ManyBabies 3: A multi-lab study of infant algebraic rule learning

Authors

Ingmar Visser * [a] i.visser@uva.nl Andreea Geambasu * [b] a.geambasu@hum.leidenuniv.nl

Christina Bergmann [c] chbergma@gmail.com Krista Byers-Heinlein [d] k.byers@concordia.ca

Frances L. Doyle [e] f.doyle@westernsydney.edu.au Erin Hannon [f] erin.hannon@unlv.edu

Scott Johnson [g] scott.johnson@ucla.edu George Kachergis [h] kachergis@stanford.edu

Jessica Kosie [i] jkosie@princeton.edu Casey Lew-Williams [i] caseylw@princeton.edu

Julien Mayor [j] julien.mayor@psykologi.uio.no Jutta L. Mueller [l] jutta.mueller@univie.ac.at

Maartje Raijmakers [a] m.e.j.raijmakers@uva.nl Mohinish Shukla [n] mohinish.s@gmail.com

Angeline Sin Mei Tsui [h] astsui@stanford.edu

Melanie Soderstrom [# p] M_Soderstrom@umanitoba.ca

Claartje Levelt [# b] c.c.levelt@hum.leidenuniv.nl

Note: *Joint first authorship, [#]Joint senior authorship

Corresponding author: Ingmar Visser, i.visser@uva.nl

Affiliations

[a] University of Amsterdam [b] Leiden University [c] Max Planck Institute for Psycholinguistics

[d] Concordia University [e] Western Sydney University [f] UNLV [g] UCLA [h] Stanford University

[i] Princeton University [j] University of Oslo [l] University of Vienna [n] Unaffiliated, USA

[p] University of Manitoba

Data sharing statement: Data and code will be made available upon publication on Github.

Highlighted text indicates details that are to be filled out after data collection.

Acknowledgements: This research has been initiated within the ManyBabies research framework and has greatly benefited from the experience gained within earlier projects therein.

Research Highlights

- Our study provides a large-scale (XXX labs; YYY participants), cross-linguistic, conceptual replication of the seminal study of Marcus et al. (1999).
- This study will provide insights into the development of algebraic rule learning for infants aged 5;0 to 12;0 months
- Our study will evaluate whether direct repetitions have a special status in algebraic rule learning.
- A phonologically balanced and universally interpretable set of stimuli was developed to assess the robustness of algebraic rule learning in infants from multiple language backgrounds.

Project timetable

Submit registered report	July 2021
Revise/resubmit registered report	Fall 2021
Start data collection	Start of 2022, or after in principle acceptance
End data collection	End of 2022, data collection period will last a full year
Submit stage 2 report	June 2023

Abstract

The ability to learn and apply rules lies at the heart of cognition. In a seminal study, Marcus, Vijayan, Rao, and Vishton (1999) reported that seven-month-old infants learned abstract rules over syllable sequences and were able to generalize those rules to novel syllable sequences. Dozens of studies have since extended on that research using different rules, modalities, stimuli, participants (human adults and non-human animals) and experimental procedures. Yet questions remain about the robustness of Marcus et al.'s (1999) core findings, as the presence of significant learning effects has been mixed. In the current study, we aimed to address this issue by testing XX infants of a wide age range (5;0-12;0 months) in a multi-laboratory (XX laboratories) replication of the Marcus et al. (1999) study. SUMMARY OF KEY FINDINGS.
Keywords: rule learning, infants, multi-laboratory replication, language, cognition.

