studies have evinced predictive relationships between early vocabulary and subsequent academic outcomes (e.g., Bleses et al., 2016; Duff, Reen, et al., 2015; Morgan et al., 2015).

Yet, concerns have been raised regarding the exclusive use of parent reports for the assessment of comprehension rather than production, especially at the earlier ages, since parents can at best infer comprehension based on children's non-verbal responses to language (Feldman et al., 2000; Houston-Price et al., 2007; Tomasello & Mervis, 1994). In addition, even when parental accuracy is high, parent reports may still be unstable over time at the "item-level" due to children's rapid gains in vocabulary during the second year of life, and may have implications when parent reports are used as the basis for vocabulary goal selection (e.g., in clinical settings; Yoder et al., 1997). For these reasons, the use of supplemental measures to parent reports is encouraged (Dale et al., 2003; Fenson et al., 1993).

A direct language measure (i.e., structured tests) can serve both as a convergent and a supplemental measure of parent reports. While many structured tests, such as the Peabody Picture Vocabulary Test (PPVT; Dunn, 2018) and the Expressive Vocabulary Test (EVT; Williams, 2018), are available to assess young children's vocabulary knowledge, direct measures that are

appropriate for assessing children below 2 years of age remain scarce, due to the inherent difficulty in maintaining children's interest and attention (Friend & Keplinger, 2003) as well as behavioural non-compliance (Kaler & Kopp, 1990). As the review of the literature suggests, whereas looking-based measures, such as the Intermodal Preferential Looking Paradigm (IPLP; Golinkoff et al., 1987; Hirsh-Pasek & Golinkoff, 1996) and the looking-while-listening procedure (LWL; Fernald et al., 2006; Fernald et al., 1998), have been successfully used with infants as young as 4-months-old by eliminating the need for a volitional response (Golinkoff et al., 2013), the passive and repetitive nature of such measures may quickly lead to boredom among older children, thus making an extensive assessment impracticable. The Computerized Comprehension Task (CCT; Friend & Keplinger, 2003), on the other hand, is a reliable and valid touchscreen-based measure designed specifically for assessing comprehension among children between 16 and 24 months of age and has been shown to be effective in maintaining children's attention as well as improving compliance (Friend & Keplinger, 2003, 2008; Friend et al., 2012; Friend & Zesiger, 2011; Hendrickson et al., 2015; Poulin-Dubois et al., 2013).

Following the approach of the CCT—in providing an engaging direct language assessment—the present study explores the viability of tablets in assessing young children's word comprehension by means of a word recognition task. The purpose of doing this is twofold. First, despite tablets and apps being increasingly commonplace among children of all ages, the use of tablet-based assessments has been primarily limited to adults and older children. Given that tablets are easy to operate even for the youngest children and additionally, given children's increasing proficiency with tablets (Abdul Aziz et al., 2014; Marsh et al., 2015), there is a need to examine how such devices can be used most effectively to collect child language data. Neumann et al. (2019), for instance, demonstrated that a tablet-based assessment could provide a valid and reliable measure of early literacy skills, at least among the older children ($M_{\text{age}} = 4.65$ years) tested in their study. Twomey et al. (2018) further showed

that children as young as 24-months-old were able to complete a tablet-based assessment of early cognitive functions.

Second, compared to traditional paper-and-pencil tests, tablet-based assessments provide a testing situation that is more engaging and motivating. While the CCT offers the same advantage, the assessment is typically administered in laboratories, where screens are often mounted on a wall or placed on a desk and thus require full arm movements, which may in turn, lead to fatigue in longer sessions (Frank et al., 2016). In contrast, tablet-based assessments require only minimal motor movements and are much more portable due to the small form factor of tablets.

In order to explore the viability of using a tablet-based measure in assessing early word comprehension, the present study employed a two-alternative forced choice (2AFC) word recognition paradigm (similar to the CCT) with Norwegian children aged between 18 and 20 months. In doing so, comparisons can be made with parent report measures of comprehension (obtained using the Norwegian adaptation of the CDI–Words and Gestures [CDI–WG], which covers development up to 20 months of age; Simonsen et al., 2014). As the CCT is only available in three languages (i.e., English, Spanish, and French), lexical items were selected from the Norwegian adaptation of the CDI–WG with varying levels of difficulty (defined based on the normative data). Within each trial, children saw on a screen two images: one representing the lexical target, and the other representing the distractor. In contrast to the CCT, in which only semantically related item pairs were used, the current design additionally examined the role of semantic relatedness on children's performance in the word recognition task, by pairing the lexical target with a distractor belonging to a different semantic category (e.g., car and cat) and with another distractor belonging to the same semantic category (e.g., car and aeroplane). Previous research has shown that early word representations are (semantically) coarse and children use a number of cues to disambiguate words. For instance, at 6 months of age, infants typically fail in disambiguating semantically/functionally related items (Bergelson & Aslin, 2017a) and at 8

months of age, they struggle to disambiguate items matched for frequency in child-directed speech (Kartushina & Mayor, 2019). Although word-referent associations undergo a progressive development through learning, they are seemingly still fragile by the end of the second year. At 18 to 24 months, children fail to disambiguate items that are both perceptually and semantically related (Arias-Trejo & Plunkett, 2010), as the presence of a perceptually and semantically similar distractor increases the burden of visual discrimination and feature overlap. In line with this study, it was expected that, children, in the present study, would be more accurate in semantically unrelated than related trials. Based on previous work using the CCT (e.g., Friend & Keplinger, 2003, 2008), accuracy was also expected to mirror the a priori difficulty levels, with accuracy decreasing with increasing difficulty. Finally, if parent reports are an accurate predictor of children's word knowledge, a positive relationship between parent-reported comprehension and children's accuracy in word recognition would be expected.

5.1.2 Method

5.1.2.1 Design

The present study used a within-subjects design. Children's comprehension of 24 lexical items of three levels of difficulty (easy, moderately difficult, and difficult; see Section 5.1.2.3.1 below) was assessed using a tablet-based 2AFC word recognition task. Lexical targets were assessed under two conditions: semantically related (i.e., the lexical target was presented with a distractor from the same semantic category) and semantically unrelated (i.e., the lexical target was presented with a distractor from a different semantic category).

5.1.2.2 Participants

Parents of 49 primarily monolingual Norwegian children (aged between 18 and 20 months) from the Greater Oslo Region, Norway, were contacted through

one of four ways: social media, leaflets distributed in a kindergarten, postal mailing lists, or email lists. After consenting to participate in the study, parents completed the Norwegian adaptation of the CDI–WG (Simonsen et al., 2014) online within one week prior to the study so that the current estimates of their child's vocabulary size could be obtained.

All children recruited were full-term at birth, had no hearing or visual impairments, and had Norwegian as their native language. Children participated in the study in one of three settings: the BabyLing laboratory, a municipal kindergarten, or online (i.e., at children's own homes).[18] In both the laboratory and the kindergarten settings, children were tested by an experimenter, whereas online, children were tested by their parents.[19] Thus, for simplicity, both the laboratory and kindergarten samples were categorised under the *lab* setting ($n = 21$; 16 females, 5 males), and the online samples, the *online* setting ($n = 28$; 15 females, 13 males). Mean age, age range, and standard deviation for each setting are detailed in Table 5.1. An additional 11 participants had to be excluded for failing to complete the task ($n = 7$; 2 lab, 5 online) and for attempting the task more than once ($n = 4$; 0 lab, 4 online). The study was reviewed and approved by the ethics committee of the Department of Psychology, University of Oslo and by the Norwegian Centre for Research Data.

Table 5.1

Age Mean, Standard Deviation, and Range

Setting	n	M_{age} (months)	SD_{age} (months)	$\text{Range}_{\text{age}}$ (months)
Lab	21	19.29	0.60	17.91–20.30
Online	28	19.63	0.63	18.60–20.60

[18]Data was initially collected in the lab and kindergarten. Due to the COVID-19 pandemic–related lockdown in Norway (Klesty & Fouche, 2020), data collection proceeded online.
[19]Parents consented to not to interfere with the task or influence their child's responses.

5.1.2.3 Apparatus and Materials

The study was conducted via a web application.[20] In the lab setting, a Samsung Galaxy Tab S4 was used run the study, whereas in the online setting, parents' own touchscreen devices were used. The Norwegian adaptation of the CDI–WG (Simonsen et al., 2014) was used as a measure of vocabulary size.

5.1.2.3.1 Lexical Items

Four highly familiar lexical items were selected for the familiarisation phase: "ball" [ball], "hus" [house], "sko" [shoe], and "tre" [tree]. For the test phase, a total of 24 lexical items were selected. Each lexical target was assessed twice, by pairing its referent with semantically related and unrelated referents as distractors. Item pairs varied in difficulty (defined a priori on the basis of the Norwegian CDI–WG normative data for 20 month-olds; Frank et al., 2017; Simonsen et al., 2014) and were comprised of an equal number of *easy* (comprehended by more than 80% of the normative sample), *moderately difficult* (comprehended by 40–80% of the normative sample), and *difficult* (comprehended by less than 40% of the normative sample) item pairs. Within each level of difficulty, there was also an equal representation of animate and inanimate referents. The list of item pairs is provided in Table 5.2.

5.1.2.3.2 Visual and Auditory Stimuli

To remove potential biases due to familiarity effects (from assessing the same item twice), visual stimuli for the test phase included 48 images of prototypical referents for the 24 lexical items assessed (i.e., two images for each item). The set of images used can be found in Appendix F (see also Appendix G for the images used in the familiarisation phase). Within each item pair, the side (left or right) on which a referent appeared was counterbalanced. All auditory stimuli used were recorded by a female native speaker of Norwegian in child-directed speech.

[20] Programmed using e-Babylab (Chapter 3).

Table 5.2

Item Pairs

Difficulty level	Semantically related	Semantically unrelated
Easy	bil [car] - fly [aeroplane]	hest [horse] - banan [banana]
	eple [apple] - banan [banana]	hund [dog] - fly [aeroplane]
	hest [horse] - ku [cow]	katt [cat] - bil [car]
	hund [dog] - katt [cat]	ku [cow] - eple [apple]
Moderate	elefant [elephant] - tiger [tiger]	elefant [elephant] - saks [scissors]
	lastebil [truck] - tog [train]	løve [lion] - tog [train]
	saks [scissors] - blyant [pencil]	sjiraff [giraffe] - lastebil [truck]
	sjiraff [giraffe] - løve [lion]	tiger [tiger] - blyant [pencil]
Difficult	elg [moose] - pingvin [penguin]	elg [moose] - pasta [pasta]
	gås [goose] - ugle [owl]	gås [goose] - shorts [shorts]
	pasta [pasta] - sukkertøy [candy]	pingvin [penguin] - sukkertøy [candy]
	shorts [shorts] - glidelås [zipper]	ugle [owl] - glidelås [zipper]

5.1.2.4 Procedure

The study began with an introductory phase, followed by a familiarisation phase and a test phase.

5.1.2.4.1 Introductory Phase

During the introductory phase, a smiley face was presented at the centre of the screen with an introductory audio "Hei! Har du lyst til å spille?" [Hi! Do you want to play?] to attract participants' attention. In order to proceed to the familiarisation phase, the experimenter/parent had to tap on the "Next" button at the bottom right corner of the screen (see Figure 5.1 for a screenshot).

Figure 5.1

Screenshot of the Introductory Phase

5.1.2.4.2 Familiarisation Phase

The familiarisation phase consisted of four 2AFC trials to: (a) ensure that participants understood the context of the task and (b) familiarise them

with the tapping paradigm. In each trial, participants were presented with a pair of highly familiar objects (placed on the left and right sides of the screen respectively) and prompted to tap on the referent for the heard lexical target X embedded in the carrier phrase "Kan du trykke på X?" [Can you touch the X?] Tapping was disabled for the first 2000 ms from the onset of the trial to prevent impulsive responses during the audio prompt that lasted between 1500 and 2000 ms. When tapping was enabled, participants had 8000 ms to respond until the subsequent trial was presented.

5.1.2.4.3 Test Phase

Before the test phase began, a smiley face was again presented at the centre of the screen, accompanied by an audio with an encouraging phrase "Da forsetter vi!" [Let's continue!] The experimenter/parent had to tap on the "Next" button to begin the test phase.

The test phase consisted of 48 2AFC trials, in which each lexical target was assessed twice (paired with either a semantically related distractor or a semantically unrelated distractor). In each trial, participants were presented with an item pair (see Table 5.2) and prompted to tap on the referent for the heard lexical target X (see carrier phrase from the familiarisation phase). Each item pair was presented twice so that each item within the pair served as both a target and a distractor. As with the familiar trials, tapping was disabled for the first 2000 ms of the trial (to prevent participants from responding before the end of the audio prompt that lasted between 1500 ms and 2000 ms), after which participants were given 8000 ms to respond until the subsequent trial was presented. Trials were presented in a random order, with three breaks interspersed throughout the test phase. During each break, a smiley face was presented in the same manner as before, accompanied by one of the following encouraging phrases: (a) "Da forsetter vi!" [Let's continue!], (b) "Nå går vi videre!" [Now, we move on!], (c) "Da har vi den neste!" [Then, we have the next (one)!], and (d) "Da er du nesten ferdig! Bra!" [You're almost done! Good!] In order to continue with the test, the experimenter/parent had to also tap on the

"Next" button at the bottom right corner of the screen. Upon completion of the test phase, the smiley face was once again presented, accompanied by an audio with the phrase "Nå er du ferdig! Kjempebra!" [Now you're done! Great!]

5.1.3 Results

The results are organised around three central questions. First, potential differences between data collected online and in-lab were considered. Second, the influence of semantic relatedness and difficulty of item pairs on children's motivation to produce a response as well as on their performance in the word recognition task were examined. Finally, the convergent relation between children's performance and parent report (CDI–WG) was assessed. In accordance with previous work using the CCT (Friend & Keplinger, 2003; Friend et al., 2012), missing responses (i.e., trials in which the child did not produce a response) were treated as errors of comprehension.

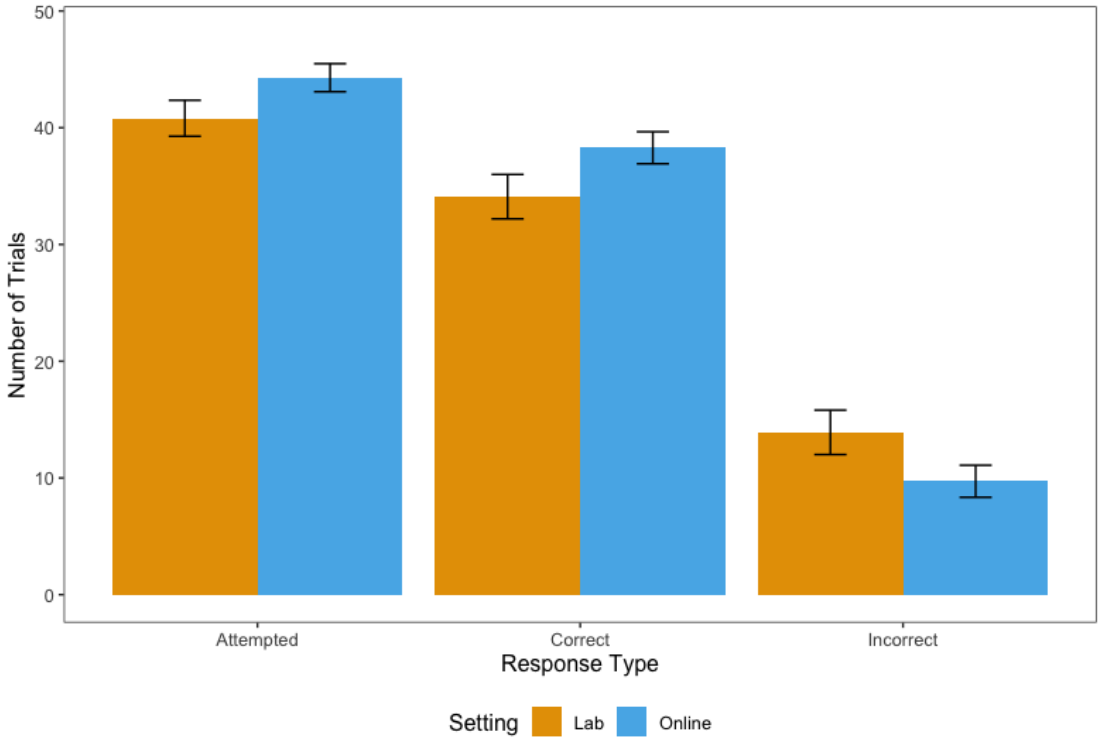
5.1.3.1 Trials Attempted

The number of trials in which a tap response was produced, regardless of whether the response was correct (i.e., tap on target) or incorrect (i.e., tap on distractor), was used as a measure of children's motivation to produce a response during the word recognition task. Results from a Welch's t-test indicated that children who were tested online ($M = 44.286, SD = 6.359$) and those who were tested in the laboratory ($M = 40.810, SD = 7.061$) did not differ significantly in the number of trials attempted; $t(40.601) = -1.779, p = .083$ (see Figure 5.2).

To assess whether children's motivation differed across semantic relatedness and difficulty of the trials, a binomial generalised linear mixed-effects model (GLMM) with a logit link function was fitted and analysed using the `mixed()` function from the *afex* package (Singmann et al., 2020), which relies on the *lme4* package (D. Bates et al., 2015) for model fitting. The model included semantic relatedness (related, unrelated), difficulty (easy, moderately difficult, difficult), children's age (in months), and the interaction between semantic

Figure 5.2

Attempted, Correct, and Incorrect Trials Across Different Settings



relatedness and difficulty as fixed effects, as well as participant and selected object as random intercepts.[21] Both semantic relatedness (-1: unrelated; 1: related) and difficulty (-1: easy; 1: moderately difficult, difficult) were sum-coded, whereas age was centred on the mean. To determine a model with a parsimonious random effect structure (Matuschek et al., 2017), the forward "best-path" approach (D. J. Barr et al., 2013) was used to test random slopes for inclusion ($\alpha = 0.20$). As none of the random slopes fell below the inclusion criterion, the random-intercepts-only model was retained:

$$\text{Attempted} \sim \text{Relatedness} * \text{Difficulty} + \text{Age} + (1|\text{Participant}) + (1|\text{Object})$$

The results are detailed in Table 5.3, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons, with p-values adjusted using the Tukey method, were conducted using the `pairs()` function in the *emmeans* package (Lenth, 2020).

As shown in Table 5.3, there were significant main effects of trial difficulty and age, with the number of trials attempted increasing with age. No significant main effect of semantic relatedness was found; neither did semantic relatedness interact with difficulty. Results from the follow-up tests indicated that children attempted significantly more easy than difficult trials ($\beta = 0.556, SE = 0.186, z = 2.995, p = .008$), while no such difference was found between easy and moderately difficult trials ($\beta = 0.363, SE = 0.189, z = 1.917, p = .134$) as well as moderately difficult and difficult trials ($\beta = 0.193, SE = 0.176, z = 1.096, p = .517$; see also Figure 5.3).

5.1.3.2 Correct Trials

Results from a Welch's t-test indicated that there was no statistically significant difference between children who were tested online ($M = 38.286, SD = 7.262$) and those who were tested in the laboratory ($M = 34.095, SD = 8.717$) in terms of the number of trials in which they

[21] The inclusion of setting (i.e., online vs. lab) and sex as fixed effects in the model did not change the conclusions and were thus omitted.

Table 5.3

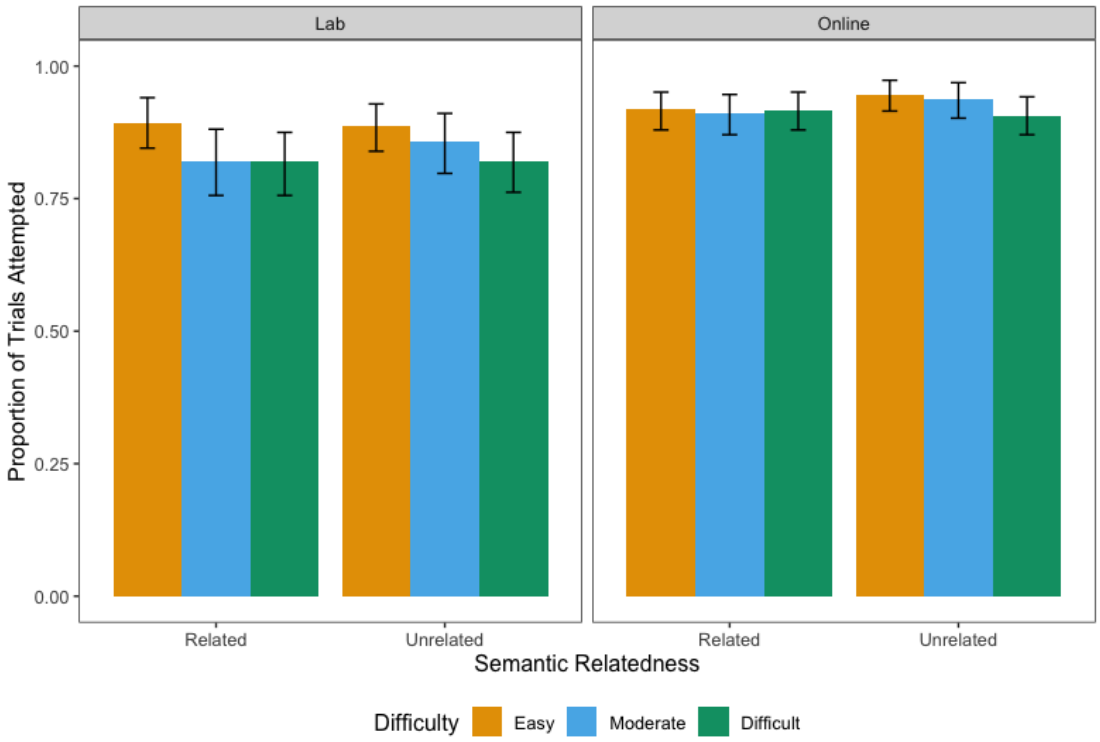
GLMM Results for Trials Attempted

	Model summary			Model comparison		
	β	*SE*	z	χ^2	*df*	p
Intercept	3.080	0.281	10.956	103.539	1	<.001***
Relatedness	-0.087	0.075	-1.163	1.355	1	.244
Difficulty				8.516	2	.014*
Moderate	-0.057	0.105	-0.542			
Difficult	-0.249	0.103	-2.432			
Age	0.949	0.395	2.402	5.686	1	.017*
Relatedness:Difficulty				1.618	2	.445
Relatedness:Moderate	-0.106	0.105	-1.006			
Relatedness:Difficult	0.116	0.102	1.136			

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 5.3

Proportion of Trials Attempted by Semantic Relatedness, Difficulty, and Setting



correctly identified the target referent; $t(38.508) = -1.787, p = .082$ (see Figure 5.2).