Introduction

The ability to learn and apply rules lies at the heart of cognition. Rules are essential in cognitive abilities, in areas as diverse as problem solving (Anderson, 1996), social cognition (Kunkel, 1997), cognitive development (Siegler, 1983), and, developing causal reasoning (Kuhn, 2012); the notion of rules and rule learning abilities are core concepts in building domain-general cognitive architectures (Anderson, 1996; Laird, 2012). Language is one area of cognition that has received a lot of attention with respect to the role of rules and rule learning. Acquiring a human language requires the ability to discover the meaningful units in the surrounding linguistic input (words), as well as the ability to discover and generalize rules combining these meaningful units at multiple levels (e.g., syntax, phonology, semantics). Crucial for both tasks, is keeping track of statistical patterns in the input. Saffran et al. (1996) have shown that eight-month-old infants can make use of a statistical learning mechanism to discover units in continuous speech. Others, however, have argued that tracking statistics is not sufficient on its own to explain rule-governed behaviours in language (and more broadly across cognition).

In their seminal study, Marcus et al. (1999) addressed whether infants could induce and generalize rules beyond the original stimulus material presented to them, and independent of surface features; that is, do infants employ a rule learning mechanism that can learn abstract relations between categories. Marcus et al. described this as “algebraic” rule learning (henceforth simply “rule learning”), because the learner must make use of open-ended abstract relationships or variables that can be expressed with arbitrary items. In their study seven-month-old infants were shown to be able to discover abstract ABB or ABA patterns in a two-minute stream of speech syllable triplets, and to generalize these patterns to sequences of novel syllables. The

authors conclude that in addition to a tool for tracking transitional probabilities (statistical learning), infants also possess a tool of manipulating variables (rule learning) to "learn about the world and attacking the problem of learning language." Many studies have since followed this paradigm to answer questions about human- or domain-specificity rule learning, and about the specific nature of the rule learning mechanism, albeit with mixed results (for a recent meta-analysis see Rabagliati, Ferguson & Lew-Williams, 2019). Thus, it remains unclear whether the rule learning effect is robust in infancy, and whether there are important causes of heterogeneity that may be driving inconsistent results in the extant literature. The primary aim of the ManyBabies 3 (MB3) project is therefore to establish the robustness of the rule learning effect in a consensus paradigm, by studying it in a large and diverse sample of infants across many laboratories. In our study we will focus on replicating Experiment 2 of the original study, which tests ABA versus ABB patterns - the motivation for this choice is discussed below. The secondary aim is to uncover potential causes of heterogeneity in the rule learning effect. Below, we first briefly describe the classical result by Marcus et al. (1999). Next, we i) discuss the relevance of rule learning per se, ii) address the motivation for studying rule learning with a large and diverse multi-laboratory approach, and iii) argue the usefulness of doing this within ManyBabies.

Rule learning

In Marcus et al. (1999), rule learning was examined in three experiments. In the first experiment, seven-month-old infants were familiarized for two minutes with 16 different three-syllable sequences that followed either an ABA or ABB rule. They were then tested on

both ABA and ABB sequences, one of the patterns thus being inconsistent with the pattern in the training stimuli. All test sequences were composed of novel syllables. When tested in the headturn preference procedure, 15 out of 16 infants showed a novelty preference, and, on average, infants listened to the inconsistent sequences 40% longer than to consistent sequences (Hedges' g = 1.13; this effect size and subsequently reported effect sizes from Marcus et al. were retrieved from http://metalab.stanford.edu/dataset/rulelearning/). In their second experiment, Marcus and colleagues further controlled for a possible correlation between the patterns of stimulus sequences and the phonetic features of the syllables (e.g., ABA patterns in the first experiment turned out to have voiced-unvoiced-voiced consonants in the syllable sequences). The authors again found that 15 out of 16 infants showed a novelty preference for the inconsistent items in the test phase and the overall effect size remained large (Hedges' g = 0.76). Finally, Marcus and colleagues examined the possibility that infants' successful discrimination of the patterns arose merely from one pattern – ABB – containing a direct repetition of syllables whereas the other – ABA – did not. If this possibility were true, then infants could have just focused on BB versus BA in order to discriminate the patterns, obviating the necessity to learn the algebraic rule. In the third experiment, the authors therefore presented a direct repetition in both conditions, by familiarizing infants with either AAB or ABB sequences and testing them using both AAB and ABB sequences. Again, it was found that infants (16 out of 16) listened longer to the inconsistent patterns in comparison to the consistent patterns with a similar effect size as reported in Experiments 1 and 2 (Hedges' g = 1.19). These findings suggested that infants focused on more than just a direct repetition of syllables. In summary, across the three experiments, Marcus and colleagues reported that seven-month-old infants could extract and