To assess whether children's accuracy differed across semantic relatedness and difficulty of the trials, a binomial GLMM with a logit link function was again fitted and analysed. The model included the same fixed effects as the previous model (i.e., semantic relatedness, difficulty, age, and the interaction between semantic relatedness and difficulty) as well as the same random intercepts (i.e., participant and selected object), with by-participant adjustment to the slope of difficulty:[22]

$$\text{Accuracy} \sim \text{Relatedness} * \text{Difficulty} + \text{Age} + (1 + \text{Difficulty}|\text{Participant}) + (1|\text{Object})$$

The results are detailed in Table 5.4, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons were conducted with p-values adjusted using the Tukey method.

As shown in Table 5.4, there were significant main effects of semantic relatedness, difficulty, and age. Specifically, children responded with higher accuracy in semantically unrelated than related trials. Children's accuracy also increased significantly with age. No significant interaction effect between semantic relatedness and difficulty was found however. Results from the follow-up tests indicated that children were significantly more accurate in easy trials relative to both moderately difficult ($\beta = 0.523, SE = 0.183, z = 2.861, p = .012$) and difficult trials ($\beta = 1.113, SE = 0.164, z = 6.799, p < .001$). Children were also significantly more accurate in moderately difficult than difficult trials ($\beta = 0.590, SE = 0.150, z = 3.924, p < .001$; see also Figure 5.4).

[22]The inclusion of setting (i.e., online vs. lab) and sex as fixed effects in the model did not change the conclusions and were thus omitted.

Table 5.4

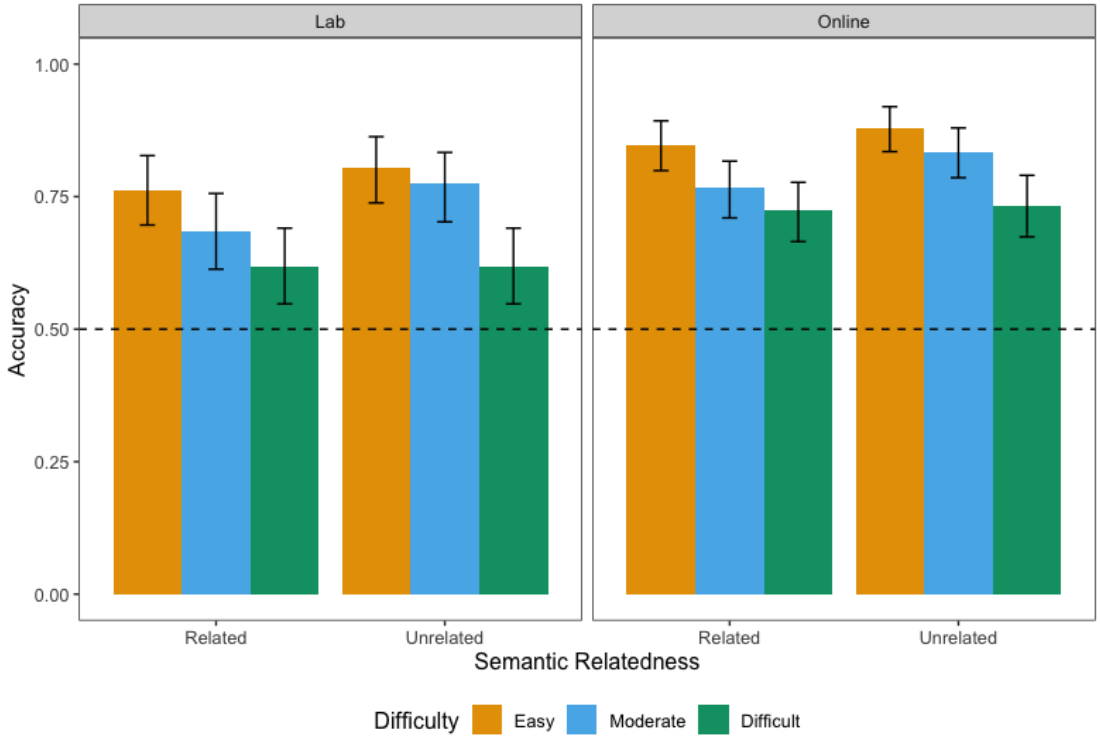
GLMM Results for Accuracy

	Model summary			Model comparison		
	β	*SE*	z	χ^2	*df*	p
Intercept	1.438	0.143	10.038	56.979	1	<.001***
Relatedness	-0.141	0.054	-2.624	6.782	1	.009**
Difficulty				36.405	2	<.001***
Moderate	0.022	0.097	0.229			
Difficult	-0.568	0.085	-6.660			
Age	0.537	0.193	2.779	7.233	1	.007**
Relatedness:Difficulty				3.887	2	.143
Relatedness:Moderate	-0.114	0.076	-1.511			
Relatedness:Difficult	0.127	0.071	1.785			

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 5.4

Accuracy by Semantic Relatedness, Difficulty, and Setting



Note. Dashed line represents chance (.50).

5.1.3.3 Convergent Validity

At the summary level, children's receptive vocabulary size, as measured by the CDI–WG, and their overall accuracy in the word recognition task significantly correlated in both unrelated ($r_{(47)} = .631, p < .001$) and related trials ($r_{(47)} = .603, p < .001$). Partialling out the effect of age further revealed that children's receptive vocabulary size accounted for a significant proportion of unique variance in their recognition accuracy, beyond that accounted for by their age in both unrelated ($r_{(46)} = .593, p < .001, R^2 = .352$) and related trials ($r_{(46)} = .538, p < .001, R^2 = .289$).

To explore the consistency between children's responses and parent-reported comprehension on the test items (i.e., parent–child agreement), item-level agreement was calculated (see Table 5.5) and a binomial GLMM with a logit link function was fitted. The model included semantic relatedness, difficulty, age, and the interaction between semantic relatedness and difficulty as fixed effects. Both semantic relatedness (-1: unrelated; 1: related) and difficulty (-1: easy; 1: moderately difficult, difficult) were sum-coded, whereas age was centred on the mean. Random intercepts included participant and selected object, with by-participant adjustments to the slopes of semantic relatedness, difficulty, and their interaction:[23]

$$\begin{aligned}\text{Agreement} \sim \text{Relatedness} * \text{Difficulty} + \text{Age} \\ + (1 + \text{Relatedness} * \text{Difficulty}|\text{Participant}) + (1|\text{Object})\end{aligned}$$

The GLMM results are detailed in Table 5.6, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons were conducted with p-values adjusted using the Tukey method.

Overall, as shown in Table 5.5, there was good item-level agreement between parent reports and children's responses, although this attenuated with increasing item difficulty. Results from the GLMM indicated that semantic relatedness, difficulty, as well as the interaction between semantic relatedness

[23]The inclusion of setting (i.e., online vs. lab) and sex as fixed effects in the model did not change the conclusions and were thus omitted.

and difficulty significantly predicted parent–child agreement, while age was not a significant predictor (see also Figure 5.5). The follow-up tests revealed that parent–child agreement was significantly higher in semantically unrelated than related easy trials ($\beta = 0.795, SE = 0.299, z = 2.662, p = .008$), but no significant differences were found across the different semantic conditions in the moderately difficult ($\beta = 0.253, SE = 0.169, z = 1.495, p = .135$) and difficult trials ($\beta = -0.166, SE = 0.164, z = -1.014, p = .311$).

Table 5.5

Item-Level Agreement Between Parent Report and Child Performance

Difficulty level	Semantically related	Semantically unrelated	Overall
Easy	.781	.827	.804
Moderate	.615	.661	.638
Difficult	.564	.538	.551
Overall	.653	.675	.664

To further examine whether item-pair comprehension status (i.e., whether the target or the distractor label was known or not known by the child as indicated by parental responses on the CDI–WG) was an accurate predictor of children's performance in the word recognition task, another binomial GLMM with a logit link function was fitted, with semantic relatedness, difficulty, item-pair comprehension status, age, and the interaction between semantic relatedness and difficulty as fixed effects. Semantic relatedness (-1: unrelated; 1: related), difficulty (-1: easy; 1: moderately difficult, difficult), and item-pair comprehension status (-1: both unknown; 1: both known, target known only, distractor known only) were sum-coded, whereas age was centred on the mean.

Table 5.6

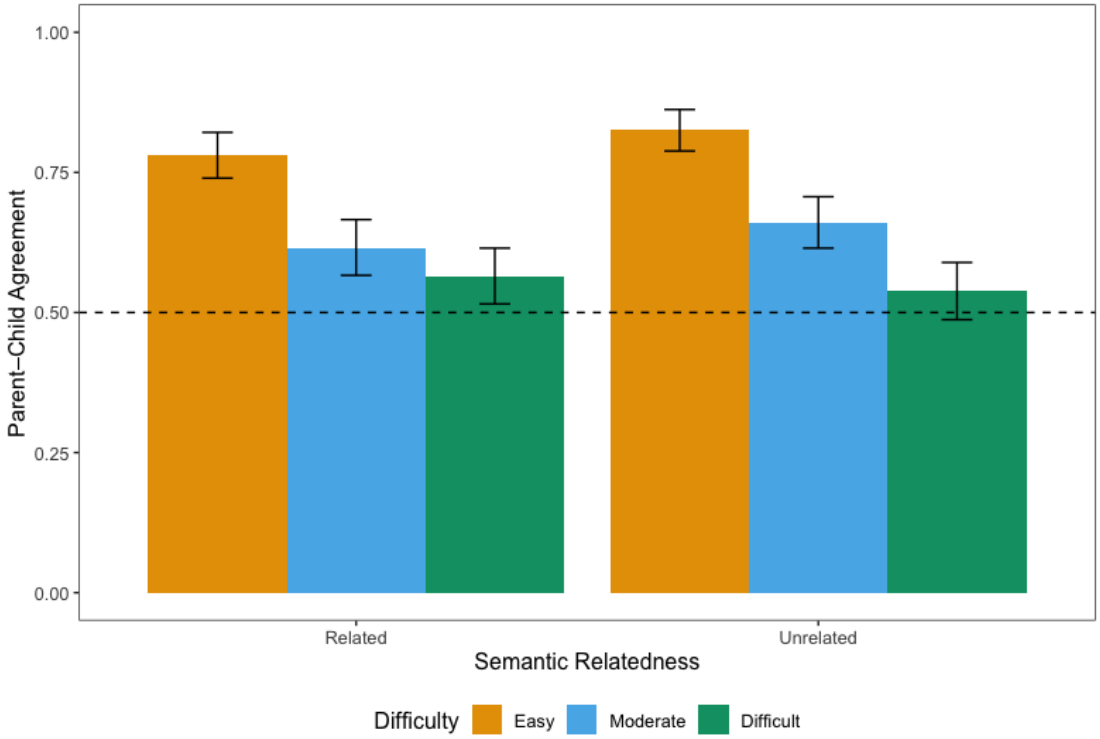
GLMM Results for Parent–Child Agreement

	Model summary			Model comparison		
	β	*SE*	z	χ^2	*df*	p
Intercept	0.921	0.163	5.663	68.207	1	<.001***
Relatedness	-0.147	0.066	-2.237	5.436	1	.020*
Difficulty				21.564	2	<.001***
Moderate	-0.240	0.168	-1.423			
Difficult	-0.752	0.182	-4.134			
Age	0.074	0.153	0.486	0.218	1	.641
Relatedness:Difficulty				9.994	2	.007**
Relatedness:Moderate	0.020	0.082	0.249			
Relatedness:Difficult	0.230	0.076	3.030			

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 5.5

Parent–Child Agreement by Semantic Relatedness and Difficulty



Note. Dashed line represents chance (.50).

Random intercepts included participant and selected object, with by-participant adjustment to the slope of difficulty:[24]

$$\text{Accuracy} \sim \text{Relatedness} * \text{Difficulty} + \text{Pair Comprehension} + \text{Age} + (1 + \text{Difficulty}|\text{Participant}) + (1|\text{Object})$$

The results are detailed in Table 5.7, with χ^2 statistics and p-values obtained using Likelihood Ratio Tests. Follow-up pairwise comparisons were conducted with p-values adjusted using the Tukey method.

As shown in Table 5.7, parent-reported item-pair comprehension was a significant predictor of children's performance, along with semantic relatedness, difficulty, and age. No significant interaction effect between semantic relatedness and difficulty was found. Results from the follow-up tests indicated that children were significantly less accurate when both target and distractor were reported as unknown compared to when both were known ($\beta = -0.628, SE = 0.190, z = -3.300, p = .005$) and when only the target was known ($\beta = -0.769, SE = 0.196, z = -3.923, p < .001$). No significant differences were found in other cases: (a) both known and target known only ($\beta = -0.141, SE = 0.195, z = -0.725, p = .887$); (b) both known and distractor known only ($\beta = -0.284, SE = 0.184, z = 1.539, p = .414$); (c) target known only and distractor known only ($\beta = 0.425, SE = 0.205, z = 2.070, p = .163$); (d) distractor known only and both unknown ($\beta = -0.344, SE = 0.186, z = 1.846, p = .252$; see also Figure 5.6).

5.1.4 Discussion

In the interest of developing a performance-based measure of comprehension during the second year of life that addresses the need for a convergent and supplemental measure of parent reports, while taking into account young children's non-compliance and limited attention capabilities (as in

[24]The inclusion of setting (i.e., online vs. lab) and sex as fixed effects in the model did not change the conclusions and were thus omitted.

Table 5.7

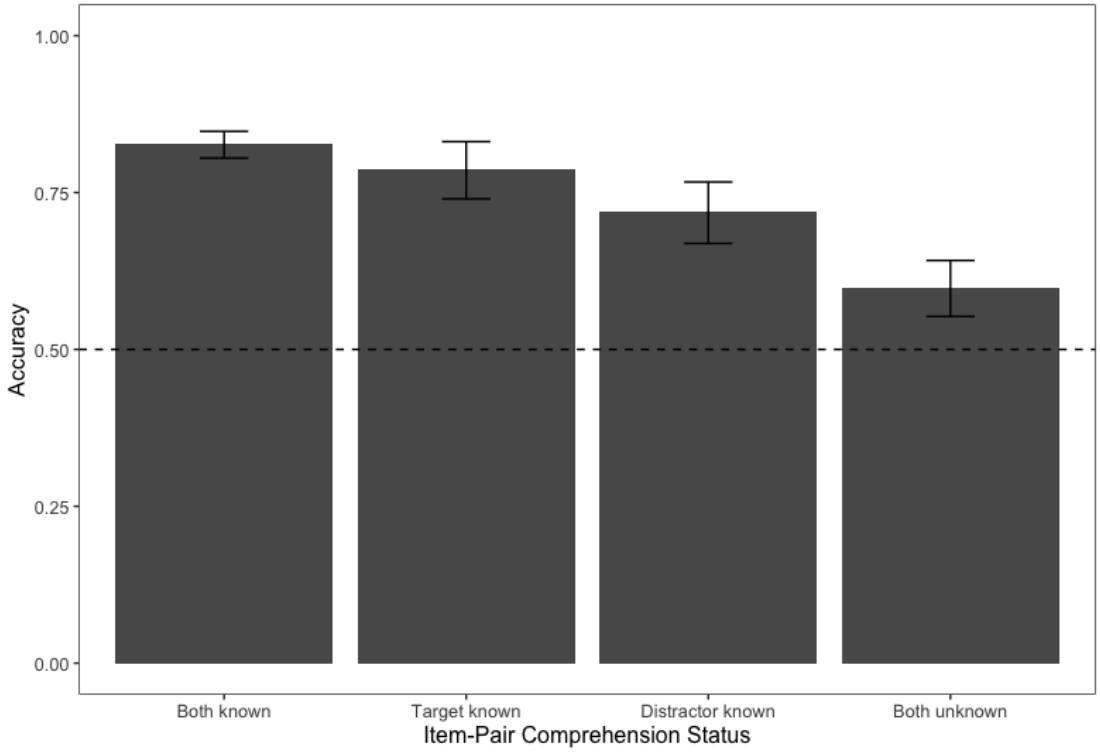
GLMM Results for Accuracy (With Parent-Reported Comprehension as Predictor)

	Model summary			Model comparison		
	β	*SE*	*z*	χ^2	*df*	*p*
Intercept	1.402	0.144	9.749	58.245	1	<.001***
Relatedness	-0.139	0.054	-2.588	6.586	1	.010*
Difficulty				14.702	2	<.001***
Moderate	0.007	0.098	0.068			
Difficult	-0.403	0.107	-3.776			
Pair comprehension				18.108	1	<.001***
Both known	0.193	0.114	1.685			
Target known	0.334	0.125	2.667			
Distractor known	-0.091	0.117	-0.778			
Age	0.511	0.181	2.817	7.428	1	.006**
Relatedness:Difficulty				4.141	2	.126
Relatedness:Moderate	-0.120	0.076	-1.581			
Relatedness:Difficult	0.132	0.072	1.832			

*$p < .05$, **$p < .01$, ***$p < .001$

Figure 5.6

Accuracy by Parent-Reported Item-Pair Comprehension Status



Note. Dashed line represents chance (.50).

Friend & Keplinger, 2003), the present study explored the viability of a tablet-based 2AFC word recognition task in assessing early word comprehension.

Children aged between 18 and 20 months were tested—either in the lab setting by an experimenter or online (i.e., at home) by their parents—on their comprehension of 24 lexical items selected from the Norwegian CDI–WG (Simonsen et al., 2014). During the task, children were asked to identify the referent for the lexical target presented alongside a distractor. Target–distractor pairs were manipulated such that each lexical target was paired once with a semantically related distractor and once with a semantically unrelated distractor. Item pairs also varied in three levels of difficulty (defined based on the Norwegian CDI–WG normative data for age-matched children).

Both the analyses on the number of trials attempted (regardless of whether the response was correct or incorrect) as well as the number of trials in which children provided a correct response revealed no significant differences between the online and lab samples, suggesting that children were equally motivated to produce a response in the task and that neither setting led to better or poorer performance. This demonstrates that remote data collection among young children with fully automatised tasks can be as efficient and reliable as in situ laboratory-based assessments. Remote administration is not only an important enabler of data collection during this time of the COVID-19 pandemic, but also provide a promising avenue for collecting developmental data with increased speed, lowered costs, and potentially, an improved sample diversity by reaching to a wider socio-demographic background than traditional laboratory-based research (Sheskin et al., 2020).

Overall, in line with Friend and Keplinger (2008), children attempted significantly more easy than difficult trials. Older children also attempted significantly more trials than younger children. Together, these findings suggest that children were responding non-randomly and bolster the support for the notion that non-responses represent children's true inability to map the lexical target to its referent, rather than their non-compliance or the lack of motivation,

while incorrect responses can be taken as evidence of partial word knowledge, and correct responses, robust word knowledge (Hendrickson et al., 2015).

With regard to the accuracy measure, children demonstrated above-chance performance throughout the task. Congruent with previous work (Friend & Keplinger, 2003, 2008), children's performance was consistent with the a priori difficulty categorisation, as their best performance was obtained for easy trials, and their worst performance, for difficult trials. As would be expected from the literature, older children also performed with greater accuracy relative to younger children.

Examining the role of semantic relatedness, it was found that children displayed more robust recognition in semantically unrelated than related trials, suggesting that, and similar to research in younger children (Bergelson & Aslin, 2017a), semantic relatedness between the target and the distractor triggered competition effects in referent selection. Although there is evidence that early word representations are semantically more specified by 18 to 20 months of age (Bergelson & Aslin, 2017b), they might still be lacking representational specificity (Arias-Trejo & Plunkett, 2010). In the present study, poorer recognition performance on some related trials could also be attributed to the increased burden of visual discrimination and feature overlap, as shown with 18- to 24-month-olds in Arias-Trejo and Plunkett (2010). For instance, in the "goose–owl" pair, both goose and owl are birds and have wings, feather, and a beak. It is also likely that children, upon hearing the lexical target, co-activated related (and thus, competing) word referents, which subsequently interfered with their lexical decision about the target. Such interference has been reported even among older children, between 3 and 9 years of age, as they took longer to provide a correct response in a visual search task when a related distractor was present than when an unrelated distractor was present (Vales & Fisher, 2019).

Comparing between children's recognition accuracy and their receptive vocabulary size as measured by the CDI–WG, significant and moderate correlations (comparable to that achieved with the CCT; Friend & Keplinger, 2008) were found across both semantic conditions, evincing acceptable

convergent validity of the word recognition task employed in the present study. Consistent with the CCT (Friend et al., 2012; Friend & Zesiger, 2011), there was also good item-level agreement between children's responses and parent reports across both semantic conditions, with easy items having the highest agreement and difficult items having the lowest agreement. The results further indicated that parent–child agreement was significantly higher in semantically unrelated than related trials, although this was only limited to easy items. This discrepancy suggests that parents' inference on their child's word comprehension is not solely based on evidence of their child's true ability to comprehend the word, but rather on the confluence of both evidence of robust word knowledge (i.e., their child's true ability to comprehend the word) and evidence of partial word knowledge (i.e., their child's ability to respond appropriately when cued by the rich context in which the word is heard, or upon recognising the sound of the word; Friend et al., 2018; Houston-Price et al., 2007; Tomasello & Mervis, 1994). Restating the finding that children were less accurate in semantically related than unrelated trials, a performance-based measure that uses semantically related target–distractor pairs can potentially tap children's strong, rather than weak, word knowledge to supplement parent reports. Nevertheless, parent-reported item-pair comprehension (i.e., whether the target or distractor label was known or not known by the child) was found to be a significant predictor of children's recognition accuracy. Specifically, compared to trials where both the target and distractor were reported by parents as "not understood" on the CDI–WG, children were more likely to respond correctly in trials where either the target or both the target and distractor were reported as "understood", indicating that parents are adequate informants of their child's language abilities.

It is important to note that the CCT uses a set of carefully selected test items consisting of an equal representation of nouns, verbs, and adjectives, whereas the present study is limited in that only nouns were considered. Nevertheless, that such encouraging results were obtained is remarkable. With a more structured way of selecting test items, tablet-based word recognition tasks

may provide a useful measure of receptive vocabulary skills in the second year of life—and potentially serve as a supplemental and convergent measure of parent reports. In this respect, one could possibly utilise recent innovations made in the development of short-form versions of parent reports (e.g., Makransky et al., 2016; Mayor & Mani, 2019) in the selection of test items, that is, to administer short forms directly to children through the use of tablet-based tasks, thus effectively eliminating the tedious process of adapting an assessment to each language—as is the case with the CCT which, despite its utility, is only available in three languages at present. In addition, future work should consider further establishing the validity and reliability of the assessment, for instance, with children from more diverse backgrounds and varied abilities, while also taking into account other properties of distractor items (beyond semantic relatedness), such as perceptual and acoustic–phonetic similarities—and to take this further, extend the method to children's productive vocabulary. Together, these pave the way for an effective and efficient means to directly assess young children's word knowledge.

5.2 Study 3

5.2.1 Introduction

As noted in the literature review, CDIs are an effective, cost-efficient set of parent report instruments for assessing early language skills in children between 8 and 37 months of age (Fenson et al., 2007). Despite their many advantages, the applicability of CDIs, due to the sheer size of the forms, is greatly restricted in many research and clinical settings, especially when a rapid assessment is needed. Completion of the forms may also be daunting to parents having low literacy skills.

To address these drawbacks, various approaches have been taken, all of which aim to provide briefer alternatives to the full forms. These include the development of short-form versions of CDIs in different languages (e.g., Fenson, Pethick, et al., 2000; Rinaldi et al., 2019), the application of item response

theory (IRT)–based computerised adaptive testing (CAT) in CDI administrations (Makransky et al., 2016), and more recently, an approach that capitalises on CDI data from language-, sex-, and age-matched children on Wordbank (Frank et al., 2017) in estimating full CDI scores based on small subsets of items sampled from the full forms (Mayor & Mani, 2019). While showing great promise, each of these approaches comes with its own limitations. For instance, the short-form version of CDI–Words and Sentences (CDI–WS) may contribute to a ceiling effect after 27 to 28 months, not to mention the substantial amount of time and effort that is required to develop such forms for each language. Whereas Makransky et al.'s (2016) approach circumvents the need for "manually" adapting tests for each language, interpretation of the scores (i.e., latent ability) clearly suffers, since scores cannot be directly mapped back to the scores most typically used for CDIs (i.e., raw vocabulary sums or percentiles). Mayor and Mani's (2019) approach, on the other hand, provides readily interpretable scores, but scores are estimated based on random item samples, which can potentially be uninformative of a child's ability.

With the aim to develop a language-general approach that produces short forms in which items are selected to be maximally informative and subsequently derives CDI estimates that are on the same scale as the full CDI scores, the present study builds upon Mayor and Mani's (2019) approach to estimating full CDI scores, by implementing a principled selection of test items in place of the random selection. More specifically, CDIs were administered as IRT-based computerised adaptive tests, as in Makransky et al. (2016). Briefly, IRT refers to a family of mathematical models for estimating the measurement properties of test items and rests on two key assumptions: (a) *unidimensionality*—an examinee's response on a test item can be explained by latent traits or abilities; and (b) *monotonicity*—an examinee's ability and their response on a test item are related by a monotonically nondecreasing function (i.e., examinees having higher ability levels should never have a lower probability of responding correctly on a well-functioning test item than examinees having lower ability levels; Embretson & Reise, 2000; Hambleton, Swaminathan, et al., 1991). In IRT

models, each test item has a difficulty parameter which describes the point on the ability scale at which the probability of getting a correct response for a test item is .50. In other words, the more difficult an item, the higher the ability that is required for an examinee to have a 50% chance of providing a correct response. Additionally, each test item can have a discrimination parameter which determines the rate at which the probability of getting a correct response vary with different ability levels. An item with high discrimination is particularly useful for detecting subtle differences in examinees' abilities. By selecting test items on the basis of these item parameters, while taking into account the examinee's ability, not only can tests be shortened and tailored to each examinee, the risk of sampling minimally informative items can also accordingly be avoided.

To validate the present approach, real-data simulations were conducted using four CDI–WS versions for which their sample sizes on Wordbank vary: American English (a very large data set; Fenson et al., 2007), Danish (a large data set; Bleses et al., 2008a), Beijing Mandarin (a medium-sized data set; Tardif et al., 2009), and Italian (a small data set; M. C. Caselli & Casadio, 1995). This, in turn, helped to examine the possibility of applying IRT and CAT to different languages as well as to languages possessing few digitalised administrations on Wordbank. Validations were performed across different age groups and sexes.

The next section details the two main components of the present approach, that is, the IRT-based selection of test items (administered via CAT) and the estimation of full CDI scores based on Mayor and Mani's (2019) model. The results were then presented, followed by a discussion on the implications of the present findings for researchers and practitioners intending to use short forms for quick and cost-effective assessments of young children's vocabulary.

5.2.2 Method

5.2.2.1 IRT-Based Item Selection and Test Administration via CAT

The first step in selecting test items is to fit a two-parameter logistic IRT model to (prior) CDI data sampled from language-, sex-, and age-matched

children on Wordbank (accessed using the *wordbankr* package; Braginsky, 2018; Frank et al., 2017). For each item on the CDI, two parameters are assigned: a *discrimination* parameter and a *difficulty* parameter. Marginal maximum likelihood estimates of item parameters are computed with the expectation–maximisation algorithm (Bock & Aitkin, 1981) using the `mirt()` function from the *mirt* package (Chalmers, 2012).

Once the item parameters have been estimated, simulation of the CAT procedure is conducted using the `mirtCAT()` function from the *mirtCAT* package (Chalmers, 2016). The CAT procedure begins by administering items with maximum information. After each response, the ability parameter of the child, estimated using the weighted likelihood estimation method (Warm, 1989), is updated. Based on the child's estimated ability at each point (i.e., at each item administered) during the test, the CAT algorithm dynamically selects the subsequent item with maximum information, thereby allowing items that are more relevant (i.e., items that can inform maximally about the child's knowledge) to be administered. In doing so, items that are minimally informative (i.e., items that are too hard or trivially easy, given the child's estimated ability level) can also be omitted and this further translates into reduced administration times. In line with Makransky et al. (2016), the CAT procedure is set to terminate based on a fixed number of test items: 5, 10, 25, 50, 100, 200, 400, and the full CDI size. In the next step, the child's responses on the items administered in CAT are used to estimate their full CDI score.

5.2.2.2 CDI Score Estimation

The method of estimating a child's full CDI score closely resembled that presented in Mayor and Mani (2019). Specifically, for each test item i responded to (either known or not known by the child), a histogram of full CDI scores of language-, sex-, and age-matched children having the same response on item i is extracted from Wordbank. A normal distribution is then fitted to each of these item-based histograms using maximum likelihood estimation. To smoothen out random fluctuations, a polynomial curve is subsequently fitted to the parameters

(i.e., mean and standard deviation) extracted from the fitted histograms respectively. Unlike Mayor and Mani who fitted cubic polynomials, a more flexible approach to polynomial curve fitting is taken here, that is, by adapting the degree of polynomials to the breadth of the distribution of the vocabulary counts.[25] Once normalised, each histogram can be thought of as the distribution of full CDI score probabilities given the response for each test item. All histograms are subsequently log-summed and from the resulting histogram, the mode retrieved. Finally, a linear transformation[26] of this mode produces the estimate of the child's full CDI score. This linear transformation is needed to ensure that the full range of CDI scores associated with language-, sex-, and age-matched children can be reached.

5.2.2.3 Real-Data Simulations

To validate the present approach, real-data simulations were conducted using four CDI–WS data sets (retrieved from Wordbank; Frank et al., 2017) of varying sizes and with relatively homogenous sample sizes across all ages: American English (Fenson et al., 2007), Danish (Bleses et al., 2008a), Beijing Mandarin (Tardif et al., 2009), and Italian (M. C. Caselli & Casadio, 1995). The American English data set was categorised as *very large-sized* for having more than 200 samples for each age, in months; the Danish data set was categorised as *large-sized* for having between 100 and 200 samples for each age; the Beijing Mandarin data set was categorised as *medium-sized* for having between 50 and 100 samples for each age; the Italian data set was categorised as *small-sized* for having fewer than 50 samples for each age.

The performance of the present approach, hereafter referred to as the *IRT version*, in estimating full CDI scores was compared to the *original version*

[25]The breadth of the distribution is quantified by computing the median absolute deviation (MAD) of vocabulary counts for each age, in months. When MAD < 100, a linear polynomial is fitted to improve generalisation, whereas when MAD > 100, a cubic polynomial is fitted to obtain a better fit.

[26]$x = N(m - min)/(max - min)$, where x is the estimated CDI score, N is the number of items on the full CDI, m is the mode, and min and max, the minimum and maximum estimated CDI scores of language-, sex-, and age-matched children respectively.

presented in Mayor and Mani (2019), as well as a baseline measure, in which estimates were computed by summing items reported as known on a *random* selection of items from the full CDI and scaling these up to the instrument size to fit the range of the full CDI scores. Estimates were derived from tests consisting of 5, 10, 25, 50, 100, 200, 400, and all items on the CDI.

In addition, comparisons were made between the IRT version and established short-form versions of CDIs (i.e., Bleses et al., 2010; Fenson, Pethick, et al., 2000; Rinaldi et al., 2019; Tardif et al., 2008), with short form estimates computed in a similar manner to the baseline measure, that is, by summing items reported as known in the short forms and scaling these up to the full CDI size.

In line with previous work using real-data simulations (i.e., Makransky et al., 2016; Mayor & Mani, 2019), three outcomes are reported here for each CDI, across both sexes and different age groups: (a) the correlation between the estimates and the full CDI score; (b) the average *SE*; and (c) reliability ($1 - SE^2$). The outcomes obtained from the original version and the baseline measure were averaged over 10 simulations, whereas those for the IRT version were based on single simulations as items are selected on the basis of each child's ability level in CATs, consequently constraining the selection of items for each child. As in Makransky et al., the following minimal thresholds for test acceptability are adopted: (a) a correlation above .95 with the full CDI, (b) an average *SE* below .20, and (c) reliability above .96.

5.2.3 Results

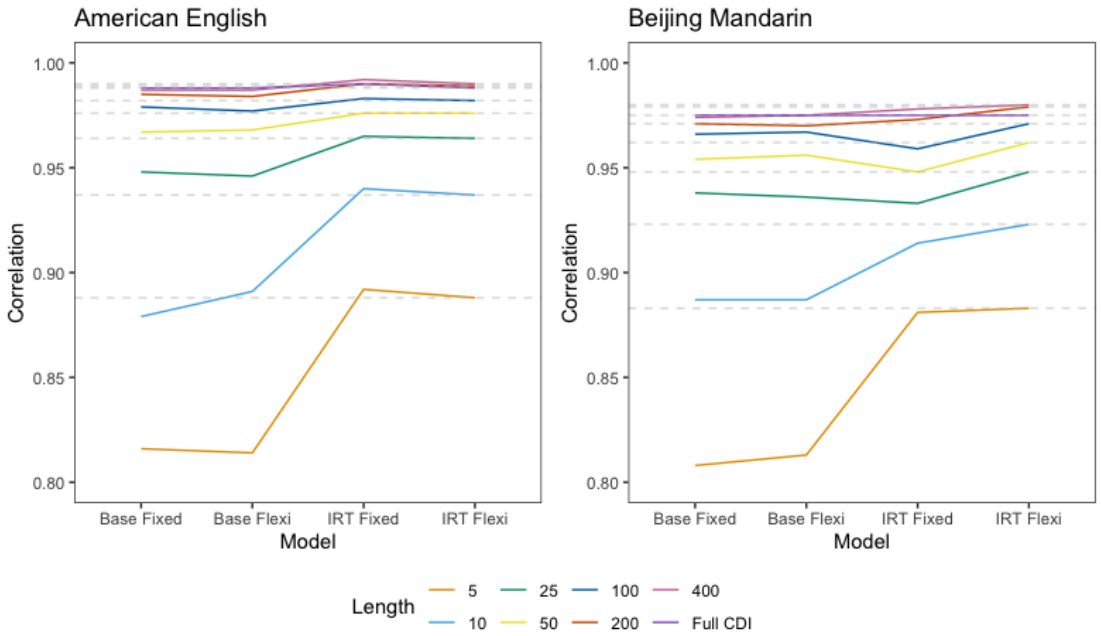
5.2.3.1 Model Selection

The IRT version differs from the original version (Mayor & Mani, 2019) in two respects: the application of IRT-based CAT and flexible polynomial fitting. Prior to selecting the final model, preliminary comparisons were made (i.e., in terms of correlations), for each step of change applied to the original version, using the very large-sized American English CDI–WS data set and the medium-sized Beijing Mandarin CDI–WS data set. That is, comparisons were

made across four versions of the model: the original version, the original version with flexible polynomial fitting (but without IRT-based CAT), the original version with IRT-based CAT (but without flexible polynomial fitting), and the original version with both IRT-based CAT and flexible polynomial fitting (i.e., the IRT version). As shown in Figure 5.7, when applied to the very large-sized data set, both the model with IRT-based CAT and the maximal model (with the combination of IRT-based CAT and flexible polynomial fitting) performed comparably well, with slightly better performance by the IRT-only model. When applied to the medium-sized data set, the application of the maximal model led to the largest improvements. Thus, the maximal model was selected as the final model.

Figure 5.7

Model Comparisons Across Different Test Lengths on the American English and Beijing Mandarin CDI–WS



Note. Base Fixed refers to the original model; Base Flexi refers to the original model with flexible polynomial fitting; IRT Fixed refers to the original model with IRT-based CAT; IRT Flexi refers to the original model with both flexible polynomial fitting and IRT-based CAT. Dashed lines represent the values of IRT Flexi at each test length.

5.2.3.2 Comparisons With the Original Version

5.2.3.2.1 American English CDI–WS

Real-data simulations were run using the very large-sized American English CDI–WS data set (Fenson et al., 2007), for each age (16–30 months) and sex, with tests consisting of 5, 10, 25, 50, 100, 200, 400, and all 680 items on the CDI. An overview of the results, along with the results reported in Makransky et al. (2016), obtained from tests with 100 items and below is provided in Figure 5.8, while the full list of values across both sexes and all test lengths can be found in Table H.1 in the appendix.

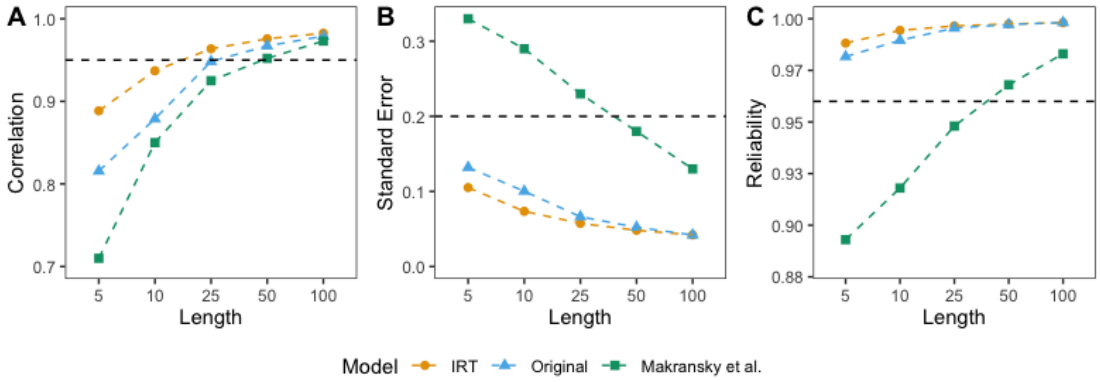
In terms of correlations, the IRT version outperformed the original version across both sexes and all test lengths, achieving correlations above .90 with just 10 items. Correlations greater than the .95 threshold for test acceptability, as suggested by Makransky et al. (2016), were achieved at 25 items. In terms of average *SE*s and reliability, performance between the IRT version and the original version was similar at 25 items and above, and at shorter tests (i.e., below 25 items), the former outperformed the latter. Furthermore, the IRT version had better correlations, average *SE*s, and reliability than the baseline measure at 50 items and below. Additional real-data simulations revealed that a correlation of .95 was already achieved at 14 items, with an average *SE* of .07 and a reliability of .995.

To further evaluate the performance of the IRT version, comparisons between the IRT version and the original version were made across five different age groups (i.e., 16–18 months, 19–21 months, 22–24 months, 25–27 months, and 28–30 months; see Table H.2 in the appendix). Once again, the IRT version outperformed the original version in terms of correlations across all age groups. Notably, at 25 items, correlations were already greater than the .95 threshold across all age groups, while in the original version, at least 50 items were required to achieve correlations of .95 and above in both the youngest (16–18 months) and the oldest (28–30 months) age groups. In line with Makransky

et al. (2016) and Mayor and Mani (2019), a marked reduction in performance was observed when the test featured fewer than 10 items.

Figure 5.8

Comparisons Between the IRT Version and the Original Version Across Different Test Lengths on the American English CDI–WS, With Makransky et al.'s (2016) Values for Reference



Note. Dashed horizontal lines at .95 in Figure A, .20 in Figure B, and .96 in Figure C represent Makransky et al.'s (2016) recommended thresholds for test acceptability. The x-axes are not linear.

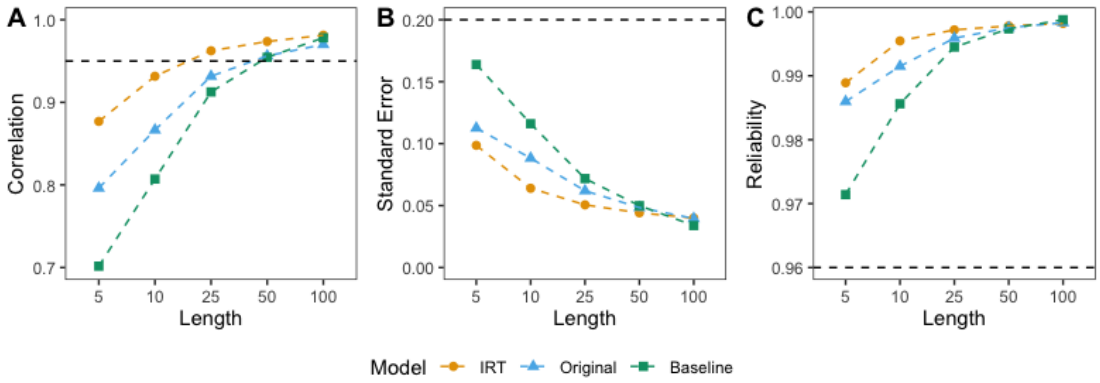
5.2.3.2.2 Danish CDI–WS

Real-data simulations were run using the large-sized Danish CDI–WS data set (Bleses et al., 2008a), for each age (16–30 months) and sex, with tests consisting of 5, 10, 25, 50, 100, 200, 400, and all 725 items on the CDI. An overview of the results obtained from tests with 100 items and below is provided in Figure 5.9, while the full list of values across both sexes and all test lengths can be found in Table H.3 in the appendix.

Similar to the American English CDI–WS data set, the IRT version outperformed the original version in terms of correlations, across both sexes and all test lengths, achieving correlations above .90 with just 10 items and correlations above the .95 threshold with 25 items. In contrast, at least 50 items were required in the original version to achieve correlations of .95 and above across both sexes. In terms of average *SE*s and reliability, consistent

Figure 5.9

Comparisons Between the IRT Version and the Original Version Across Different Test Lengths on the Danish CDI–WS, With Random Lists as Baseline



Note. Dashed horizontal lines at .95 in Figure A, .20 in Figure B, and .96 in Figure C represent Makransky et al.'s (2016) recommended thresholds for test acceptability. The x-axes are not linear.

improvements relative to the original version were also observed for the IRT version. Furthermore, the IRT version had better correlations, average *SE*s, and reliability than the baseline measure at 50 items and below. Additional real-data simulations revealed that a correlation of .95 was already achieved at 17 items, with an average *SE* of .06 and a reliability of .997.

5.2.3.2.3 Beijing Mandarin CDI–WS

Real-data simulations were run using the medium-sized Beijing Mandarin CDI–WS data set (Tardif et al., 2009), for each age (16–30 months) and sex, with tests consisting of 5, 10, 25, 50, 100, 200, 400, and all 799 items on the CDI. An overview of the results obtained from tests with 100 items and below is provided in Figure 5.10, while the full list of values across both sexes and all test lengths can be found in Table H.4 in the appendix.

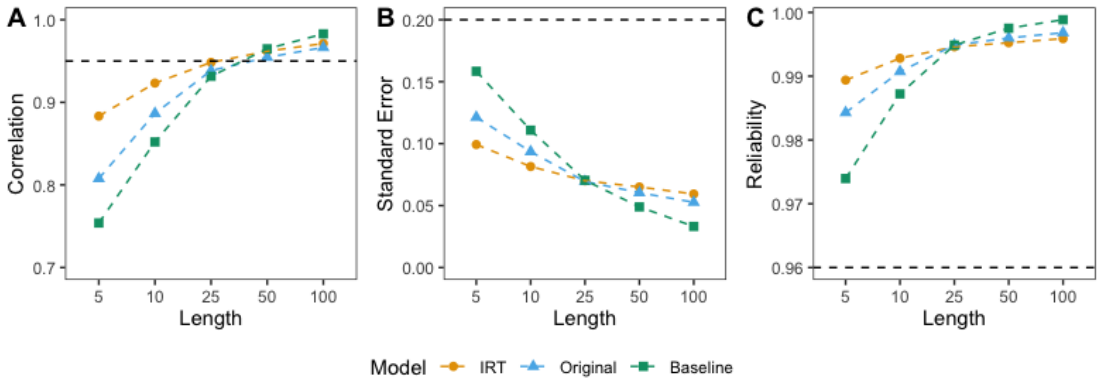
As with the American English and Danish CDI–WS data sets, the IRT version generally outperformed the original version across both sexes and all test lengths, with similar or better correlations, average *SE*s, and reliability. With a reduced sample size, correlations of above the .95 threshold were achieved at 50 items for females and at 25 items for males. In comparison to the baseline measure, the IRT version had higher correlations at 25 items and below, with similar or better average *SE*s and reliability. Additional real-data simulations revealed that a correlation of .95 was achieved at 36 items for females, with an average *SE* of .08 and a reliability of .993, and at 23 items for males, with an average *SE* of .09 and a reliability of .992.