generalize simple abstract rules from syllable sequences, and that this ability did not appear to be driven by the phonetic make-up of the sequences or the mere presence of a direct repetition in the syllable sequences.

Multi-lab replication of rule learning

The classical result by Marcus et al. (1999) has sparked dozens of follow-up studies and their paper has been cited over 1000 times. Further, it is frequently mentioned in textbooks on language acquisition that seven-month-old infants have rule learning abilities (Gleason & Ratner, 2016; Hoff, 2014; Kail & Fayol, 2015; Lust, 2006; O'Grady, 2005; Saxton, 2010; Sun, 2008). Thus, it appears that rule learning is a cornerstone finding in the development of infant cognition. In this section, we will outline the rationale for further investigating rule learning with a large and diverse multi-laboratory study.

First, the centrality of rule learning and rule representation in cognition can hardly be overstated. The ability to represent rules lies at the very core of cognitive processing as conceived in the classical view of cognition; manipulating cognitive representations by (symbolic) rules is believed to be required to explain the powerful cognitive abilities that humans have in several areas, such as playing chess, solving reasoning problems, and doing mathematics (Chomsky, 1980; Fodor, 1981; Fodor & Pylyshyn, 1988; Pylyshyn, 1984). The ability to flexibly switch between rules has also been studied extensively as part of cognitive executive functioning in both adults (where the Wisconsin Card-Sorting Task is popular in the examination of switching; Kopp et al. 2019) and during development (where the Dimensional Change Card Sort

is frequently used; see Zelazo, 2015 for a review and Doebel & Zelazo, 2015 for a meta-analysis). Finding out if, when, and how rule learning develops informs fundamental debates about the architecture and development of the human mind (Marcus, 2001; Pothos, 2005).

Second, one of the unique cognitive abilities of human beings is language, characterized by an intricate system of rules. Rule representations and rule learning have not only been studied in human adults and infants, but also across a variety of different species to find out whether the rule-representing ability underlying language is indeed uniquely human or whether it is shared by other species. The Marcus et al. (1999) paradigm has been used in a number of studies with non-human animals (among others: Hauser & Glynn, 2009; de la Moro & Toro 2013; Chen et al., 2015), and, to date, there is only sparse evidence that non-humans can learn algebraic rules (van Heijningen et al., 2013; Ten Cate & Okanoya 2012; Santolin et al., 2016; Spierings & ten Cate, 2016). In order to make fair and insightful comparisons between human and non-human learning abilities it is important to assess the robustness of the original finding.

Third, in order to test whether the ability of rule extraction and generalization is domain-general, subsequent studies have extended the rule learning paradigm to different modalities. Overall, these findings suggest that infants are capable of extracting and generalizing simple rules from visual stimuli (e.g., Johnson et al., 2009, Ferguson et al., 2018), from non-linguistic sound stimuli such as tones or animal sounds (Marcus et al., 2007), and from bimodal visual-auditory stimuli (e.g., Frank et al., 2009; Tsui et al., 2015). In a meta-analysis, Rabagliati et al. (2019) reported an average positive - indicating a preference for items inconsistent with the training items - and significant effect size of Hedges' *g* equal to 0.25 for

infants' rule learning ability across 91 studies, suggesting that infants demonstrate a rule learning ability in a variety of modalities. However, this average effect size was much smaller than those reported in Marcus et al. (1999), which had Hedges' *g*'s = 1.13, .76, and 1.196 respectively for their three experiments. Importantly, Rabagliati et al.'s meta-analysis showed that there is large heterogeneity in looking time preferences across studies, ranging from large novelty preferences to large familiarity preferences. Further, Rabagliati et al. (2019) found that meaningful and speech-like stimuli yielded the largest effect of infant rule learning studies, but they also had a very high residual heterogeneity in the meta-regression models, even when other variables were controlled for (e.g., infants' age, experimental design factors). Thus, how and why infants' rule learning varied across studies remains largely unknown.