5.2.3.2.4 Italian CDI–WS

Real-data simulations were run using the small-sized Italian CDI–WS data set (M. C. Caselli & Casadio, 1995), for each age (18–30 months) and sex, with tests consisting of 5, 10, 25, 50, 100, 200, 400, and all 670 items on the CDI. For this particular data set, the original version (with cubic polynomial fitting) was unable to reliably estimate full CDI scores. Thus, the results reported here were obtained using the original model with flexible polynomial fitting. An

Figure 5.10

Comparisons Between the IRT Version and the Original Version Across Different Test Lengths on the Beijing Mandarin CDI–WS, With Random Lists as Baseline



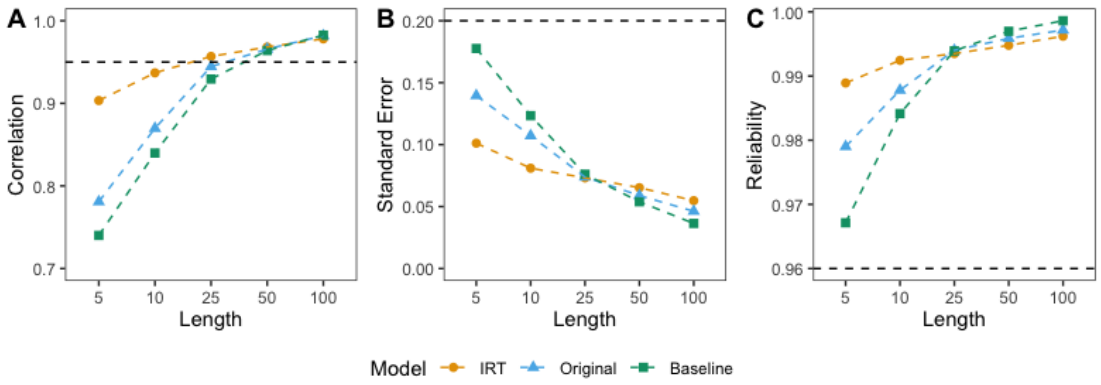
Note. Dashed horizontal lines at .95 in Figure A, .20 in Figure B, and .96 in Figure C represent Makransky et al.'s (2016) recommended thresholds for test acceptability. The x-axes are not linear.

overview of the results obtained from tests with 100 items and below is provided in Figure 5.11, while the full list of values across both sexes and all test lengths can be found in Table H.5 in the appendix.

In terms of correlations, the IRT version generally outperformed the original version (except at 50-, 100-, and 200-item tests among females, where the original version had slightly higher correlations). Nevertheless, despite the small sample size, correlations above the .95 threshold were again achieved with just 25 items, across both sexes. While the original version achieved the same for females, 50 items were required for males to achieve correlations above .95. In terms of average *SE*s and reliability, the IRT version had similar or better performance than the original version at 25 items and below. In comparison to the baseline measure, correlations of the IRT version were higher at 50 items and below, with comparable, if not better, average *SE*s and reliability. Additional real-data simulations revealed that a correlation of .95 was already achieved at 15 items, with an average *SE* of .08 and a reliability of .993.

Figure 5.11

Comparisons Between the IRT Version and the Original Version Across Different Test Lengths on the Italian CDI–WS, With Random Lists as Baseline



Note. The original version here refers to the original version with flexible polynomial fitting. Dashed horizontal lines at .95 in Figure A, .20 in Figure B, and .96 in Figure C represent Makransky et al.'s (2016) recommended thresholds for test acceptability. The x-axes are not linear.

5.2.3.3 Comparisons With Established Short-Form Versions of CDIs

5.2.3.3.1 American English CDI–WS

Comparisons were made between the IRT version and the short-form version of the American English CDI–WS (Form A; Fenson, Pethick, et al., 2000), with random lists as the baseline measure, across five different age groups (i.e., 16–18 months, 19–21 months, 22–24 months, 25–27 months, and 28–30 months). In accordance with the number of test items in the short form, 100-item tests were used in the real-data simulations.

As indicated in Table 5.8, the IRT version performed better than the short-form version in terms of correlations in the younger and middle age groups (between 16 and 24 months), whereas the short-form version performed better in the older age groups (between 25 and 30 months), with similar average *SE*s and reliability overall. The baseline measure outperformed the IRT version between 22 and 30 months as well as the short-form version across all age groups, with better correlations, average *SE*s, and reliability in general.

5.2.3.3.2 Danish CDI–WS

Comparisons were made between the IRT version and the short-form version of the Danish CDI–WS (Bleses et al., 2010), with random lists as the baseline measure, across five different age groups (i.e., 16–18 months, 19–21 months, 22–24 months, 25–27 months, and 28–30 months). In accordance with the number of test items in the short form, 100-item tests were used in the real-data simulations.

As indicated in Table 5.9, the IRT version performed better than both the short-form version and the baseline measure in terms of correlations in the younger and middle age groups (between 16 and 24 months), while the short-form version had the best performance after 24 months. In comparison to both the short-form version and the baseline measure, the IRT version had similar, if not slightly poorer, average *SE*s and reliability overall.

5.2.3.3.3 Beijing Mandarin CDI–WS

Comparisons were made between the IRT version and the short-form version of the Beijing Mandarin CDI–WS (Tardif et al., 2008), with random lists as the baseline measure, across five different age groups (i.e., 16–18 months, 19–21 months, 22–24 months, 25–27 months, and 28–30 months). In accordance with the number of test items in the short form, 110-item tests were used in the real-data simulations.

As indicated in Table 5.10, the IRT version had poorer correlations in comparison to both the short-form version and the baseline measure, except in the youngest age group (16–18 months). Average *SE*s and reliability were also poorer than the other two approaches overall.

5.2.3.3.4 Italian CDI–WS

The final comparisons were made between the IRT version and the short-form version of the Italian CDI–WS (Rinaldi et al., 2019), with random lists as the baseline measure, across four different age groups (i.e., 18–21 months, 22–24 months, 25–27 months, and 28–30 months). In accordance with the number of test items in the short form, 100-item tests were used in the real-data simulations.

As indicated in Table 5.11, the IRT version performed better than both the short-form version and the baseline measure in terms of correlations in the younger age groups (between 18 and 24 months), while the short-form version had the best performance after 24 months. Average *SE*s and reliability were comparable across all three approaches.

5.2.4 Discussion

In view of the limitations of extant short-form versions of CDIs, the present study aimed to develop a language-general approach that produces short forms in which items are selected to be maximally informative and derives CDI estimates that are on the same scale as the full CDI scores. To realise this aim,

Table 5.8

Comparisons Between the IRT Version and Fenson, Pethick, et al.'s (2000) Short-Form Version of the American CDI–WS Across Different Age Groups, With Random 100-Item Lists as Baseline

Age group	IRT version			Short-form version			Baseline		
(months)	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.
16–18	.982	.03	.999	.954	.04	.998	.975	.03	.999
19–21	.990	.03	.999	.973	.05	.997	.985	.04	.999
22–24	.985	.04	.998	.984	.05	.997	.988	.04	.998
25–27	.978	.05	.997	.986	.06	.997	.988	.04	.999
28–30	.978	.05	.997	.985	.04	.998	.987	.04	.999

Note. Avg. *SE* = average standard error; Rel. = reliability.

Table 5.9

Comparisons Between the IRT Version and Bleses et al.'s (2010) Short-Form Version of the Danish CDI–WS Across Different Age Groups, With Random 100-Item Lists as Baseline

Age group	IRT version			Short-form version			Baseline		
(months)	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.
16–18	.986	.02	.999	.968	.02	1.000	.972	.02	1.000
19–21	.978	.05	.997	.969	.03	.999	.973	.03	.999
22–24	.990	.04	.999	.983	.04	.998	.982	.04	.998
25–27	.981	.05	.997	.984	.05	.997	.983	.04	.998
28–30	.971	.06	.997	.985	.05	.997	.980	.04	.998

Note. Avg. *SE* = average standard error; Rel. = reliability.

Table 5.10

Comparisons Between the IRT Version and Tardif et al.'s (2008) Short-Form Version of the Beijing Mandarin CDI–WS Across Different Age Groups, With Random 110-Item Lists as Baseline

Age group	IRT version			Short-form version			Baseline		
(months)	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.
16–18	.986	.06	.995	.980	.04	.999	.979	.04	.999
19–21	.984	.05	.998	.990	.05	.998	.990	.05	.998
22–24	.963	.07	.995	.981	.04	.998	.986	.04	.998
25–27	.961	.06	.997	.979	.04	.998	.983	.04	.999
28–30	.979	.06	.996	.981	.03	.999	.976	.03	.999

Note. Avg. *SE* = average standard error; Rel. = reliability.

Table 5.11

Comparisons Between the IRT Version and Rinaldi et al.'s (2019) Short-Form Version of the Italian CDI–WS Across Different Age Groups, With Random 100-Item Lists as Baseline

Age group	IRT version			Short-form version			Baseline		
(months)	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.
18–21	.981	.04	.998	.972	.03	.999	.975	.03	.999
22–24	.990	.03	.999	.983	.04	.998	.982	.04	.998
25–27	.981	.04	.998	.984	.05	.997	.983	.05	.998
28–30	.971	.05	.998	.985	.05	.997	.980	.05	.998

Note. Avg. *SE* = average standard error; Rel. = reliability.

the present approach, that is, the IRT version, builds upon Mayor and Mani's (2019) approach to estimating full CDI scores with the application of IRT-based CAT that adapts to the child's ability by dynamically selecting test items to be maximally informative (as in Makransky et al., 2016). The performance of the IRT version was evaluated by conducting real-data simulations for each age (in months) and sex, using four CDI–WS versions having varying sample sizes on Wordbank (Frank et al., 2017): American English (a very large data set; Fenson et al., 2007), Danish (a large data set; Bleses et al., 2008a), Beijing Mandarin (a medium-sized data set; Tardif et al., 2009), and Italian (a small data set; M. C. Caselli & Casadio, 1995). In addition, the performance of the IRT version was compared to three other approaches: Mayor and Mani's model (in a novel implementation, in R; R Core Team, 2018), established short forms (i.e., Bleses et al., 2010; Fenson, Pethick, et al., 2000; Rinaldi et al., 2019; Tardif et al., 2008), as well as a baseline measure (i.e., the sum of vocabulary counts on a random sample of items from the full CDI).

Overall, the IRT version met the minimal thresholds for test acceptability (correlations above .95 with the full CDI, average *SE*s below .20, and reliability above .96, as suggested in Makransky et al., 2016) with tests consisting of fewer than 17 items. The only exception to this was the Beijing Mandarin data set, for which the thresholds were only met with 36-item tests for females and 23-item tests for males. Further inspection on the data set revealed that the female data had a much lower variation (quantified by MAD) relative to the male data. Specifically, from 23 months of age onwards, the female data was more left-skewed than the male data, that is, most females in the sample had high CDI scores. In contrast, males had scores that continued to vary until about 27 months, when a majority of them, like females, began to reach the ceiling. The implication of this is twofold: first, a larger and more representative sample may be needed for females; second, items in the CDI may be too easy, especially for females above 23 months of age, thus reaching the ceiling earlier than males.

Nevertheless, results from the real-data simulations suggest that the IRT version can reliably estimate children's full CDI scores with tests consisting of as

few as 25 items for the most part, regardless of language and sex. Analyses conducted across different age groups (ranging from 16 to 30 months) using the American English data set extend this finding, suggesting that a 25-item test can be suitably used with children across all age groups.

Across all four CDIs, both sexes, and different test lengths, the IRT version compared favourably with Mayor and Mani's (2019) model, in terms of correlations, average *SE*s, and reliability. In other words, the estimates elicited via the IRT version have a closer match to children's full CDI scores. In comparison to the baseline measure, the IRT version had better correlations, average *SE*s, and reliability for all short tests (i.e., tests having 50 items and below). Remarkably, starting at 50 items, the baseline measure achieved correlations above .95, with good average *SE*s and reliability across all four CDIs. At 100 items, the baseline measure also performed similarly to established short forms. Such impressive results should be attributed to the high internal consistency of CDIs (Bleses et al., 2008a; Fenson et al., 2007; Tardif et al., 2009).

The final comparisons were made between the IRT version and established short forms, also with random lists as the baseline measure. Here, tests consisting of 100 items (110 items for Beijing Mandarin) were used, in accordance with the number of test items in established short forms. Overall, all three approaches met the minimal threshold for test acceptability across all CDIs and age groups, with the IRT version typically outperforming established short forms in the younger and middle age groups (i.e., between 16 and 24 months), except for the Beijing Mandarin data set. In the older age groups (i.e., between 25 and 30 months), both established short forms and the baseline measure had better performance than the IRT version. It is noteworthy, though, that the development of short forms for even just a single language can be laborious. Crucially, the objective is to provide a briefer format that effectively reduces administration time—a 100-item test may still pose an obstacle to parents with low literacy skills and even more so in situations requiring multiple tests to be administered (e.g., in a busy clinical setting or in a multilingual environment). The IRT version, on the other hand, is able to provide reliable estimates of full

CDI scores with just a small fraction of test items (14 to 25 items), while offering the advantage of being generalisable, inasmuch as it can be applied to CDIs of any language, as long as sufficient CDI data is available.

The results reported here are based on real-data simulations and thus call for a full assessment of the psychometric properties of the IRT version with new participants to, for instance, establish its test–retest reliability as well as its concurrent validity and predictive validity. With empirical data, lower correlations can be expected as a result of parents responding differently to the same item in the full and short forms, as demonstrated in Mayor and Mani (2019). Furthermore, since items are presented in a semantically unstructured order in the IRT version, as opposed to the more structured full CDI forms that group items according to their semantic classes, it is possible that parents' response behaviour may likewise be affected. Therefore, the essential next steps are to investigate the psychometric properties of the present approach with new participants as well as to examine the differences in parents' response behaviour.

Finally, the IRT version relies on the availability of CDI data from children with matching key demographics (e.g., language, age, and sex) to attain good performance. Current findings suggest that the IRT version is able to reliably estimate full CDI scores with as few as 15 items—even when having a small data set (with fewer than 50 samples in each age, in months)—effectively cutting administration time to a mere couple of minutes. Thus, the public sharing of data collected is instrumental in enabling access to and reuse of these data, which in turn allow for the establishment of computerised adaptive tests that are tailored to each child.

5.3 Summary

This chapter described two studies relevant to early word knowledge assessment. The first explored the viability of tablets in assessing young children's word knowledge by means of a word recognition task that is similar to the CCT (Friend & Keplinger, 2003). Overall, preliminary data suggests that a tablet-based word recognition task can be a useful performance-based measure of

receptive vocabulary skills in the second year of life—and potentially serve as a supplemental and convergent measure of parent reports, though there remains specifics in the design of the assessment (e.g., the selection of test items) that need to be further explored and improved. In the second study, an approach to producing short-form versions of CDIs was presented. The approach administers CDIs as IRT-based CAT (as in Makransky et al., 2016) and derives estimates of full CDI scores based on Mayor and Mani's (2019) work. Real-data simulations conducted using adaptations of the CDI–WS in four different languages revealed that correlations exceeding .95 with full CDI administrations were reached with as few as 15 test items, with high levels of reliability, even when CDIs (e.g., Italian) have smaller samples in online repositories, for instance, with around 50 samples for each age, in months. The next chapter discusses the key findings of this thesis in relation to the four research questions laid out in Chapter 2.

CHAPTER 6. GENERAL DISCUSSION

This chapter discusses the main findings from the present studies in relation to the four research questions laid out in Chapter 2:

1. Can young children learn words using tablets?
2. What are the factors that may affect young children's learning from tablets?
3. How can young children's word knowledge be assessed using tablets?
4. How can short-form versions of the MacArthur–Bates Communicative Development Inventories (CDI) be further developed to more efficiently estimate early word knowledge?

The implications of the findings and research limitations are also discussed along with possible avenues for future research, with a view to informing three communities: (a) those who are concerned with the educational potential of tablet apps during early childhood, including parents, early language researchers, educators, and app developers; (b) researchers interested in expanding their toolkit for collecting developmental data to include web technology– and tablet-based methods; as well as (c) researchers and practitioners seeking alternatives for quick and cost-effective assessments of early vocabulary.

6.1 Overview of Main Findings

6.1.1 Questions 1 and 2: Early Word Learning Using Tablets

Addressing the first and second research questions, Chapter 4 presented a series of three studies which examined the educational potential of tablet apps in the word learning domain for children aged 2 to 3 years. Consistent with previous work (Choi & Kirkorian, 2016; Kirkorian, Choi, et al., 2016), the results

from Study 1A suggest a passive advantage in terms of recognition accuracy (of novel word–referent associations) among 30- and 40-month-olds but no such advantage was found among 24-month-olds. Put differently, giving children active control over their learning experiences did not appear to benefit children across the three age groups but passive watching led to better performance among older children. One possible explanation for active children's poorer performance is that interacting with the app by tapping takes up valuable cognitive resources, which could have otherwise been allocated to support information encoding and retention (i.e., a competence deficit). Alternatively, it may be that active children continue to indicate their preferences during the test phase, treating this as the learning phase, despite having learnt the novel word–referent associations (i.e., a performance deficit). Using a more implicit measure of children's eye movements, Study 1B attempted to clarify the competence–performance distinction. While Study 1B replicated the findings in Study 1A with a new group of 30-month-olds from a different cultural and linguistic background, no differences were found across both active and passive conditions in terms of their gaze behaviour during the test phase, that is, both groups of children fixated the target equally. This was despite passive children fixating the target more than their active peers during the learning phase. In other words, the findings suggest that children learnt equally across both conditions, but there may be performance, rather than competence, costs associated with active selection in tasks designed as in these studies.

6.1.2 Question 3: Early Word Knowledge Assessment Using Tablets

The third research question was addressed in Chapter 5. Preliminary data obtained from Study 2 indicates that children (aged between 18 and 20 months) were responding above chance in the two-alternative forced choice (2AFC) word recognition task which assessed word comprehension and that their performance was consistent with a priori trial difficulty, broadly mirroring findings from previous work using the Computerized Comprehension Task (CCT; Friend & Keplinger, 2003, 2008). Children also showed more robust recognition

in semantically unrelated (i.e., trials in which the target and distractor are from different semantic categories) than related trials (i.e., trials in which the target and distractor are from the same semantic category) possibly due to competition effects of semantic relatedness which interfered with children's lexical decision about the target. Indeed, Arias-Trejo and Plunkett (2010) found that children aged between 18 and 24 months performed worse in responding to named target images in the presence of semantically similar competitors. Crucially, in line with the CCT (Friend & Keplinger, 2008), children attempted more easy than difficult trials, suggesting that non-responses reflect word knowledge—that children are unable to distinguish the target from the distractor—rather than their non-compliance or lack of motivation. The word recognition task also evinced acceptable convergent validity with the CDI–Words and Gestures (CDI–WG) as well as good item-level agreement between parent reports and children's responses. Examining parent–child agreement in relation to semantic relatedness and difficulty, it was found that agreement was significantly higher in semantically unrelated than related trials when these were categorised as easy trials. This discrepancy suggests that parents may not always discriminate between words that are truly understood (i.e., strong, decontextualised word–referent associations) and words that are recognised in the presence of familiar or supportive cues (i.e., weak word–referent associations; Friend et al., 2018; Houston-Price et al., 2007; Tomasello & Mervis, 1994). Nevertheless, parents are still adequate informants of their child's language abilities as parent-reported comprehension was found to be a significant predictor of children's recognition accuracy.

While the focus of the study was not to examine potential environmental influences (i.e., whether the study was conducted in-lab or remotely), the finding that lab and online samples did not differ significantly in terms of their motivation (as indexed by the number of trials attempted) and recognition accuracy offered a glimpse into the possibility of having parents administer such assessments to their own child at home.

6.1.3 Question 4: Short-Form Versions of CDIs

Turning now to the fourth and final question, Study 3 presented a language-general approach to producing short-form versions of CDIs with test items that are maximally informative by combining item response theory (IRT)–based computerised adaptive testing (CAT; as in Makransky et al., 2016) which adapts to the ability of each child with Mayor and Mani's (2019) approach which estimates full CDI scores based on prior CDI data from language-, sex-, and age-matched children. Results from real-data simulations demonstrated that the approach compared favourably with Mayor and Mani's approach, producing estimates that match more closely full CDI scores. While the approach did not always outperform established short forms (i.e., Bleses et al., 2010; Fenson, Pethick, et al., 2000; Rinaldi et al., 2019; Tardif et al., 2008) at 100- or 110-item tests, the development of 100-item short forms is labour-intensive. Additionally, as such forms take a one-size-fits-all approach, they may fail to account for individual differences in children and in the parents completing the forms, whereas CATs allow tests that are tailored to each child. Importantly, the objective is to reduce test lengths; 100-item tests may still be daunting for parents having low literacy skills and too time-consuming in cases requiring multiple tests to be completed or when a rapid assessment is desirable (e.g., in a multilingual environment or in a busy clinical or research setting). On the other hand, the approach presented here was able to efficiently estimate full CDI scores with tests having just a small fraction of items (14 to 25 items) on the full CDI—regardless of language, sex, and age—achieving correlations above .95 with full CDI administrations, with high levels of reliability, even when prior CDI data is limited to a small sample (e.g., around 50 samples per month-age).

6.2 Research Implications

This research was motivated by the need to examine ways in which the unique affordances of tablets and apps can contribute to young children's early language development in light of their proliferation in young children's lives.

From the word learning viewpoint, Chapter 4 revealed that children as young as 24 months are capable of learning novel word–referent associations through a tablet app, regardless of whether pseudo-social contingent interactions (i.e., tapping on objects to hear their names, rather than merely observing) or active choice is involved. While such finding may assuage parents' and educators' concerns about the educational potential of tablet apps for young children (at least in the word learning domain), the finding of a performance, rather than competence, deficit among 30- and 40-month-olds who had active control on their course of learning relative to those who did not, suggests that there may not always be systematic benefits associated with active or self-directed learning in "educational" apps and more specifically, that such apps may not be adequately tapping into children's learning progress. Depending on the structure of the learning experience, pseudo-social contingency and self-direction may differentially impact children's performance in certain tasks, without having much impact on their learning competence. For instance, Kirkorian, Choi, et al. (2016) found that pseudo-social contingency (i.e., letting children tap on the object or tap on anywhere on the screen) had the same negative impact on 27.5- to 32-month-olds' performance, even when they were only taught a single word. On the other hand, 3- to 5-year-olds benefited from specific- but not general-contingency when they were taught a single word, that is, they performed better when they learnt the word–referent association by tapping on the object than on a button (Partridge et al., 2015). Furthermore, the same study revealed that self-direction (i.e., letting children decide on the order in which objects were labelled) improved children's performance, although this was only limited to tasks involving fewer objects. It is worth noting that these studies, as well as the present studies used custom-made apps that are not commercially available. Apps that are available, either for free or for a fee, in online app stores (e.g., Google Play Store, Apple App Store) typically come with many interactive features, such as sound effects and animations, that research suggests, may distract young children from the desired learning goals when these features deplete cognitive resources (Parish-Morris et al., 2013; Takacs et al., 2015).

However, when used appropriately, such enhancements have been found to promote engagement during learning (Smeets & Bus, 2015). Offering yet another perspective, a very recent study found that the inclusion of simple interactive features (both relevant and irrelevant) were neither helpful nor harmful for word learning and story comprehension as children performed similarly in both kinds of tasks (Etta and Kirkorian, 2019; see also Bus et al., 2015 for a review).

Confronted with different perspectives of educational apps and the vast selections in the "chaotic Wild West of digital apps" (Guernsey et al., 2012, p. 15), parents and educators who are seeking critical information about how digital, especially interactive media (e.g., apps) can be leveraged to support young children's learning and development can turn to resources such as Common Sense Media[27] and Children's Technology Review[28] that provide advice on best practices and evaluations of digital media in helping parents and educators make informed decisions about digital media selection and use. Beyond that, it is also imperative that app developers/publishers collaborate with educators and researchers, or at least, take into consideration research-based information—for instance, by recruiting the four "pillars" of learning (i.e., active "minds-on" participation, social interaction, sustained engagement, and meaningful connections; Hirsh-Pasek et al., 2015)—in designing developmentally appropriate, high-quality apps that set the stage for effective learning in both formal and informal learning environments.

From the word knowledge assessment viewpoint, Chapter 5 showed that children as young as 18 months can engage meaningfully with a tablet-based CCT-like assessment, with minimal verbal instruction and child–administrator interaction. The encouraging results obtained further suggest that such assessments have scope for deriving a direct measure of early vocabulary comprehension that can supplement parent reports, thereby addressing concerns relating to the exclusive use of parent reports and allowing a more complete picture of children's early language development. While only 24 lexical items were assessed in the study, using a "one-shot" design, it is possible to assess as

[27]https://www.commonsensemedia.org/

[28]https://reviews.childrenstech.com/ctr/home.php

many as 48 items under 10 minutes, in keeping with Semmelmann et al.'s (2016) recommendation that child-directed tablet-based tasks should be below 15 minutes in length. Coupled with the use of semantically related target–distractor pairs, a "one-shot" design can be useful for tapping children's strong, rather than weak word–referent associations (Styles & Plunkett, 2008). In the "shorts–zipper" pair, for instance, a zipper can be found on a pair of shorts and both items are likely to be encountered when a child is getting dressed. Given that the child only has one chance to respond, a stronger word–referent association—beyond knowing that a zipper is related to the "dressing up" routine—is needed for the child to distinguish a zipper from a pair of shorts.

As noted in Chapter 5, the assessment can also benefit from a more structured way of selecting test items to be administered so that those that are less informative of children's abilities can be omitted, thereby increasing the quality of test items, while also reducing the length of the assessment. For instance, if a child is asked to pick out a "truck" from a "train" but fails to do so, this could mean that the child knows neither the word "truck" nor "train" and thus, both words can possibly be omitted in subsequent trials for this child and other words can be assessed instead. The generic approach to producing very short (fewer than 25 test items) CDIs that adapt to each child's ability with a dynamic selection of test items presented in the same chapter may lend itself well in this regard. By combining the approach with the child-directed assessment, the advantages related to its application to parent reports can likewise be reaped—including (a) the automated and adaptive selection of test items for each child, (b) the reduction in administration time, as well as (c) the convenient adaptation of the assessment to any language with sufficient CDI data available on online repositories (e.g., Wordbank; Frank et al., 2017)—and importantly, all of these can be achieved without compromising on the accuracy and precision of the full CDIs.

From the methodology viewpoint, the present research extends previous findings on the viability of tablets in early developmental research (i.e., Frank et al., 2016; Semmelmann et al., 2016), illustrating that tablets can be used to

collect data even among children as young as 18 months when (a) care is taken to familiarise them with the experimental task, (b) the experiment is kept below 15 minutes in length, and (c) the required gestures for responding are developmentally appropriate (e.g., tapping would be a more intuitive gesture than pinching to a 1-year-old; Sesame Workshop, 2012).

Furthermore, Study 1B and Study 2 exemplified the use of e-Babylab (the authoring tool presented in Chapter 3) in creating and running online browser-based experiments. The capability of e-Babylab to create experiments that simultaneously record participants' explicit (e.g., screen touches) and implicit responses (e.g., eye movement) was also demonstrated in Study 1B in which children's eye movements were recorded (using the built-in front-facing camera of a tablet) as they responded in the word learning task by tapping on objects shown on-screen.

By applying web technology to the tablet-based experiments, that is, by programming these experiments as web applications and hosting them online, the present research additionally demonstrated the advantages of web technology–based experimentation. More specifically, putting the experiments online has enabled the collection of data in three different countries (i.e., Germany, Malaysia, and Norway) and allowed experiments to be "brought to the participants"; for instance, in Malaysia, children were tested at their respective childcare centres. Notably, when the COVID-19 pandemic shut everything down in Norway, data collection, which was initially carried out by an experimenter in the laboratory and at the kindergarten, could still proceed with minimal disruption: because the study was hosted online, all that had to be done was to send the URL of the study to parents and ask them to administer the study to their own child at home.

6.3 Research Limitations and Future Directions

Although the present findings have shed light on young children's word learning from tablets as well as the potential use of tablets as a means to assess early word knowledge, there remains several pressing questions that future

research needs to address. First, the studies were conducted among monolingual children. Thus, the findings may not generalise to non-monolingual populations as children exposed to more than one language employ different processes in word learning (e.g., Byers-Heinlein, 2017; Kan & Kohnert, 2008; Yoshida et al., 2011). With regard to word knowledge assessment, the convergent validity of the word recognition task with parent report is likewise limited to monolinguals. Future research should thus examine its use with non-monolingual populations so as to maximise the potential opportunities for advancing our understanding of their language development process early in life. The use of the CCT, for instance, has provided preliminary evidence of a translation facilitation effect in French–English bilinguals' lexical access at 22 months of age, showing that the simultaneous activation of both dominant and non-dominant languages emerges early in development (Poulin-Dubois et al., 2018).

Second, the present analyses have not considered potential differences in socioeconomic status (SES) which may influence children's screen media exposure (Rideout, 2017), their executive functioning (Lawson et al., 2018), and potentially, in turn, their performance in the present studies. For instance, Russo-Johnson et al. (2017) found that children from low SES families learnt better in the tablet-based word learning task by dragging the labelled object than tapping, likely because the former gesture was a more distinctive, meaningful action than the latter for children from low SES families, who spent on average more than double the amount of time using touchscreens than children from middle and high SES families. Thus, further work is required to determine whether and how these factors will affect children's performance in tablet-based tasks, especially when considering the use of tablet-based assessments.

Third, recalling the timing issue pertaining to web experiments discussed in Chapter 2, specifically that reaction time (RT) overestimations vary with different browsers and devices, Study 2 has, for this reason, not considered children's RT, although this measure would have additionally allowed children's speed of word processing to be examined in relation to their vocabulary

development.[29] Such haptic measure of children's processing speed has been shown to be as sensitive as looking time measures and has been successfully used in the study of lexical access in young monolinguals and bilinguals (e.g., DeAnda et al., 2018; Legacy et al., 2016, 2018; Poulin-Dubois et al., 2013; Poulin-Dubois et al., 2018).

Fourth, as discussed in Chapter 5, the results from Study 3 are based on real-data simulations. Thus, this warrants a full assessment of the psychometric properties of the approach presented with new participants in order to establish its test–retest reliability and validity using an array of validity tests, while keeping in mind the potential inconsistencies in parents' response behaviour across the full and short forms (Mayor & Mani, 2019).

Fifth, while Study 2 and Study 3 lay the groundwork for two different measures of early word knowledge (i.e., a performance-based measure and a parent report measure), it is worth noting that the performance-based measure, that is, the tablet-based word recognition assessment, relies on the use of pictorial (and possibly animated or video) representations of lexical items. This means that even if function words (e.g., question words, pronouns, prepositions) have similar discrimination parameters as nouns, adjectives, and verbs (as is the case for the American English CDI–WG; Frank et al., 2021)—and are therefore equally likely as the latter three to be administered in IRT-based CATs, only the latter three can easily be pictured and thus be included in such assessments.

Finally, to broaden the application of e-Babylab, future work could incorporate automatic, webcam-based eye-tracking algorithms (e.g., Papoutsaki et al., 2016; Valliappan et al., 2020) in recording gaze data. It is important to note, though, that the gaze detection performance of such algorithms, at their present state, is susceptible to head movements and body repositioning (Semmelmann & Weigelt, 2018; Valliappan et al., 2020), and may thus be unsuitable for use with young children. Other factors such as illumination conditions, participant's distance from the webcam, system performance,

[29]RT analysis was feasible in Study 1 as the same device and browser were used within each of the studies and RT overestimations generally vary little within any single configuration used.

browser, and webcam quality may likewise affect the performance of such algorithms.

6.4 Conclusion

In summary, the message emerging from the present research is that during early childhood, tablets and apps are a double-edged sword: on the one hand, with appropriate design considerations, the unique affordances of tablets and apps can be harnessed to support learning as well as to provide a valuable performance-based measure of receptive vocabulary skills in the second year of life and as a supplemental and convergent measure of parent reports. On the other hand, depending on the app structure, placing the child in the role of an active, self-guided learner in the context of tablet-based learning may not always be beneficial as this may detract from successful task performance, albeit without having much impact on the child's learning competence.

Another equally important message, relating to the use of tablets for data acquisition—an aspect that early developmental research has, perhaps, often overlooked, is that tablets can be an invaluable tool for collecting developmental data. Because of the highly intuitive touchscreen interface, tablets can be used to collect data from children as young as 18 months. In comparison to preferential looking or eye-tracking paradigms, tablets also offer a more engaging and interactive experience, thus alleviating the difficulty in maintaining young children's interest and attention. When coupled with web technology, tablet-based methods further reduce constraints relating to the geographical location of the research institution, such as regional or even national borders, and allow for the same study to be conducted in different countries, thereby paving the way for cross-cultural collaborations.

That being said, many questions remain about the intricacies of how tablet (or in general, touchscreen) devices can be used to young children's benefit and research has yet to keep pace with their rapid adoption in homes with young children as well as their continuous evolution. In this regard, it is instrumental that researchers, educators, and app developers join forces to establish a strong

evidence base that informs best practices regarding touchscreen use in early childhood as well as the design of high-quality, educational apps.

Until a comprehensive road map is built, parents (or caregivers), who assume the role of mediators of touchscreen devices, are thus encouraged to engage with young children during touchscreen use, for instance, by providing an appropriate amount of guidance or by relating screen content to daily routines, rather than rely on these devices as a standalone educational tool, so as to make the experience educational while entertaining.

REFERENCES

Abdul Aziz, N. A., Mat Sin, N. S., Batmaz, F., Stone, R., & Chung, P. W. H. (2014). Selection of touch gestures for children's applications: Repeated experiment to increase reliability. *International Journal of Advanced Computer Science and Applications*, *5*(4). https://doi.org/10.14569/ijacsa.2014.050415

Adelaar, K. A. (1992). *Proto Malayic: The reconstruction of its phonology and parts of its lexicon and morphology*. Dept. of Linguistics, Research School of Pacific Studies, the Australian National University.

American Academy of Pediatrics. (2016). Media and young minds. *Pediatrics*, *138*(5), e20162591. https://doi.org/10.1542/peds.2016-2591

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505–522. https://doi.org/10.1177/0002764204271506

Anderson, D., & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. *Journal of Deaf Studies and Deaf Education*, *7*(2), 83–106. https://doi.org/10.1093/deafed/7.2.83

Anwyl-Irvine, A. L., Dalmaijer, E., Hodges, N., & Evershed, J. (2020). *Online timing accuracy and precision: A comparison of platforms, browsers, and participant's devices*. PsyArXiv. https://doi.org/10.31234/osf.io/jfeca

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). 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

Apple Inc. (2019). Apple expands Everyone Can Code to bring more coding resources to teachers and students.

https://www.apple.com/newsroom/2019/11/apple-expands-everyone-can-code-to-bring-more-coding-resources-to-teachers-and-students/

Arias-Trejo, N., & Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word-referent identification. *Journal of Experimental Child Psychology*, *105*(1–2), 63–80. https://doi.org/10.1016/j.jecp.2009.10.002

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *64*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Baldwin, D. A. (2000). Interpersonal understanding fuels knowledge acquisition. *Current Directions in Psychological Science*, *9*(2), 40–45.

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

Baldwin, D. A., & Moses, L. J. (2001). Links between social understanding and early word learning: Challenges to current accounts. *Social Development*, *10*(3), 309–329. https://doi.org/10.1111/1467-9507.00168

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. https://doi.org/10.3758/s13428-014-0530-7

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barr, R. (2008). Attention and learning from media during infancy and early childhood. In S. L. Calvert & B. J. Wilson (Eds.), *The handbook of children, media, and development* (pp. 143–165). Blackwell. https://doi.org/10.1002/9781444302752.ch7

Barr, R. (2010). Transfer of learning between 2D and 3D sources during infancy: Informing theory and practice. *Developmental Review*, *30*(2), 128–154. https://doi.org/10.1016/j.dr.2010.03.001

Barr, R., & Hayne, H. (1999). Developmental changes in imitation from television during infancy. *Child Development*, *70*(5), 1067–1081. https://doi.org/10.1111/1467-8624.00079

Barr, R., Moser, A., Rusnak, S., Zimmermann, L., Dickerson, K., Lee, H., & Gerhardstein, P. (2016). The impact of memory load and perceptual cues on puzzle learning by 24-month olds. *Developmental Psychobiology*, *58*(7), 817–828. https://doi.org/10.1002/dev.21450

Barr, R., Muentener, P., & Garcia, A. (2007). Age-related changes in deferred imitation from television by 6- to 18-month-olds. *Developmental Science*, *10*(6), 910–921. https://doi.org/10.1111/j.1467-7687.2007.00641.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *12*(5-6), 507–584. https://doi.org/10.1080/016909697386628

Begus, K., Gliga, T., & Southgate, V. (2014). Infants learn what they want to learn: Responding to infant pointing leads to superior learning. *PLoS ONE*, *9*(10), e108817. https://doi.org/10.1371/journal.pone.0108817s

Bergelson, E., & Aslin, R. N. (2017a). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921. https://doi.org/10.1073/pnas.1712966114

Bergelson, E., & Aslin, R. N. (2017b). Semantic specificity in one-year-olds' word comprehension. *Language Learning and Development*, *13*(4), 481–501. https://doi.org/10.1080/15475441.2017.1324308

Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary

learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology*, *55*(1), 803–832. https://doi.org/10.1146/annurev.psych.55.090902.141601

Blakey, E., Visser, I., & Carroll, D. J. (2016). Different executive functions support different kinds of cognitive flexibility: Evidence from 2-, 3-, and 4-year-olds. *Child Development*, *87*(2), 513–526. https://doi.org/10.1111/cdev.12468

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476. https://doi.org/10.1017/S0142716416000060

Bleses, D., Vach, W., Jørgensen, R. N., & Worm, T. (2010). The internal validity and acceptability of the Danish SI-3: A language-screening instrument for 3-year-olds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 490–507. https://doi.org/10.1044/1092-4388(2009/08-0132)

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008a). The Danish Communicative Developmental Inventories: Validity and main developmental trends. *Journal of Child Language*, *35*(3), 651–669. https://doi.org/10.1017/S0305000907008574

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008b). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, *35*(3), 619–650. https://doi.org/10.1017/S0305000908008714

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Booth, A. E., & Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. *Child*

Development, *80*(1), 15–22.
https://doi.org/10.1111/j.1467-8624.2008.01242.x

Braginsky, M. (2018). *wordbankr: Accessing the Wordbank database* (Version 0.3.0) [R package].
https://CRAN.R-project.org/package=wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind: Discoveries in Cognitive Science*, *3*, 52–67.
https://doi.org/10.1162/opmi_a_00026

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Buchanan, T. (2007). Personality testing on the internet: What we know, and what we do not. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of internet psychology* (pp. 447–460). Oxford University Press.
https://doi.org/10.1093/oxfordhb/9780199561803.013.0028

Bugbee Jr., A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, *23*(1), 87–100. https://doi.org/10.1080/08886504.1990.10781945

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.
https://doi.org/10.1177/1745691610393980

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367–407). Macmillan.
https://doi.org/10.1002/j.2330-8516.1988.tb00291.x

Bus, A. G., Takacs, Z. K., & Kegel, C. A. (2015). Affordances and limitations of electronic storybooks for young children's emergent literacy.

Developmental Review, *35*, 79–97. https://doi.org/10.1016/j.dr.2014.12.004

Byers-Heinlein, K. (2017). Bilingualism affects 9-month-old infants' expectations about how words refer to kinds. *Developmental Science*, *20*(1), e12486. https://doi.org/10.1111/desc.12486

Callaghan, M. N., & Reich, S. M. (2018). Are educational preschool apps designed to teach? An analysis of the app market. *Learning, Media and Technology*, *43*(3), 280–293. https://doi.org/10.1080/17439884.2018.1498355

Canfield, R. L., Smith, E. G., Brezsnyak, M. P., & Snow, K. L. (1997). Information processing through the first year of life: A longitudinal study using the visual expectation paradigm. *Monographs of the Society for Research in Child Development*, *62*(2), i–160. https://doi.org/10.2307/1166196

Carver, L. J., Meltzoff, A. N., & Dawson, G. (2006). Event-related potential (ERP) indices of infants' recognition of familiar and unfamiliar objects in two and three dimensions. *Developmental Science*, *9*(1), 51–62. https://doi.org/10.1111/j.1467-7687.2005.00463.x

Caselli, M. C., & Casadio, P. (1995). *Il primo vocabolario del bambino: Guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita [The child's first words: Guide to the use of the MacArthur questionnaire for the assessment of communication and language in the first years of life]* (Vol. 5). Franco Angeli.

Caselli, M. C., Casadio, P., & Bates, E. (1999). A comparison of the transition from first words to grammar in English and Italian. *Journal of Child Language*, *26*(1), 69–111. https://doi.org/10.1017/S0305000998003687

Caselli, N. K., Lieberman, A. M., & Pyers, J. E. (2020). The ASL-CDI 2.0: An updated, normed adaptation of the MacArthur Bates Communicative Development Inventory for American Sign Language. *Behavior Research Methods*, *52*(5), 2071–2084. https://doi.org/10.3758/s13428-020-01376-6

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009

Castro, R., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2008). Human active learning. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Proceedings of the 21st international conference on neural information processing systems* (pp. 241–248). Curran Associates.

CDI Advisory Board. (2015). Adaptations in other languages. http://mb-cdi.stanford.edu/adaptations.html

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038. https://doi.org/10.3758/s13428-019-01273-7

Choi, K., & Kirkorian, H. L. (2016). Touch or watch to learn? Toddlers' object retrieval using contingent and noncontingent video. *Psychological Science*, *27*(5), 726–736. https://doi.org/10.1177/0956797616636110

Choi, K., Kirkorian, H. L., & Pempek, T. A. (2018). Understanding the transfer deficit: Contextual mismatch, proactive interference, and working memory affect toddlers' video-based transfer. *Child Development*, *89*(4), 1378–1393. https://doi.org/10.1111/cdev.12810

Clark, R. (1974). Performing without competence. *Journal of Child Language*, *1*(1), 1–10. https://doi.org/10.1017/S0305000900000040

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120–131. https://doi.org/10.1017/xps.2014.5

Clynes, A., & Deterding, D. (2011). Standard Malay (Brunei). *Journal of the International Phonetic Association*, *41*(2), 259–268. https://doi.org/10.1017/S002510031100017X

Conboy, B. T., & Thal, D. J. (2006). Ties between the lexicon and grammar: Cross-sectional and longitudinal studies of bilingual toddlers. *Child Development*, *77*(3), 712–735. https://doi.org/10.1111/j.1467-8624.2006.00899.x

Coronavirus: UK lockdown extended for 'at least' three weeks. (2020, April 16). https://www.bbc.com/news/uk-52313715

Couse, L. J., & Chen, D. W. (2010). A tablet computer for young children? Exploring its viability for early childhood education. *Journal of Research on Technology in Education*, *43*(1), 75–96. https://doi.org/10.1080/15391523.2010.10782562

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., Orben, A., Parsons, S., & Schulte-Mecklenbeck, M. (2018). *7 easy steps to Open Science: An annotated reading list*. PsyArXiv. https://doi.org/10.31234/osf.io/cfzyx

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127. https://doi.org/10.3758/BF03203646

Dale, P. S., Price, T. S., Bishop, D. V., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient delay at 3 and 4

years. *Journal of Speech, Language, and Hearing Research*, *46*(3), 544–560. https://doi.org/10.1044/1092-4388(2003/044)

Dale, P. S., Reznick, J. S., & Thal, D. J. (1998). A parent report measure of language development for three-year-olds. *Infant Behavior and Development*, *21*(Suppl.), 370. https://doi.org/10.1016/S0163-6383(98)91583-1

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, *104*, 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

DeAnda, S., Hendrickson, K., Zesiger, P., Poulin-Dubois, D., & Friend, M. (2018). Lexical access in the second year: A study of monolingual and bilingual vocabulary development. *Bilingualism: Language and Cognition*, *21*(2), 314–327. https://doi.org/10.1017/S1366728917000220

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, *238*(4833), 1556–1557. https://doi.org/10.1126/science.2446392

DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, *62*(4), 736. https://doi.org/10.2307/1131174

DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., Strouse, G. A., & O'Doherty, K. (2010). Do babies learn from baby media? *Psychological Science*, *21*(11), 1570–1574. https://doi.org/10.1177/0956797610384145

DeLoache, J. S., Pierroutsakos, S. L., & Uttal, D. H. (2003). The origins of pictorial competence. *Current Directions in Psychological Science*, *12*(4), 114–118. https://doi.org/10.1111/1467-8721.01244

Deocampo, J. A., & Hudson, J. A. (2005). When seeing is not believing: Two-year-olds' use of video representations to find a hidden toy. *Journal of Cognition and Development*, *6*(2), 229–258. https://doi.org/10.1207/s15327647jcd0602_4

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *International Journal of Language & Communication Disorders*, *43*(4), 361–389. https://doi.org/10.1080/13682820701546854

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A crosslinguistic study of the relationship between grammar and lexical development. *Journal of Child Language*, *32*(4), 759–786. https://doi.org/10.1017/s0305000905007105

Dickerson, K., Gerhardstein, P., Zack, E., & Barr, R. (2013). Age-related changes in learning across early childhood: A new imitation task. *Developmental Psychobiology*, *55*(7), 719–732. https://doi.org/10.1002/dev.21068

Diener, M. L., Pierroutsakos, S. L., Troseth, G. L., & Roberts, A. (2008). Video versus reality: Infants' attention and affective responses to video and live presentations. *Media Psychology*, *11*(3), 418–441. https://doi.org/10.1080/15213260802103003

Dink, J., & Ferguson, B. (2018). *eyetrackingR* (Version 0.1.8) [R package]. http://www.eyetracking-R.com

Django documentation: Migrations. (n.d.). https://docs.djangoproject.com/en/3.0/topics/migrations/

Donker, A., & Reitsma, P. (2007). Young children's ability to use a computer mouse. *Computers & Education*, *48*(4), 602–617. https://doi.org/10.1016/j.compedu.2005.05.001

Duff, F. J., Nation, K., Plunkett, K., & Bishop, D. V. (2015). Early prediction of language and literacy problems: Is 18 months too early? *PeerJ*, *3*, e1098. https://doi.org/10.7717/peerj.1098

Duff, F. J., Reen, G., Plunkett, K., & Nation, K. (2015). Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry*, *56*(8), 848–856. https://doi.org/10.1111/jcpp.12378

Dunn, D. M. (2018). *Peabody Picture Vocabulary Test–Fifth Edition*. Pearson Assessment.