Fourth, studies examining infant rule learning have mainly come from well-resourced laboratories in Western societies (Rabagliati et al., 2019). Although there are a few notable exceptions, (i.e. a few studies in Asia, such as Tsui et al., 2016; Tseng et al., 2018), this raises concerns about whether infant rule learning (as described in the Marcus et al., 1999 paradigm) is a universal ability that can be generalized to different infant populations across the diversity of human infant experiences.

In summary, it appears that our understanding of infant rule learning remains very limited. To address this research gap, we aim to carry out a large-scale replication study to examine infant rule learning across a wider age range (i.e. 5;0 to 12;0 months) and with populations from many different language backgrounds, in order to identify factors that explain the large variations of effect sizes in the literature and to improve generalizability of the findings.

Rule learning in ManyBabies

The current ManyBabies study builds on existing findings in other ManyBabies studies. The first ManyBabies project – MB1 – investigated infants' preference for infant-directed speech (IDS) over adult-directed speech (ADS), and has generated a number of insights regarding the influence of so-called "laboratory factors", participant characteristics, and methodological approaches on the strength of infant looking time effect sizes (ManyBabies Consortium, 2020) . However, as the first study of its kind, it is not clear to what extent these findings would generalize to other research questions and procedural details. For example, in MB1, the strongest effect size was found using the "Headturn Preference Procedure" (HPP) where infants must make a clear head turn toward the stimulus, compared to approaches that required only eye movement responses, central fixation and eye-tracking. However, there are a number of possible explanations for the higher effect sizes in HPP, some of which could generalize to other studies (e.g., that HPP involves a more active response from the infant and is therefore more sensitive), while others would not (e.g., a correlation between the particular skill set of the specific laboratories using HPP and the question being tested – laboratories with more resources or more experience with infant language studies may use HPP more than laboratories with fewer resources or less related experience). The current ManyBabies study allows us to continue to test the robustness and generalizability of that finding.

In the present ManyBabies study, we can also build on the findings of MB1 in three meaningful ways. First, the rule learning phenomenon being tested in the present study is structurally simpler than the phenomenon tested in MB1. In MB1, IDS and ADS stimuli differed

along a variety of dimensions related to the IDS/ADS distinction (e.g., pitch range, vowel durations, pauses), and along other dimensions (e.g., affect); thus, the underlying ability being examined in the present study is more clearly defined than in MB1. Specifically, in the current study, generalization across syllable patterns will be assessed, with stimuli carefully constructed to control for several acoustic-phonetic properties (see Stimuli section below for details). Second, rule learning is expected to be more universal in its expression than the preference for IDS, which is thought to vary based on infants' own linguistic experiences (ManyBabies consortium, 2020; Byers-Heinlein et al., 2020). By examining a (theoretically) universal phenomenon, we will be able to better tease apart methodological factors from population-specific factors (which may be heavily confounded across laboratories). Third, dissimilar to MB1, the stimuli in the current study are designed to be comparable in their naturalness cross-culturally. While in MB1 all infants regardless of their background were tested on stimuli clearly created in North American English, we designed stimuli for the current study to sound reasonably "natural" to infants across a wide variety of environments, countries/cultures, and language experiences (see Stimuli Section below for further details).