Ellis, B. J., & Symons, D. (1990). Sex differences in sexual fantasy: An evolutionary psychological approach. *Journal of Sex Research*, *27*(4), 527–555. https://doi.org/10.1080/00224499009551579

Ellis Weismer, S. (2007). Typical talkers, late talkers, and children with specific language impairment: A language endowment spectrum? In R. Paul (Ed.), *Language disorders from a developmental perspective: Essays in honour of Robin S. Chapman* (pp. 83–101). Erlbaum.

Ellis Weismer, S., & Evans, J. L. (2002). The role of processing limitations in early identification of specific language impairment. *Topics in Language Disorders*, *22*(3), 15–29. https://doi.org/10.1097/00011363-200205000-00004

Embretson, S. E., & Reise, S. P. (2000). *Multivariate Applications Books Series. Item response theory for psychologists.* Erlbaum.

Etta, R. A., & Kirkorian, H. L. (2019). Children's learning from interactive eBooks: Simple irrelevant features are not necessarily worse than relevant ones. *Frontiers in Psychology*, *9*, 2733. https://doi.org/10.3389/fpsyg.2018.02733

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development*, *76*(4), 856–868. https://doi.org/10.1111/j.1467-8624.2005.00882.x

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, *71*(2), 310–322. https://doi.org/10.1111/1467-8624.00146

Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., Pethick, S., & Reilly, J. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. Singular.

Fenson, L., Marchman, V., Thal, D., Dale, P., Reznick, J., & Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Brookes.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000). Measuring variability in early child language: Don't shoot the messenger. *Child Development*, *71*(2), 323–328. https://doi.org/10.1111/1467-8624.00147

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics*, *21*(1), 95–116. https://doi.org/10.1017/S0142716400001053

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology*, *42*(1), 98. https://doi.org/10.1037/0012-1649.42.1.98

Fernald, A., Pinto, J. P., Swingley, D., Weinbergy, 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., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjective–noun phrases. *Cognitive Psychology*, *60*(3), 190–217. https://doi.org/10.1016/j.cogpsych.2009.12.002

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017, July 10–13). *LabVanced: A unified JavaScript framework for online studies* [Paper presentation]. https://www.labvanced.com/publication.html

Flynn, E., & Whiten, A. (2008). Imitation of hierarchical structure versus component details of complex actions by 3- and 5-year-olds. *Journal of Experimental Child Psychology*, *101*(4), 228–240. https://doi.org/10.1016/j.jecp.2008.05.009

Forget-Dubois, N., Dionne, G., Lemelin, J.-P., Pérusse, D., Tremblay, R. E., & Boivin, M. (2009). Early child language mediates the relation between home environment and school readiness. *Child Development*, *80*(3), 736–749. https://doi.org/10.1111/j.1467-8624.2009.01294.x

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2015). Free viewing gaze behavior in infants and adults. *Infancy*, *21*(3), 262–287. https://doi.org/10.1111/infa.12119

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press. https://langcog.github.io/wordbank-book/

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1–17. https://doi.org/10.1080/15248372.2015.1061528

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160–170. https://doi.org/10.1016/j.cognition.2008.11.010

Friend, M., & Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behavior Research Methods, Instruments, & Computers*, *35*(2), 302–309. https://doi.org/10.3758/bf03202556

Friend, M., & Keplinger, M. (2008). Reliability and validity of the Computerized Comprehension Task (CCT): Data from American English and Mexican Spanish infants. *Journal of Child Language*, *35*(1), 77–98. https://doi.org/10.1017/s0305000907008264

Friend, M., Schmitt, S. A., & Simpson, A. M. (2012). Evaluating the predictive validity of the Computerized Comprehension Task: Comprehension predicts production. *Developmental Psychology*, *48*(1), 136–148. https://doi.org/10.1037/a0025511

Friend, M., Smolak, E., Liu, Y., Poulin-Dubois, D., & Zesiger, P. (2018). A cross-language study of decontextualized vocabulary comprehension in toddlerhood and kindergarten readiness. *Developmental Psychology*, *54*(7), 1317. https://doi.org/10.1037/dev0000514

Friend, M., Smolak, E., Patrucco-Nanchen, T., Poulin-Dubois, D., & Zesiger, P. (2019). Language status at age 3: Group and individual prediction from vocabulary comprehension in the second year. *Developmental Psychology*, *55*(1), 9. https://doi.org/10.1037/dev0000617

Friend, M., & Zesiger, P. (2011). Une réplication systématique des propriétés psychométriques du Computerized Comprehension Task dans trois langues [A systematic replication of the psychometric properties of the CCT in three languages]. *Enfance*, *63*(3), 329–344. https://doi.org/10.4074/S0013754511003053

Galeote, M., Checa, E., Sánchez-Palacios, C., Sebastián, E., & Soto, P. (2016). Adaptation of the MacArthur–Bates Communicative Development Inventories for Spanish children with Down Syndrome: Validity and

reliability data for vocabulary. *American Journal of Speech-Language Pathology*, *25*(3), 371–380. https://doi.org/10.1044/2015_AJSLP-15-0007

Garaizar, P., Vadillo, M. A., & López-de-Ipiña, D. (2014). Presentation accuracy of the web revisited: Animation methods in the HTML5 era. *PLoS ONE*, *9*(10), e109812. https://doi.org/10.1371/journal.pone.0109812

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Ghassabian, A., Rescorla, L., Henrichs, J., Jaddoe, V. W., Verhulst, F. C., & Tiemeier, H. (2014). Early lexical development and risk of verbal and nonverbal cognitive delay at school age. *Acta Paediatrica*, *103*(1), 70–80. https://doi.org/10.1111/apa.12449

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

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

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–339. https://doi.org/10.1177/1745691613484936

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Green, B. F. (1988). Construct validity of computer-based tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 77–86). Erlbaum.

Guernsey, L., Levine, M., Chiong, C., & Severns, M. (2012). *Pioneering literacy in the digital Wild West: Empowering parents and educators*. Joan Ganz Cooney Center. https://www.joanganzcooneycenter.org/wp-content/uploads/2012/12/GLR_TechnologyGuide_final.pdf

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481. https://doi.org/10.1177/1745691612454304

Gurteen, P. M., Horne, P. J., & Erjavec, M. (2011). Rapid word learning in 13-and 17-month-olds in a naturalistic two-word procedure: Looking versus reaching measures. *Journal of Experimental Child Psychology*, *109*(2), 201–217. https://doi.org/10.1016/j.jecp.2010.12.001

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid linguistic ambiguity resolution in young children with autism spectrum disorder: Eye tracking evidence for the limits of weak central coherence. *Autism Research*, *8*(6), 717–726. https://doi.org/10.1002/aur.1487

Haith, M. M., Wentworth, N., & Canfield, R. L. (1993). The formation of expectations in early infancy. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (pp. 251–297). Ablex.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.

Hambleton, R. K., Zaal, J. N., & Pieters, J. P. M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 341–366). Kluwer. https://doi.org/10.1007/978-94-009-2195-5_12

Han, Z., He, Q., & von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology*, *10*, 2461. https://doi.org/10.3389/fpsyg.2019.02461

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3173574.3174023

Harlaar, N., Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2008). Why do preschool language abilities correlate with later reading? A twin study. *Journal of Speech, Language, and Hearing Research*, *51*(3), 688–705. https://doi.org/10.1044/1092-4388(2008/049)

Hassan, H. (2020, April 23). Coronavirus: Malaysia extends movement curbs by two weeks to May 12. *The Straits Times*. https://www.straitstimes.com/asia/se-asia/coronavirus-malaysia-pm-muhyiddin-says-to-extend-movement-curbs-by-two-weeks-to-may-12

Hayne, H., Herbert, J., & Simcock, G. (2003). Imitation from television by 24-and 30-month-olds. *Developmental Science*, *6*(3), 254–261. https://doi.org/10.1111/1467-7687.00281

Hendrickson, K., Mitsven, S., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2015). Looking and touching: What extant approaches reveal about the structure of early word knowledge. *Developmental Science*, *18*(5), 723–735. https://doi.org/10.1111/desc.12250

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (in press). lab.js: A free, open, online study builder. *Behavior Research Methods*. https://doi.org/10.31234/osf.io/fqr49

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Henrichs, J., Rescorla, L., Schenk, J. J., Schmidt, H. G., Jaddoe, V. W., Hofman, A., Raat, H., Verhulst, F. C., & Tiemeier, H. (2011). Examining continuity of early expressive vocabulary development: The Generation R Study. *Journal of Speech, Language, and Hearing Research*, *54*(3), 854–869. https://doi.org/10.1044/1092-4388(2010/09-0255)

Hilbig, B. E. (2016). Reaction time effects in lab- versus web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). The intermodal preferential looking paradigm: A window onto emerging language comprehension. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Language, speech, and communication. Methods for assessing children's syntax* (pp. 105–124). MIT Press.

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the Science of Learning. *Psychological Science in the Public Interest*, *16*(1), 3–34. https://doi.org/10.1177/1529100615569721

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

Hourcade, J. P., Bullock-Rest, N. E., & Hansen, T. E. (2012). Multitouch tablet applications and activities to enhance the social skills of children with autism spectrum disorders. *Personal and Ubiquitous Computing*, *16*, 157–168. https://doi.org/10.1007/s00779-011-0383-3

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(4), 701–724. https://doi.org/10.1017/s0305000907008124

Hudson, J. A., & Sheffield, E. G. (1999). The role of reminders in young children's memory development. In L. Balter & C. S. Tamis-LeMonda (Eds.), *Child psychology: A handbook of contemporary issues* (pp. 193–214). Psychology Press.

IBM Cloud Education. (2019). Containerization. https://www.ibm.com/cloud/learn/containerization

Irwin, J. R., Carter, A. S., & Briggs-Gowan, M. J. (2002). The social-emotional development of "late-talking" toddlers. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(11), 1324–1332. https://doi.org/10.1097/00004583-200211000-00014

Janetzko, D. (2017). Nonreactive data collection online. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 76–91). SAGE. https://doi.org/10.4135/9781473957992.n5

JATOS server during the COVID-19 pandemic. (2020, October 29). Retrieved October 31, 2020, from https://www.jatos.org/JATOS-server-during-the-COVID-19-pandemic.html

Jing, M., & Kirkorian, H. L. (2020). Video deficit in children's early learning. In J. Van den Bulck (Ed.), *The international encyclopedia of media psychology*. John Wiley & Sons. https://doi.org/10.1002/9781119011071.iemp0239

Johnson, E. K., & Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychological Research*, *75*, 35–42. https://doi.org/10.1007/s00426-010-0285-4

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods*, *21*(4), 526–541. https://doi.org/10.1037/met0000099

Kaler, S. R., & Kopp, C. B. (1990). Compliance and comprehension in very young toddlers. *Child Development*, *61*(6), 1997–2003. https://doi.org/10.2307/1130853

Kan, P. F., & Kohnert, K. (2008). Fast mapping by bilingual preschool children. *Journal of Child Language*, *35*(3), 495–514. https://doi.org/10.1017/S0305000907008604

Kartushina, N., & Mayor, J. (2019). Word knowledge in six-to nine-month-old Norwegian infants? Not without additional frequency cues. *Royal Society Open Science*, *6*(9), 180711. https://doi.org/10.1098/rsos.180711

Kemp, N., Scott, J., Bernhardt, B. M., Johnson, C. E., Siegel, L. S., & Werker, J. F. (2017). Minimal pair word learning and vocabulary size: Links with later language skills. *Applied Psycholinguistics*, *38*(2), 289–314. https://doi.org/10.1017/S0142716416000199

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PLoS ONE*, *14*(9), e0221802. https://doi.org/10.1371/journal.pone.0221802

Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen? *Child Development Perspectives*, *12*(3), 210–214. https://doi.org/10.1111/cdep.12290

Kirkorian, H. L., Anderson, D. R., & Keen, R. (2012). Age differences in online processing of video: An eye movement study. *Child Development*, *83*(2), 497–507. https://doi.org/10.1111/j.1467-8624.2011.01719.x

Kirkorian, H. L., Choi, K., & Pempek, T. A. (2016). Toddlers' word learning from contingent and noncontingent video on touch screens. *Child Development*, *87*(2), 405–413. https://doi.org/10.1111/cdev.12508

Kirkorian, H. L., Lavigne, H. J., Hanson, K. G., Troseth, G. L., Demers, L. B., & Anderson, D. R. (2016). Video deficit in toddlers' object retrieval: What eye movements reveal about online cognition. *Infancy*, *21*(1), 37–64. https://doi.org/10.1111/infa.12102

Kirkorian, H. L., Pempek, T. A., & Choi, K. (2016). The role of online processing in young children's learning from interactive and noninteractive digital media. In R. Barr & D. N. Linebarger (Eds.), *Media exposure during infancy and early childhood* (pp. 65–89). Springer. https://doi.org/10.1007/978-3-319-45102-2_5

Klesty, V., & Fouche, G. (2020, March 24). Norway extends coronavirus curbs until April 13. *Reuters*. https://uk.reuters.com/article/us-health-coronavirus-norway-restriction/norway-extends-coronavirus-curbs-until-april-13-idUSKBN21B2ED

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 264–269. https://doi.org/10.3758/bf03204824

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 35–60). Academic Press. https://doi.org/10.1016/B978-012099980-4/50003-4

Krcmar, M., Grela, B., & Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. *Media Psychology*, *10*(1), 41–63. https://doi.org/10.1080/15213260701300931

Kucirkova, N. (2014). iPads in early education: Separating assumptions and evidence. *Frontiers in Psychology*, *5*, 715. https://doi.org/10.3389/fpsyg.2014.00715

Kyllonen, P. C. (1991). Principles for creating a computerized test battery. *Intelligence*, *15*(1), 1–15. https://doi.org/10.1016/0160-2896(91)90019-A

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lauricella, A. R., Pempek, T. A., Barr, R., & Calvert, S. L. (2010). Contingent computer interactions for young children's object retrieval success. *Journal of Applied Developmental Psychology*, *31*(5), 362–369. https://doi.org/10.1016/j.appdev.2010.06.002

Law, J., & Roy, P. (2008). Parental report of infant language skills: A review of the development and application of the Communicative Development Inventories. *Child and Adolescent Mental Health*, *13*(4), 198–206. https://doi.org/10.1111/j.1475-3588.2008.00503.x

Lawson, G. M., Hook, C. J., & Farah, M. J. (2018). A meta-analysis of the relationship between socioeconomic status and executive function

performance among children. *Developmental Science*, *21*(2), e12529. https://doi.org/10.1111/desc.12529

Lee, J. (2011). Size matters: Early vocabulary as a predictor of language and literacy competence. *Applied Psycholinguistics*, *32*(1), 69–92. https://doi.org/10.1017/S0142716410000299

Legacy, J., Zesiger, P., Friend, M., & Poulin-Dubois, D. (2016). Vocabulary size, translation equivalents, and efficiency in word recognition in very young bilinguals. *Journal of Child Language*, *43*(4), 760–783. https://doi.org/10.1017/S0305000915000252

Legacy, J., Zesiger, P., Friend, M., & Poulin-Dubois, D. (2018). Vocabulary size and speed of word recognition in very young French–English bilinguals: A longitudinal study. *Bilingualism: Language and Cognition*, *21*(1), 137–149. https://doi.org/10.1017/S1366728916000833

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.4.5) [R package]. https://CRAN.R-project.org/package=emmeans

Lewis, J., & Fowler, M. (2014). Microservices. https://martinfowler.com/articles/microservices.html

Litman, L., Robinson, J., Rosen, Z., Rosenzweig, C., Waxman, J., & Bates, L. M. (2020). The persistence of pay inequality: The gender pay gap in an anonymous online labor market. *PLoS ONE*, *15*(2), e0229383. https://doi.org/10.1371/journal.pone.0229383

Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolík, F., Butcher, M., Chondrogianni, V., Hreich, E. K., Messarra, C., Razak, R. A., Treffers-Daller, J., Yap, N. T., Abboud, L., Talebi, A., Gureghian, M., Tuller, L., & Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *PLoS ONE*, *14*(8), e0220611. https://doi.org/10.1371/journal.pone.0220611

Luyster, R., Lopez, K., & Lord, C. (2007). Characterizing communicative development in children referred for autism spectrum disorders using the

MacArthur–Bates Communicative Development Inventory (CDI). *Journal of Child Language*, *34*(3), 623–654. https://doi.org/10.1017/s0305000907008094

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI:WS). *Journal of Speech, Language, and Hearing Research*, *59*(2), 281–289. https://doi.org/10.1044/2015_JSLHR-L-15-0202

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–F16. https://doi.org/10.1111/j.1467-7687.2008.00671.x

Marjanovič-Umek, L., Fekonja-Peklaj, U., & Podlesek, A. (2013). Characteristics of early vocabulary and grammar development in Slovenian-speaking infants and toddlers: A CDI-adaptation study. *Journal of Child Language*, *40*(4), 779–798. https://doi.org/10.1017/S0305000912000244

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122. https://doi.org/10.1037/a0032108

Marsh, J., Plowman, L., Yamada-Rice, D., Bishop, J., Lahmar, J., Scott, F., Davenport, A., Davis, S., French, K., Piras, M., Robinson, P., Thornhill, S., & Winter, P. (2015). *Exploring play and creativity in pre-schooler's use of apps: Final project report.* http://www.techandplay.org/reports/TAP_Final_Report.pdf

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of*

Memory and Language, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mayne, A. M., Yoshinaga-Itano, C., & Sedey, A. L. (1999). Receptive vocabulary development of infants and toddlers who are deaf or hard of hearing. *Volta Review*, *100*, 29–52.

Mayne, A. M., Yoshinaga-Itano, C., Sedey, A. L., & Carey, A. (1998). Expressive vocabulary development of infants and toddlers who are deaf or hard of hearing. *Volta Review*, *100*, 1–28.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur–Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, *51*(5), 2248–2255. https://doi.org/10.3758/s13428-018-1146-0

McGuigan, N., Whiten, A., Flynn, E., & Horner, V. (2007). Imitation of causally necessary versus unnecessary tool use by 3-and 5-year-old children. *Cognitive Development*, *22*(3), 353–364. https://doi.org/10.1016/j.cogdev.2007.01.001

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of the psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 614–620. https://doi.org/10.3758/BF03195541

Miller, R., Schmidt, K., Kirschbaum, C., & Enge, S. (2018). Comparability, stability, and reliability of internet-based mental chronometry in domestic and laboratory settings. *Behavior Research Methods*, *50*(4), 1345–1358. https://doi.org/10.3758/s13428-018-1036-5

Mindell, J. A., Sadeh, A., Wiegand, B., How, T. H., & Goh, D. Y. (2010). Cross-cultural differences in infant and toddler sleep. *Sleep Medicine*, *11*(3), 274–280. https://doi.org/10.1016/j.sleep.2009.04.012

Mohd Don, Z., Knowles, G., & Yong, J. (2008). How words can be misleading: A study of syllable timing and "stress" in Malay. *Linguistics Journal*, *3*(2), 66–81.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Development*, *86*(5), 1351–1370. https://doi.org/10.1111/cdev.12398

Mumme, D. L., & Fernald, A. (2003). The infant as onlooker: Learning from emotional reactions observed in a television scenario. *Child Development*, *74*(1), 221–237. https://doi.org/10.1111/1467-8624.00532

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. *Psychological experiments on the internet* (pp. 61–87). Elsevier. https://doi.org/10.1016/B978-012099980-4/50004-6

Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017). Baby FaceTime: Can toddlers learn from online video chat? *Developmental Science*, *20*(4), e12430. https://doi.org/10.1111/desc.12430

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods*, *43*(2), 353–362. https://doi.org/10.3758/s13428-011-0069-9

NetMarketShare. (2021). Browser market share. https://bit.ly/2N0JYIm

Neumann, M. M., Worrall, S., & Neumann, D. L. (2019). Validation of an expressive and receptive tablet assessment of early literacy. *Journal of Research on Technology in Education*, *51*(4), 326–341. https://doi.org/10.1080/15391523.2019.1637800

Nielsen, M., Simcock, G., & Jenkins, L. (2008). The effect of social engagement on 24-month-olds' imitation from live and televised models. *Developmental Science*, *11*(5), 722–731. https://doi.org/10.1111/j.1467-7687.2008.00722.x

Ofcom. (2012). *Children and parents: Media use and attitudes report*. Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0020/56324/main.pdf

Ofcom. (2013). *Children and parents: Media use and attitudes report*. Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0018/53514/research07oct2013.pdf

Ofcom. (2014). *Children and parents: Media use and attitudes report*. Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0027/76266/childrens_2014_report.pdf

Ofcom. (2020). *Children and parents: Media use and attitudes report*. Office of Communications. https://www.ofcom.org.uk/__data/assets/pdf_file/0023/190616/children-media-use-attitudes-2019-report.pdf

Ogston, P. L., Mackintosh, V. H., & Myers, B. J. (2011). Hope and worry in mothers of children with an autism spectrum disorder or Down syndrome. *Research in Autism Spectrum Disorders*, *5*(4), 1378–1384. https://doi.org/10.1016/j.rasd.2011.01.020

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-Lemonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, *31*(3), 587–608. https://doi.org/10.1017/s0305000904006270

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI)* (pp. 3839–3845).

Parish-Morris, J., Mahajan, N., Hirsh-Pasek, K., Golinkoff, R. M., & Collins, M. F. (2013). Once upon a time: Parent–child dialogue and storybook reading in the electronic era. *Mind, Brain, and Education*, *7*(3), 200–211. https://doi.org/10.1111/mbe.12028

Partridge, E., McGovern, M. G., Yung, A., & Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In D. C. Noelle,

R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Patrucco-Nanchen, T., Friend, M., Poulin-Dubois, D., & Zesiger, P. (2019). Do early lexical skills predict language outcome at 3 years? A longitudinal study of French-speaking children. *Infant Behavior and Development*, *57*, 101379. https://doi.org/10.1016/j.infbeh.2019.101379

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pérez-Rojas, A. E., Bartholomew, T. T., Lockard, A. J., & González, J. M. (2019). Development and initial validation of the Therapist Cultural Comfort Scale. *Journal of Counseling Psychology*, *66*(5), 534–549. https://doi.org/10.1037/cou0000344

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*(1), 408–411. https://doi.org/10.3758/s13428-015-0577-0

Plant, R. R., & Quinlan, P. T. (2013). Could millisecond timing errors in commonly used equipment be a cause of replication failure in some neuroscience studies? *Cognitive, Affective, & Behavioral Neuroscience*, *13*(3), 598–614. https://doi.org/10.3758/s13415-013-0166-6

Poulin-Dubois, D., Bialystok, E., Blaye, A., Polonia, A., & Yott, J. (2013). Lexical access and vocabulary development in very young bilinguals. *International Journal of Bilingualism*, *17*(1), 57–70. https://doi.org/10.1177/1367006911431198

Poulin-Dubois, D., Kuzyk, O., Legacy, J., Zesiger, P., & Friend, M. (2018). Translation equivalents facilitate lexical access in very young bilinguals. *Bilingualism: Language and Cognition*, *21*(4), 856–866. https://doi.org/10.1017/S1366728917000657

Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2020). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavior Research Methods*, *52*(3), 1371–1382. https://doi.org/10.3758/s13428-019-01321-2

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., & White, L. E. (2012). *Neuroscience*. Sinauer Associates.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354–360. https://doi.org/10.1016/j.chb.2015.12.049

Reich, S. M., Yau, J. C., & Warschauer, M. (2016). Tablet-based eBooks for young children: What does the research say? *Journal of Developmental & Behavioral Pediatrics*, *37*(7), 585–591. https://doi.org/10.1097/DBP.0000000000000335

Reilly, S., Wake, M., Ukoumunne, O. C., Bavin, E., Prior, M., Cini, E., Conway, L., Eadie, P., & Bretherton, L. (2010). Predicting language outcomes at 4 years of age: Findings from Early Language in Victoria Study. *Pediatrics*, *126*(6), e1530–e1537. https://doi.org/10.1542/peds.2010-0254

Reimers, S. (2007). The BBC internet study: General methodology. *Archives of Sexual Behavior*, *36*, 147–161. https://doi.org/10.1007/s10508-006-9143-2

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Reips, U.-D. (2006). Web-based methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 73–85). American Psychological Association. https://doi.org/10.1037/11383-006

Reips, U.-D. (2007). The methodology of internet-based experiments. In A. N. Joinson, K. Y. A. McKenna, T. Postmes, & U.-D. Reips (Eds.), *Oxford handbook of internet psychology* (pp. 373–390). Oxford University Press.

Reips, U.-D., & Lengler, R. (2005). The Web Experiment List: A web service for the recruitment of participants and archiving of internet-based experiments. *Behavior Research Methods*, *37*(2), 287–292. https://doi.org/10.3758/BF03192696

Reiß, M., Krüger, M., & Krist, H. (2019). Theory of mind and the video deficit effect: Video presentation impairs children's encoding and understanding of false belief. *Media Psychology*, *22*(1), 23–38. https://doi.org/10.1080/15213269.2017.1412321

Rescorla, L., & Dale, P. S. (2013). *Late talkers: Language development, interventions, and outcomes*. Brookes.

Rescorla, L., Ratner, N. B., Jusczyk, P., & Jusczyk, A. M. (2005). Concurrent validity of the Language Development Survey: Associations with the MacArthur–Bates Communicative Development Inventories. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/1058-0360(2005/016)

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. *Progress in Brain Research*, *253*, 243–262. https://doi.org/10.1016/bs.pbr.2020.06.005

Reznick, J. S., & Goldfield, B. A. (1994). Diary vs. representative checklist assessment of productive vocabulary. *Journal of Child Language*, *21*(2), 465–472. https://doi.org/10.1017/s0305000900009351

Rideout, V. (2017). *The Common Sense census: Media use by kids age zero to eight*. Common Sense Media.

Rinaldi, P., Pasqualetti, P., Stefanini, S., Bello, A., & Caselli, M. C. (2019). The Italian Words and Sentences MB-CDI: Normative data and concordance between complete and short forms. *Journal of Child Language*, *46*(3), 546–566. https://doi.org/10.1017/S0305000919000011

Ring, E. D., & Fenson, L. (2000). The correspondence between parent report and child performance for receptive and expressive vocabulary beyond infancy. *First Language*, *20*, 141–159. https://doi.org/10.1177/014272370002005902

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Development*, *85*(3), 956–970. https://doi.org/10.1111/cdev.12166

Roseberry, S., Hirsh-Pasek, K., Parish-Morris, J., & Golinkoff, R. M. (2009). Live action: Can young children learn verbs from video? *Child Development*, *80*(5), 1360–1375. https://doi.org/10.1111/j.1467-8624.2009.01338.x

Ruggeri, A., Markant, D. B., Gureckis, T. M., & Xu, F. (2016). Active control of study leads to improved recognition memory in children. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Russo-Johnson, C., Troseth, G., Duncan, C., & Mesghina, A. (2017). All tapped out: Touchscreen interactivity and young children's word learning. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00578

Schmidt, M. E., Crawley-Davis, A. M., & Anderson, D. R. (2007). Two-year-olds' object retrieval based on television: Testing a perceptual account. *Media Psychology*, *9*(2), 389–409. https://doi.org/10.1080/15213260701291346

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 274–279. https://doi.org/10.3758/BF03204826

Schmidt, W. C. (2001). Presentation accuracy of web animation methods. *Behavior Research Methods, Instruments, & Computers*, *33*(2), 187–200. https://doi.org/10.3758/BF03195365

Schmitt, K. L., & Anderson, D. R. (2002). Television and reality: Toddlers' use of visual information from video to guide behavior. *Media Psychology*, *4*(1), 51–76. https://doi.org/10.1207/s1532785xmep0401_03

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS ONE*, *8*(6), e67769. https://doi.org/10.1371/journal.pone.0067769

Semmelmann, K., Hönekopp, A., & Weigelt, S. (2017). Looking tasks online: Utilizing webcams to collect video data from home. *Frontiers in Psychology*, *8*, 1582. https://doi.org/10.3389/fpsyg.2017.01582

Semmelmann, K., Nordt, M., Sommer, K., Röhnke, R., Mount, L., Prüfer, H., Terwiel, S., Meissner, T. W., Koldewyn, K., & Weigelt, S. (2016). U can touch this: How tablets can be used to study cognitive development. *Frontiers in Psychology*, *7*, 1021. https://doi.org/10.3389/fpsyg.2016.01021

Semmelmann, K., & Weigelt, S. (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, *49*(4), 1241–1260. https://doi.org/10.3758/s13428-016-0783-4

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Semuels, A. (2018, January 23). The internet is enabling a new kind of poorly paid hell. *The Atlantic*. https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Sesame Workshop. (2012). *Best practices: Designing touch tablet experiences for preschoolers*. Joan Ganz Cooney Center. https://joanganzcooneycenter.org/wp-content/uploads/2020/02/SesameWorkshop-2012.pdf

Sheehan, K. J., & Uttal, D. H. (2016). Children's learning from touch screens: A dual representation perspective. *Frontiers in Psychology*, *7*, 1220. https://doi.org/10.3389/fpsyg.2016.01220

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Li, F.-F., 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

Shuler, C. (2012). *iLearn II: An analysis of the education category of Apple's App Store*. Joan Ganz Cooney Center. https://joanganzcooneycenter.org/wp-content/uploads/2012/01/ilearnii.pdf

Sim, Z. L., Tanner, M. M., Alpert, N. Y., & Xu, F. (2015). Children learn better when they select their own data. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. *First Language*, *34*(1), 3–23. https://doi.org/10.1177/0142723713510997

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *afex: Analysis of factorial experiments* (Version 0.27-2) [R package]. https://CRAN.R-project.org/package=afex

Smeets, D. J., & Bus, A. G. (2015). The interactive animated e-book as a word learning device for kindergartners. *Applied Psycholinguistics*, *36*(4), 899–920. https://doi.org/https://doi.org/10.1017/S0142716413000556

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 496–505. https://doi.org/10.3758/BF03210601

Stieger, S., & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*(6), 1488–1495. https://doi.org/10.1016/j.chb.2010.05.013

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Stolt, S., Haataja, L., Lapinleimu, H., & Lehtonen, L. (2009). Associations between lexicon and grammar at the end of the second year in Finnish children. *Journal of Child Language*, *36*(4), 779–806. https://doi.org/10.1017/S0305000908009161

Strouse, G. A., & Ganea, P. A. (2017). Parent–toddler behavior and language differ when reading electronic and print picture books. *Frontiers in Psychology*, *8*, 677. https://doi.org/10.3389/fpsyg.2017.00677

Strouse, G. A., & Samson, J. E. (2021). Learning from video: A meta-analysis of the video deficit in children ages 0 to 6 years. *Child Development*, *92*(1), e20–e38. https://doi.org/10.1111/cdev.13429

Strouse, G. A., & Troseth, G. L. (2008). "Don't try this at home": Toddlers' imitation of new skills from people on video. *Journal of Experimental Child Psychology*, *101*(4), 262–280. https://doi.org/10.1016/j.jecp.2008.05.010

Strouse, G. A., Troseth, G. L., O'Doherty, K. D., & Saylor, M. M. (2018). Co-viewing supports toddlers' word learning from contingent and noncontingent video. *Journal of Experimental Child Psychology*, *166*, 310–326. https://doi.org/10.1016/j.jecp.2017.09.005

Styles, S., & Plunkett, K. (2008). What is 'word understanding' for the parent of a one-year-old? Matching the difficulty of a lexical comprehension task to

parental CDI report. *Journal of Child Language*, *36*(4), 895–908. https://doi.org/10.1017/S0305000908009264

Suddendorf, T. (2003). Early representational insight: Twenty-four-month-olds can use a photo to find an object in the world. *Child Development*, *74*(3), 896–904. https://doi.org/10.1111/1467-8624.00574

Suddendorf, T., Simcock, G., & Nielsen, M. (2007). Visual self-recognition in mirrors and live videos: Evidence for a developmental asynchrony. *Cognitive Development*, *22*(2), 185–196. https://doi.org/10.1016/j.cogdev.2006.09.003

Sumer, B., Grabitz, C., & Küntay, A. (2017). Early produced signs are iconic: Evidence from Turkish Sign Language. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 3273–3278). Cognitive Science Society.

Swingley, D., & Fernald, A. (2002). Recognition of words referring to present and absent objects by 24-month-olds. *Journal of Memory and Language*, *46*(1), 39–56. https://doi.org/10.1006/jmla.2001.2799

Szagun, G., Stumper, B., & Schramm, S. A. (2014). *Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform) [Questionnaire on Early Language Development (FRAKIS) and FRAKIS-K (Short Form Version)]*. Pearson Assessment.

Takacs, Z. K., Swart, E. K., & Bus, A. G. (2015). Benefits and pitfalls of multimedia and interactive features in technology-enhanced storybooks: A meta-analysis. *Review of Educational Research*, *85*(4), 698–739. https://doi.org/https://doi.org/10.3102/0034654314566989

Tardif, T., Fletcher, P., Liang, W.-L., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. *Journal of Child Language*, *36*(5), 1115–1144. https://doi.org/10.1017/S0305000908009185

Tardif, T., Fletcher, P., Zhang, Z.-X., & Liang, W.-L. (2008). *The Chinese Communicative Development Inventory (Putonghua and Cantonese versions): Manual, forms, and norms*. Peking University Medical Press.

Thal, D., DesJardin, J. L., & Eisenberg, L. S. (2007). Validity of the MacArthur–Bates Communicative Development Inventories for measuring language abilities in children with cochlear implants. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/1058-0360(2007/007)

The OWASP Foundation. (2017). *OWASP Top 10 - 2017*. https://owasp.org/www-pdf-archive/OWASP_Top_10-2017_%5C%28en%5C%29.pdf.pdf

Tomasello, M., & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development*. https://doi.org/10.1111/j.1540-5834.1994.tb00186.x

Troseth, G. L. (2003). TV guide: Two-year-old children learn to use video as a source of information. *Developmental Psychology*, *39*(1), 140–150. https://doi.org/10.1037/0012-1649.39.1.140

Troseth, G. L. (2010). Is it life or is it Memorex? Video as a representation of reality. *Developmental Review*, *30*(2), 155–175. https://doi.org/10.1016/j.dr.2010.03.007

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Development*, *69*(4), 950–965. https://doi.org/10.1111/j.1467-8624.1998.tb06153.x

Troseth, G. L., Flores, I., & Stuckelman, Z. D. (2019). When representation becomes reality: Interactive digital media and symbolic development. *Advances in Child Development and Behavior*, *56*, 65–108. https://doi.org/10.1016/bs.acdb.2018.12.001

Troseth, G. L., Pierroutsakos, S. L., & DeLoache, J. S. (2004). From the innocent to the intelligent eye: The early development of pictorial competence. *Advances in Child Development and Behavior*, *32*, 1–35. https://doi.org/10.1016/s0065-2407(04)80003-x

Troseth, G. L., Russo, C. E., & Strouse, G. A. (2016). What's next for research on young children's interactive media? *Journal of Children and Media*, *10*(1), 54–62. https://doi.org/10.1080/17482798.2015.1123166

Troseth, G. L., Saylor, M. M., & Archer, A. H. (2006). Young children's use of video as a source of socially relevant information. *Child Development*, *77*(3), 786–799. https://doi.org/10.1111/j.1467-8624.2006.00903.x

Troseth, G. L., Strouse, G. A., Verdine, B. N., & Saylor, M. M. (2018). Let's chat: On-screen social responsiveness is not sufficient to support toddlers' word learning from video. *Frontiers in Psychology*, *9*, 2195. https://doi.org/10.3389/fpsyg.2018.02195

Twomey, D. M., Wrigley, C., Ahearne, C., Murphy, R., De Haan, M., Marlow, N., & Murray, D. M. (2018). Feasibility of using touch screen technology for early cognitive assessment in children. *Archives of Disease in Childhood*, *103*(9), 853–858. https://doi.org/10.1136/archdischild-2017-314010

Vales, C., & Fisher, A. V. (2019). When stronger knowledge slows you down: Semantic relatedness predicts children's co-activation of related items in a visual search paradigm. *Cognitive Science*, *43*(6), e12746. https://doi.org/10.1111/cogs.12746

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-18360-5

van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. Springer.

van der Zee, T., & Reich, J. (2018). Open Education Science. *AERA Open*, *4*(3), 1–15. https://doi.org/10.1177/2332858418787466

van Oostendorp, M. (2020). Germanic syllable structure. In M. T. Putnam & B. R. Page (Eds.), *The Cambridge handbook of Germanic linguistics* (pp. 33–48). Cambridge University Press.

Vlach, H. A., & Sandhofer, C. M. (2011). Developmental differences in children's context-dependent word learning. *Journal of Experimental Child Psychology*, *108*(2), 394–401. https://doi.org/10.1016/j.jecp.2010.09.011

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*(4), 569–586. https://doi.org/10.1016/j.jecp.2012.08.001

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, *45*(1), 108–115. https://doi.org/10.3758/s13428-012-0224-y

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition*. Pearson Assessment.

Wechsler, D. (2009). *Wechsler Memory Scale–Fourth Edition*. Pearson Assessment.

Westerlund, M., Berglund, E., & Eriksson, M. (2006). Can severely language delayed 3-year-olds be identified at 18 months? Evaluation of a screening version of the MacArthur–Bates Communicative Development Inventories. *Journal of Speech, Language, and Hearing Research*, *49*(2), 237–247. https://doi.org/10.1044/1092-4388(2006/020)

Williams, K. T. (2018). *Expressive Vocabulary Test–Third Edition*. Pearson Assessment.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice*, *8*(3), 5–10. https://doi.org/10.1111/j.1745-3992.1989.tb00324.x

Woolfe, T., Herman, R., Roy, P., & Woll, B. (2010). Early vocabulary development in deaf native signers: A British Sign Language adaptation of the Communicative Development Inventories. *Journal of Child Psychology and Psychiatry*, *51*(3), 322–331. https://doi.org/10.1111/j.1469-7610.2009.02151.x

World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020.

https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Yoder, P. J., Warren, S. F., & Biggar, H. A. (1997). Stability of maternal reports of lexical comprehension in very young children with developmental delays. *American Journal of Speech-Language Pathology*, *6*(1), 59–64. https://doi.org/10.1044/1058-0360.0601.59

Yoshida, H., Tran, D., Benitez, V., & Kuwabara, M. (2011). Inhibition and adjective learning in bilingual and monolingual children. *Frontiers in Psychology*, *2*, 210. https://doi.org/10.3389/fpsyg.2011.00210

Zambrana, I. M., Pons, F., Eadie, P., & Ystrom, E. (2014). Trajectories of language delay from age 3 to 5: Persistence, recovery and late onset. *International Journal of Language & Communication Disorders*, *49*(3), 304–316. https://doi.org/10.1111/1460-6984.12073

Zettersten, M., & Saffran, J. R. (2019). Sampling to learn words: Adults and children sample words that reduce referential ambiguity. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the Cognitive Science Society* (pp. 1261–1267). Cognitive Science Society.

APPENDIX A. OVERVIEW OF ACTIVELY MAINTAINED TOOLS FOR WEB-BASED STUDIES

Table A.1

Overview of Actively Maintained Tools for Web-Based Studies

Type	Open source	Cost	Main features
Experiment builder			
jsPsych[1]	✓	Free	– jsPsych (de Leeuw, 2015) is a JavaScript library that provides a variety of pre-programmed components (termed *plugins*) to simplify the programming of common tasks in behavioural experiments (e.g., measuring RT, displaying text instructions, and displaying stimuli). – The *core* library handles the execution of experiments. – An empty plugin template is provided to allow new plugins to be created.
lab.js[2]	✓	Free	– lab.js (Henninger et al., in press) is a web application (i.e., no installation is required). – A GUI is provided. This means that experiments can be programmed without writing a single line of code. – Alternatively, an HTML editor is provided to allow forms and questionnaires to be created and also to increase the flexibility in designing experiments (e.g., automatic scaling of web page contents to adapt to different screen sizes).

Continued on next page

Type	Open source	Cost	Main features
			– Every component of an experiment can be exported to be reused in other experiments. – Experiment components can be further customised (e.g., to add custom logic) through JavaScript. – Code base is open, thus lab.js can be customised and extended as needed.
OpenSesame Web[3]	✓	Free	– OpenSesame (Mathôt et al., 2012) is a desktop application that provides a GUI for creating experiments. – Programming of complex tasks is possible through JavaScript (online experiments) or Python (offline experiments) scripts. – OpenSesame Web is a JavaScript library that enables experiments created with OpenSesame to be run online.
PsychoPy+PsychoJS[4]	✓	Free	– PsychoPy (Peirce et al., 2019) is a desktop application that provides both a GUI (*Builder*) and a code editor (*Coder*), for creating experiments. – Programming of complex tasks is possible through Python scripts. – PsychoPy3, together with PsychoJS (PsychoPy's JavaScript library), allows "standard" experiments (i.e., experiments using images, text, and keyboards) to be exported as online experiments.

Continued on next page

Type	Open source	Cost	Main features
Tatool Web[5]	✓	Free	– Tatool (von Bastian et al., 2013) began as a Java-based desktop application and is superseded by Tatool Web (the online version). – The *Experiment Editor* provides a GUI for creating experiments. – The *Task Library* provides a range of experimental paradigms (e.g., Stroop, Choice RT, Item Recognition) that can be customised and/or combined to be used in designing an experiment. – Custom tasks can be programmed (using JavaScript, HTML, and CSS) and used on a local Tatool instance. To use custom tasks on Tatool Web, one can either make the tasks publicly available in the Tatool Task Library in exchange for free hosting on Tatool Web or pay a small hosting fee.
Study management system			
JATOS[6]	✓	Free	– "Just Another Tool for Online Studies" (JATOS; Lange et al., 2015) is a GUI-based web application that allows experiments to be hosted, run, and managed on researchers' own servers. – Its GUI allows easy communication with the server and the database, thus eliminating the need for using the commonly used command line interface. – Multiuser access is supported through individual password-protected accounts.

Continued on next page

Type	Open source	Cost	Main features
			– JATOS is straightforward to set up locally, while some technical knowledge is required to install it on a server. – JATOS currently offers its server for free (until January 2021) to support the scientific community during the COVID-19 pandemic ("JATOS server during the COVID-19 pandemic", 2020).
Open Lab[7]	✓	Mixed	– Open Lab is integrated with lab.js and allows experiments created with lab.js to be hosted and run online. – Experiments, participants, and data collected can be easily managed through its GUI. – Different pricing plans are available. The basic version (one experiment with a maximum of 300 participants) is free to use.
Pavlovia[8]	✗	Paid	– Pavlovia is a web-based platform to host, run, and manage online experiments as well as to manage participants and data. – Experiments created using a variety of experiment builders are supported, including PsychoPy+PsychoJS, jsPsych, and lab.js. – A repository of online experiments (that are made public) is provided, enabling access to an experiment and its code base.
Participant recruitment services			

Continued on next page

Type	Open source	Cost	Main features
MTurk[9]	-	Paid	– MTurk is a crowdsourcing marketplace that allows various tasks (including survey and experiment participation) to be outsourced to online workers on a pay-per-task model. – MTurk charges a 20% fee on the amount paid to workers and an additional 20% if the task is to be assigned to 10 or more workers. – MTurk can either be accessed using a GUI (i.e., the Requester user interface), a command line interface (i.e., a text-based interface that offers more flexibility), or the application programming interface (API; allows MTurk functions to be integrated programmatically). – A large set of templates are provided to get service requesters started with task creation. – An active pool of potential participants is maintained. – A reputation system is employed to ensure data quality. – A basic built-in online survey tool is provided.
Prime Panels[10]	-	Paid	– Prime Panels is an aggregate of online research panels that features a vast participant base.