Current Study Design

The original motivation for Marcus et al. (1999) to examine rule learning was to understand the computational bases of language constructions that appear to operate over variables such as "verb phrase". To this end, the stimuli were language-like and consisted of "words" (speech syllables, created by a Bell-Labs synthesizer) arranged in "sentences" (trisyllabic sequences). As in the original study, we too employ

language-like stimuli. Past studies have revealed that 7-month-olds were better able to extract rules from sequences of both speech and nonspeech sequences (e.g., musical tones, animal sounds, or varying timbres) during the test phase if they had been familiarized with those rules in sequences of speech (Marcus et al., 2007). Further, the meta-analysis of Rabagliati et al. (2019) showed that speech-like stimuli yielded the largest effect of infant rule learning. Another key reason for using language-like stimuli is that much of the rule learning literature informs debates in language learning, and hence using speech stimuli is important when adding to these debates. We did, however, decide to deviate from the original stimuli in several ways, explained in detail below in Stimuli section.

Studies of infant rule learning have produced mixed results with respect to the role of adjacency in the learnability of a simple abstract repetition rule (Schonberg et al., 2018). In experiments with abstract visual stimuli, direct repetitions seemed to facilitate rule learning such that some rules (e.g., ABB) were easier to learn than others (e.g., ABA; Johnson et al., 2009). Age and the position of the repetition both played a role here too: in Johnson et al. (2009), 8-month-olds could only learn from ABB visual sequences, while 11-month-olds could also learn from AAB sequences. Neither age group could learn from ABA sequences. However, with pictures of dogs and cats, 7-month-olds could learn from both ABA and ABB sequences (Saffran et al. 2007). In auditory rule learning studies, there were mixed results as well. In Marcus et al. (1999), 7-month-olds familiarized with ABA sequences of syllables generalized to new syllables when contrasted with ABB, and vice-versa; that is, infants who learned ABA appeared

to recognize this pattern in new items, and identified ABB as a novel pattern. This was also found in Gerken (2006) with 9-month-olds, in Marcus et al. (2007) with 7.5 month-olds, and in Thiessen (2012) with infants between 6.5 and 8.0 months old. However, newborns only responded to the difference between ABB and ABC sequences, not to the difference between ABA and ABC sequences (Gervain et al., 2008). Moreover, in a complex hierarchical task where the As and Bs in ABA and AAB patterns consisted of further ABA or AAB patterns, 7-month-olds could only learn AAB rules, not ABA rules (Kovács, 2014). When 12-month-olds were tested on learning both ABA and AAB rules simultaneously, monolinguals could only learn AAB, not ABA, but bilinguals could learn both (Kovács & Mehler, 2009b). There thus appears to be a facilitating effect of direct repetition in previous rule learning studies. In addition to this, it appears that humans, and some other animal species too, are sensitive to repetition and it has been suggested that there is a specialized primitive devoted to processing identity relations (Endress, Nespor, & Mehler, 2009). While both ABB and ABA are examples of repetition patterns, this primitive appears to be especially sensitive to the processing of direct repetitions as in **AA**B or A**BB**. In order to be able to find out more about the role of direct versus non-adjacent repetition in rule learning and given above mentioned previous findings in rule learning, we decided to focus on replicating Experiment 2 of Marcus et al. (1999), where ABA vs ABB is tested, across age and mono-/multilingualism using a relatively easy comparison of strings that can be distinguished by the presence of either direct or non-adjacent repetition.

Following MB1, laboratories will employ either the headturn preference procedure (HPP), or single screen setups with the central fixation paradigm (CF) or eye-tracking (ET) to study the magnitude of the learning effect in each of these paradigms. This will ensure that we can maximize the number of laboratories able to contribute data in the present study. Within these paradigms, we will minimize procedural flexibility wherever possible, while recognizing that laboratories will vary in what is technically possible to implement. See the Procedure Section for discussion and details[1].