Continued on next page

Type	Open source	Cost	Main features
Prolific[11]	-	Paid	– Very narrow segments of the population (e.g., people who are in therapy and have attended at least three sessions; Pérez-Rojas et al., 2019) can be sampled from multiple sample providers. – Samples are more diverse and are less familiar with common behavioural science experimental manipulations compared to MTurk samples (Chandler et al., 2019). – An active pool of potential participants is maintained. – Prescreening methods are employed to ensure data quality. – Prolific (Palan & Schitter, 2018) is developed with researchers in mind, to provide a subject pool for research. – Over 100 demographic filters are provided to prescreen participants. – In respect of fair pay, Prolific requires participants to be paid a minimum wage of £5.00/$6.50 per hour. A 33% service fee is charged on top of the amount paid to participants. – Samples are more diverse and are less familiar with common behavioural science experimental manipulations compared to MTurk samples (Peer et al., 2017).

Continued on next page

Type	Open source	Cost	Main features
Sona Systems[12]	-	Paid	– Nationally representative samples are available (the United Kingdom and the United States only). – An active pool of potential participants is maintained. – Sona is a participant pool management system for universities. – Prescreening ensures only eligible participants can take part in studies. – Sona allows researchers to manage the allocation of course credits to participants. – Sona is integrated with many popular third party applications (e.g., Qualtrics, Inquisit, LimeSurvey). – An API is available to allow Sona functions to be integrated programmatically into custom applications. – A basic built-in online survey tool is provided.
Integrated services			
Gorilla[13]	✗	Paid	– Gorilla (Anwyl-Irvine, Massonnié, et al., 2020) is a complete development platform on which experiments can be created, hosted, run, and managed online.

Continued on next page

Type	Open source	Cost	Main features
			– Gorilla features three GUIs: the *Questionnaire Builder*, the *Task Builder*, and the *Experiment Builder*. The Experiment Builder allows the logic of an experiment (constructed using Questionnaires or Tasks or a combination of both) to be defined. Thus, complex experiment/task designs can be achieved without writing a single line of code. – Coding is also possible through the *Code Editor* (for programming an entire experiment from scratch), *Task Builder Scripts* (for adding custom scripts into Tasks), or the *Questionnaire Script Widget* (for enhancing Questionnaires). – Gorilla is integrated with a number of participant recruitment systems, such as MTurk, Prolific, and Sona. – Ready-to-use samples and a wide range of classic tasks, including attention, cognition, decision making, executive function, etc. are provided. – Builder interfaces and code editor are free to use but Gorilla charges $1.08 per participant. Academic subscriptions are also available.
Inquisit Web[14]	✗	Paid	– Inquisit is a desktop application for designing and running psychological experiments and measures, either offline (i.e., locally) or online (via Inquisit Web).

Continued on next page

Type	Open source	Cost	Main features
			– Experiments are programmed using Inquisit's own scripting language that is easier to use in comparison to HTML and JavaScript. – To run an experiment online, the experiment scripts are uploaded to the Millisecond server and then accessed from the Inquisit Web app (this needs to be downloaded by the participant). This allows experiments to be run using the high-performance native system components, thereby acheiving timing accuracy that is superior to JavaScript. – The *Millisecond Test Library* features hundreds of well-known cognitive tests and neuropsychological paradigms.
Labvanced[15]	Partially	Mixed	– LabVanced (Finger et al., 2017) is a web application that provides a GUI for creating, hosting, running, and managing online experiments and questionnaires. – The *Experiment Library* is where experiments are published. It also features a set of templates to get users started. – Using the *Event System*, complex logic can be implemented without writing a single line of code. – Real-time multiplayer experiments (e.g., economic games) are supported. – Eye-tracking via webcam is supported.

Continued on next page

Type	Open source	Cost	Main features
			– A local instance of LabVanced can be set up (on Windows and Linux machines) to run offline studies. – Different pricing plans are available. The free user license includes one published study, 300 MB storage, and 10 data recordings. – As of 1 September 2020, only the experiment presentation part is open source.[16]
PsyToolkit[17]	✓	Free	– PsyToolkit is available online (Stoet, 2017) and offline (Linux; Stoet, 2010). – An extensive library of cognitive psychological experiments and psychological questionnaires allows online experiments or questionnaires to be quickly set up. – Experiments or questionnaires that are not available in the library can be programmed using a dedicated scripting language. – Online experiments and questionnaires are hosted on the Psytoolkit web server.
Testable[18]	✗	Mixed	– Testable (Rezlescu et al., 2020) is a web application that allows users to create, host, run, and manage online experiments and questionnaires.

Continued on next page

Type	Open source	Cost	Main features
			– Experiments and questionnaires can be quickly created by filling in a natural language form or for further customisation, by filling in a spreadsheet with information on trials and questions. – Experimental logic (e.g., conditional cases, the staircase procedure) can also be implemented using the same spreadsheet. – Ready-to-use templates are available in the *Testable Library*. This is also where experiments can be publicly shared. – Participants can be recruited via *Testable Minds*, Testable's own participant pool for psychology experiments (chargeable). – Real-time multiplayer experiments (e.g., economic games) are supported via *Testable Arena*. – Different pricing plans are available. The basic plan that includes an unlimited number of experiments, 100 MB storage, and 20 data recordings is available for free.

[1] https://www.jspsych.org/ [2] https://lab.js.org/ [3] https://osdoc.cogsci.nl/ [4] https://www.psychopy.org/index.html

[5] https://www.tatool-web.com/ [6] https://www.jatos.org/ [7] https://open-lab.online/ [8] https://pavlovia.org/

[9] https://www.mturk.com/ [10] https://www.cloudresearch.com/products/prime-panels/

[11] https://www.prolific.co/ [12] https://www.sona-systems.com/default.aspx [13] https://gorilla.sc/

[14] https://www.millisecond.com/products/inquisit6/weboverview.aspx [15] https://www.labvanced.com/

[16] https://github.com/Labvanced [17] https://www.psytoolkit.org/ [18] https://www.testable.org/

APPENDIX B. SAMPLE HTML TEMPLATE

```
{% extends "experiments/base.html" %}
{% load static %}
<!-- Change study name here. -->
{% block title %}Online Study{% endblock %}

{% block content %}
<div class="container" id="information">
  <div class="row">
      <div class="col text-center">
        <!-- Change study name here. -->
        <h1>Online Study</h1>
      </div>
  </div>
  <div class="row">
    <div class="col">
      <div class="card">
        <div class="card-body text-justify">
          <p class="card-text">
          <!-- Change content of the welcome page/information sheet here. -->

            Dear parents,<br /><br />

            Welcome to the Online Study.<br /><br />

            If you wish to participate in this study with your child, please
            carefully go through the following information about the study:<br />
            - The aim of this study is to XXX. <br />
            - To be eligible to participate in this study, your child must be XXX
              years old.<br />
            - In order to evaluate this online study, we will need video
              recordings and these will be recorded using your computer's
              webcam. Thus, to participate, you must be using a computer or a
              laptop with a webcam and be ready to allow access to the
              webcam for recording. The videos are transmitted via a secure
              connection (TLS encryption, 256 bit) directly to the university's
              servers, where they are stored under the highest security
              standards.<br />
            - During the study, your child needs to be seated so that he/she
              can be properly seen on the webcam recording. <br />
            - We will ask you a few questions beforehand and your personal
              data will be stored separately from the data and videos of the
```

```
42              study.<br />
43            - The study is only compatible with Firefox and Google Chrome
44              browsers. Please use one of these browsers. <br />
45            - You may withdraw from the study at any time without providing
46              a reason. During the entire study, an "Exit" button will be
47              visible at the bottom right of the screen. Click on this if in any
48              case you wish to terminate the study. <br />
49            - You can also request for your data to be deleted at any time. To
50              do so, please send an email to XXX and state the exact name
51              you entered in the participant form which will be presented next.
52              <br /><br />
53
54            If you agree to participate in the study, please click on "Next"
55            below. Before we begin, we will ask you a few more questions and
56            carry out some technical checks. <br /><br />
57
58            We look forward to your participation!
59          </p>
60          <form action="{% url 'experiments:browserCheck'  experiment.id %}"
61            method="post">
62            {% csrf_token %}
63            <div class="text-center">
64              <button type="submit" class="btn btn-primary"
65                id="nextbutton">Next</button>
66            </div>
67          </form>
68        </div>
69      </div>
70    </div>
71  </div>
72</div>
73{% endblock %}
```

APPENDIX C. GERMAN PHONOTACTIC RULES AND CONSTRAINTS

All novel words used in the study obey the phonotactic rules and constraints of German. These words are all disyllabic and stressed on the first syllable:

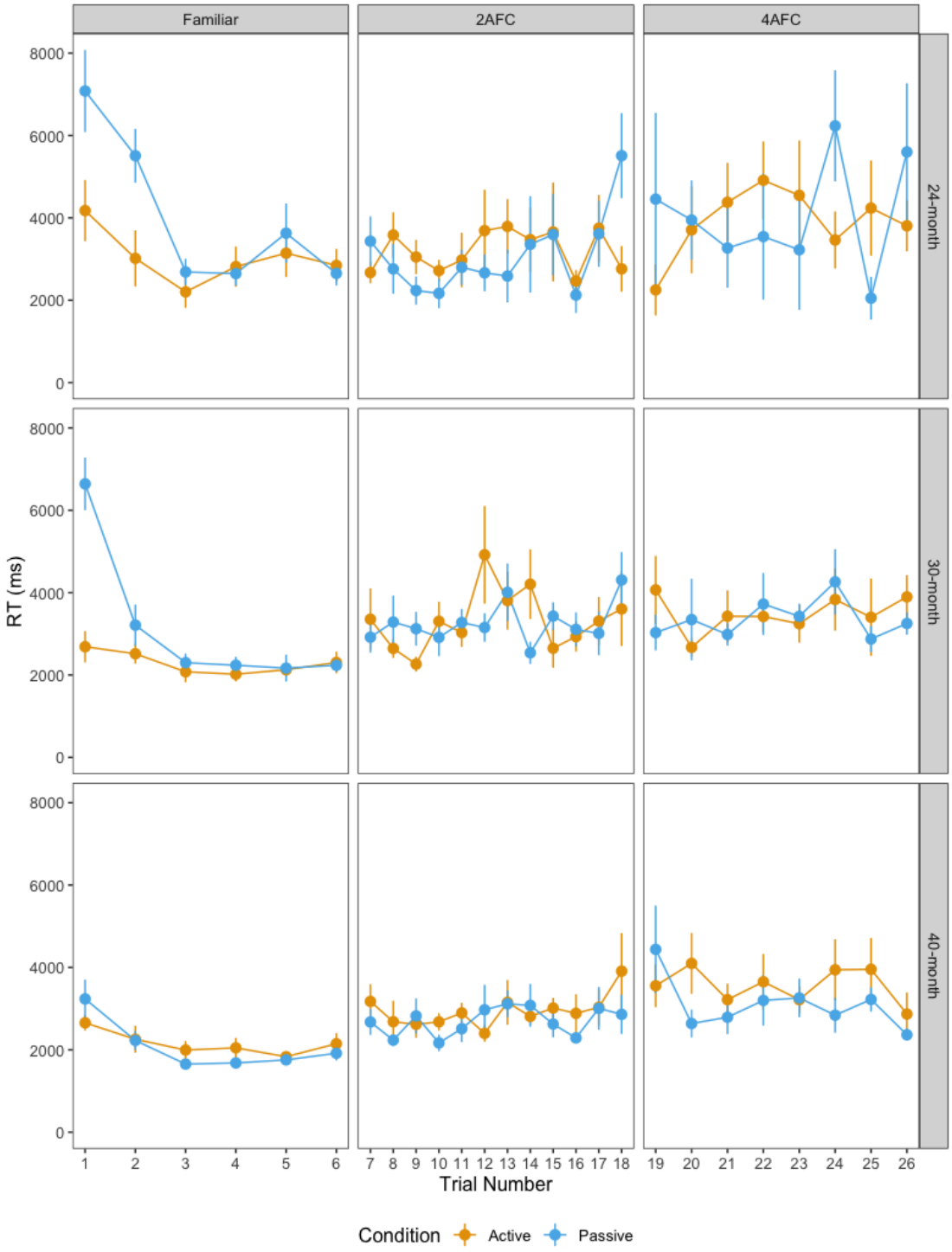
['batʃa], ['foːma], ['koːlat], ['vidɛks]

As outlined in van Oostendorp (2020), German syllables consist of a consonant onset, a vocalic nucleus, and a consonant coda. However, the nuclear vowel is obligatory, so that an empty coda, as in the first syllable of "Kolat" and "Foma", as well as the second syllable of "Batscha" and "Foma", is acceptable. The consonant clusters used in the novel words are also common in words that young children encounter in their everyday lexical environment: [tʃ] in "Batscha" appears in words like "Rutsche" [slide] or "Matsch" [mud], while [ks] in "Widex" appears in "Hexe" [witch] and "sechs" [six].

APPENDIX D. SUPPLEMENTARY FIGURES FOR STUDY 1A

Figure D.1

RT by Trial Number and Age Group



Note. Only trials in which children responded correctly are considered.

Figure D.2

Accuracy by Trial Number and Age Group

Note. Dashed line represents chance (.50) in the familiar and 2AFC test phases; dotted line represents chance (.25) in the 4AFC test phase.

APPENDIX E. MALAY PHONOTACTIC RULES AND CONSTRAINTS

All novel words used in the study obey the phonotactic rules and constraints of Malay. These words are all disyllabic with no lexical stresses but instead with a rise-fall pitch movement (where its start is indicated by [']):

['banuŋ], ['ifi], ['mipo], ['pafka]

A majority of the Malay lexicon is based on disyllabic root morphemes (Adelaar, 1992). In general, the Malay accent lacks stress but is instead, characterised by various intonation patterns, such as a rise-fall pitch movement that is commonly found across the penultimate and final syllables of a word (Mohd Don et al., 2008). As outlined in Clynes and Deterding (2011), syllables have the C_1VC_2 structure, where both C_1 and C_2 are optional consonants and V is a monophthong. Thus, the syllables making up the words "banung", "ifi", and "pafka" are valid. While in the native lexis, only /i/, /u/, and /a/ are allowed in final open syllables (Clynes & Deterding, 2011), /o/ in "mipo" is also found in loanwords like "solo" [solo] and "koko" [cocoa].

APPENDIX F. VISUAL STIMULI IN THE TEST PHASE





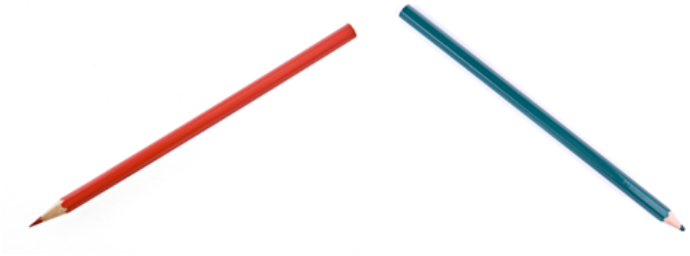



















APPENDIX G. VISUAL STIMULI IN THE FAMILIARISATION PHASE








APPENDIX H. COMPARISONS BETWEEN THE IRT VERSION AND THE ORIGINAL VERSION ACROSS BOTH SEXES AND DIFFERENT TEST LENGTHS ON THE CDI–WS, WITH RANDOM LISTS AS BASELINE

Table H.1

Comparisons Between the IRT Version and the Original Version Across Both Sexes and Different Test Lengths on the American English CDI–WS, With Random Lists as Baseline

Length	Females			Males			Baseline		
	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.	r with full CDI	Avg. *SE*	Rel.
680	.988 (.988)	.03 (.03)	.999 (.999)	.989 (.989)	.03 (.03)	.999 (.999)	1.000	.00	1.000
400	.990 (.987)	.03 (.03)	.999 (.999)	.990 (.987)	.03 (.03)	.999 (.999)	.998	.01	1.000
200	.988 (.985)	.03 (.04)	.999 (.999)	.989 (.985)	.04 (.04)	.999 (.999)	.993	.02	.999
100	.982 (.979)	.04 (.04)	.998 (.998)	.982 (.978)	.04 (.04)	.998 (.998)	.985	.04	.999
50	.976 (.968)	.05 (.05)	.997 (.997)	.976 (.966)	.05 (.05)	.997 (.997)	.967	.05	.997
25	.963 (.950)	.06 (.07)	.996 (.995)	.964 (.946)	.06 (.07)	.997 (.995)	.936	.07	.994
10	.937 (.884)	.07 (.10)	.994 (.990)	.937 (.873)	.07 (.10)	.994 (.989)	.856	.12	.985
5	.891 (.820)	.11 (.13)	.988 (.982)	.886 (.812)	.10 (.13)	.989 (.982)	.765	.17	.970

Note. Results obtained from the original version are reported in parentheses. Avg. *SE* = average standard error; Rel. = reliability.

Table H.2

Correlations of the IRT Version and the Original Version With the American English CDI–WS Across Different Test Lengths and Age Groups

Length	16–18	19–21	22–24	25–27	28–30
680	.97 (.97)	.99 (.99)	1.00 (1.00)	1.00 (1.00)	.98 (.98)
400	.98 (.96)	.99 (.99)	1.00 (1.00)	.99 (1.00)	.98 (.98)
200	.99 (.96)	.99 (.99)	.99 (.99)	.99 (.99)	.98 (.98)
100	.98 (.95)	.99 (.98)	.99 (.99)	.98 (.99)	.98 (.98)
50	.98 (.94)	.99 (.97)	.98 (.98)	.97 (.98)	.97 (.96)
25	.96 (.92)	.98 (.95)	.97 (.96)	.96 (.96)	.95 (.94)
10	.92 (.81)	.95 (.87)	.95 (.90)	.94 (.90)	.92 (.89)
5	.87 (.74)	.92 (.82)	.92 (.84)	.89 (.85)	.84 (.82)

Note. Results obtained from the original version are reported in parentheses. Age groups are reported in months.

Table H.3

Comparisons Between the IRT Version and the Original Version Across Both Sexes and Different Test Lengths on the Danish CDI–WS, With Random Lists as Baseline

Length	Females			Males			Baseline		
	r with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.
725	.982 (.982)	.03 (.04)	.999 (.998)	.983 (.983)	.03 (.04)	.999 (.998)	1.000	.00	1.000
400	.985 (.980)	.03 (.04)	.999 (.998)	.987 (.981)	.03 (.04)	.999 (.998)	.997	.01	1.000
200	.985 (.977)	.03 (.04)	.999 (.998)	.985 (.979)	.04 (.04)	.998 (.998)	.990	.02	.999
100	.981 (.969)	.04 (.05)	.998 (.998)	.981 (.971)	.04 (.05)	.998 (.998)	.978	.03	.999
50	.974 (.957)	.04 (.06)	.998 (.997)	.974 (.956)	.05 (.05)	.998 (.997)	.955	.05	.997
25	.964 (.931)	.05 (.07)	.997 (.995)	.961 (.932)	.05 (.07)	.997 (.995)	.913	.07	.995
10	.924 (.863)	.06 (.09)	.996 (.991)	.939 (.870)	.06 (.09)	.995 (.991)	.807	.12	.986
5	.866 (.792)	.10 (.12)	.989 (.985)	.888 (.801)	.10 (.11)	.989 (.986)	.702	.16	.971

Note. Results obtained from the original version are reported in parentheses. Avg. *SE* = average standard error; Rel. = reliability.

Table H.4

Comparisons Between the IRT Version and the Original Version Across Both Sexes and Different Test Lengths on the Beijing Mandarin CDI–WS, With Random Lists as Baseline

Length	Females			Males			Baseline		
	r with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.
799	.976 (.976)	.05 (.06)	.997 (.994)	.974 (.974)	.04 (.05)	.997 (.997)	1.000	.00	1.000
400	.981 (.975)	.05 (.06)	.997 (.994)	.979 (.973)	.05 (.05)	.997 (.997)	.997	.01	1.000
200	.980 (.971)	.05 (.07)	.997 (.994)	.978 (.970)	.06 (.06)	.996 (.996)	.993	.02	1.000
100	.969 (.964)	.06 (.07)	.996 (.994)	.974 (.968)	.06 (.06)	.996 (.996)	.983	.03	.999
50	.957 (.950)	.06 (.07)	.995 (.993)	.967 (.959)	.07 (.07)	.995 (.995)	.965	.05	.998
25	.942 (.930)	.07 (.08)	.995 (.991)	.955 (.947)	.07 (.07)	.994 (.994)	.932	.07	.995
10	.916 (.871)	.08 (.11)	.994 (.987)	.930 (.902)	.09 (.09)	.992 (.991)	.852	.11	.987
5	.873 (.790)	.10 (.13)	.990 (.979)	.893 (.826)	.10 (.13)	.989 (.983)	.754	.16	.974

Note. Results obtained from the original version are reported in parentheses. Avg. *SE* = average standard error; Rel. = reliability.

Table H.5

Comparisons Between the IRT Version and the Original Version Across Both Sexes and Different Test Lengths on the Italian CDI–WS, With Random Lists as Baseline

Length	Females			Males			Baseline		
	r with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.	*r* with full CDI	Avg. *SE*	Rel.
670	.993 (.993)	.02 (.03)	.999 (.998)	.996 (.996)	.03 (.03)	.997 (.999)	1.000	.00	1.000
400	.992 (.992)	.03 (.04)	.999 (.998)	.995 (.994)	.04 (.03)	.997 (.999)	.998	.01	1.000
200	.987 (.989)	.04 (.04)	.998 (.998)	.990 (.990)	.05 (.04)	.996 (.998)	.992	.02	.999
100	.976 (.983)	.05 (.05)	.997 (.997)	.981 (.981)	.06 (.05)	.996 (.997)	.983	.04	.999
50	.965 (.970)	.06 (.06)	.995 (.996)	.971 (.962)	.07 (.06)	.994 (.996)	.964	.05	.997
25	.954 (.950)	.07 (.08)	.994 (.994)	.960 (.939)	.08 (.08)	.993 (.993)	.929	.08	.994
10	.943 (.877)	.08 (.11)	.993 (.987)	.931 (.862)	.08 (.11)	.992 (.986)	.840	.12	.984
5	.912 (.797)	.10 (.15)	.990 (.976)	.895 (.765)	.10 (.16)	.988 (.973)	.740	.18	.967

Note. Results obtained from the original version with flexible polynomial fitting are reported in parentheses. Avg. *SE* = average standard error; Rel. = reliability.