The large majority of rule learning studies have focused on testing seven-month-old infants, although there are some exceptions (Dawson & Gerken, 2009; Frank et al. 2009; Gerken, 2006; Johnson et al. 2009; Schonberg et al., 2018). As a result, it has been difficult to study the impact of age on the learning effect, although the meta-analysis by Rabagliati et al. (2019) investigated this factor and did not find an age effect. In this study, we test infants between 5;0 and 12;0 months of age. Thus, the present study will include the age ranges that have been tested in the majority of studies about rule learning that are published to date. By studying a wide age range, it becomes possible to investigate the developmental trajectory of rule learning. Our findings in this respect will also help inform a larger question in developmental research, namely whether rule learning develops with age. Interestingly, MB1 found an increase in effect size in the preference for IDS over development. This increase could be due to a developmental change specifically in infants' preference for IDS, but could also be due

[1] Detailed instructions for participating labs can be found in the ManyBabies 3 Lab Manual; note that this is not complete yet.

to an age effect that is more widespread across experimental contexts (Bergmann et al., 2018; ManyBabies Consortium, 2020). In the latter case, the present study may be able to capture a more general effect of more robust behavioural responding.

As there are several important reasons for deviating from the original Marcus et al. paradigm, this ManyBabies study should be viewed as a conceptual, rather than a direct, replication. We are aware of a study being conducted concurrently in The Netherlands (with overlap in authorship with the current study), that more directly replicates Experiment 2 of Marcus et al. (1999) with the original stimuli, the headturn preference paradigm, tested in 7-month olds. The present study is a complementary initiative, and will provide a highly informative comparison with the Netherlands' team's replication of Experiment 2. Further, the present study allows us to examine the robustness of rule learning findings and probe the dimensions of its generalisability.

Research Questions and Hypotheses

The rule learning "effect" is defined as infants' ability to generalize an abstract pattern from one set of stimuli to another, operationalized as a post-familiarization preference (via headturn or visual fixation) for a novel pattern (Marcus et al., 1999; Rabagliati et al., 2018).

Our primary research questions are the following:

1. What is the magnitude and the variability of the rule learning effect in infancy?
2. Is the rule learning effect modulated by age, linguistic background, or experimental paradigm (e.g., HPP vs. CF)?
3. Is rule learning facilitated by direct repetition?

Question 1: What is the magnitude and the variability of the rule learning effect in infancy?

In MB3 (as in MB1), we employ a consistent set of stimuli across labs to reduce the likelihood that stimulus parameters will account for any observed differences in the effect across sites. These stimuli were designed to be appropriate for use across languages with diverse phonologies. In a previous meta-analysis, large heterogeneity of effects was found across labs, which in part may have been the result of differing stimulus materials (Rabagliati et al., 2019). This possible confound is eliminated here by standardizing the experiment materials. Eventual heterogeneity beyond what can be expected due to sampling error in our findings could hence be attributed to either of two main factors: 1) differences between labs that are not captured in the specification of the experiment's procedures, or 2) heterogeneity in the rule learning effect itself, i.e., it being a highly variable effect within and between infants resulting in high sample-to-sample differences in its magnitude. If indeed heterogeneity is limited after controlling for testing paradigm, linguistic background, and age, this would imply that rule learning is a robust phenomenon, and perhaps universal in its development in infancy. Finding relatively greater heterogeneity, in contrast, would imply that rule learning is not a robust ability, or that our ability to measure it is influenced by factors outside those we aimed to control. A third possibility is that the magnitude of the effect varies in a systematic fashion according to testing paradigm, language background, or age, as described below. Such a finding could eventually shed light on the nature of the mechanisms underlying rule learning. Related to this first question the research hypotheses are:

Hypothesis 1a: Rule learning is evidenced by a novelty preference for inconsistent patterns at test.

Hypothesis 1b: After controlling for age, linguistic background, and experimental paradigm, there is little residual heterogeneity between labs.

Importantly, beyond testing these hypotheses, it will be informative to have a robust estimate of the overall effect size as well as an estimate of residual heterogeneity should it be present.

Question 2: Is the rule learning effect modulated by age, linguistic background or experimental paradigm (HPP/CF/ET)?

First, there are not many studies that have extensively reported on the effect of age on the success of rule learning. Generally it would be reasonable to expect that older infants outperform younger infants as the ability to process information develops, in which case the overall effect will strengthen with age. In visual rule learning, the effect appeared to increase with age (Johnson et al., 2009; Schonberg et al., 2018). Comparatively, Dawson and Gerken (2009) found that rule learning weakened between 4 and 7 months when musical chords or tones were used as stimuli. Further, the meta-analysis of Rabagliati et al. (2019) revealed a slight, nonsignificant decline in performance with age. These mixed findings may be related to the use of stimuli from different domains, but this also calls for investigating the effect of age on rule learning in a large-scale replication study. In the current study, we expect that older infants will perform better than younger infants in the rule learning experiment, for two reasons: Rabagliati et al.'s (2019)

meta-analysis has suggested that meaningful stimuli improved infants' rule learning, thus we expect that our stimuli (language stimuli) will have the same positive effect, as language is meaningful to all infants; in addition, since MB1 found evidence of an increase in language related processing skills with age, we expect to find such an effect here too.

Second, we will investigate whether the rule learning effect varies by language background. The absence of differences by language would suggest that infant rule learning has universal characteristics. However, if the rule learning effect varies systematically by language background, this could be caused by non-universal variation in the cognitive processes underlying rule learning, which may be due to infants' different experiences during development that influence these cognitive processes. Alternatively, although we have explicitly tried to create stimuli that are equally interpretable across languages, if the rule learning effect varies across languages then there might be language-specific effects of our stimuli. In addition to variation based on linguistic differences, infants may perform differently based on the number of languages they are exposed to. It has been suggested that infants from multilingual households outperform infants from monolingual environments in rule learning (Kovács & Mehler, 2009a, 2009b) from enhanced experience and efficiency with "task switching," or shifting between dimensions and sets (Prior & MacWhinney, 2010). However, the literature about whether bi- or multilingual infants have an advantage in switching and abstracting rules is mixed. Some studies have found that bilingual infants are better at rule-switching in comparison to monolingual infants (Comishen, Bialystok & Adler, 2019; Kovács & Mehler, 2009a, 2009b; Pour lliaei, Killam, Dal Ben, & Byers-Heinlein, 2021), while other studies found little evidence for a bilingual advantage in switching or rule learning in infants (e.g., D'Souza, Brady, Haensel,

D'Souza, 2020; Kalashnikova et al., 2020, Tsui & Fennell, 2019). We therefore do not have a specific prediction regarding whether a bi-/multilingual rule learning advantage in infants younger than 12 months will be found, but because both monolingual and bi- or multilingual infants will participate, our study will provide information on this issue.

Third, we will ask whether the rule learning effect is modulated by the testing paradigm. In particular, we test the hypothesis that HPP offers a better or more reliable assessment of infant rule learning than methods using eye movements in response to a single screen (i.e., CF and ET) as the dependent variable. As noted previously, HPP yielded a larger effect size for infant-directed speech preference in the MB1 study (ManyBabies Consortium, 2020), though the reasons for the difference remain unclear. Theoretically, HPP may be a more sensitive test of learning and discrimination because it provides a clearer response behavior — a headturn to one of two locations, as opposed to a difference in look duration to a single, central target — hence yielding a larger response consistent with signal detection theory (Green & Swets, 1966). Alternatively, but relatedly, HPP is a more active task that may better allow infants to indicate preference, as opposed to looking towards a single central screen, where infants could be disinterested in a stimulus or task but continue to look forward. While this study on its own cannot rule out potential confounds such as idiosyncratic differences in laboratory experiences with methods or populations, it will provide important data regarding the robustness of the stronger effect in HPP found in MB1.

An important auxiliary hypothesis is that there will be habituation at test, and that this habituation occurs more strongly in older infants than in younger infants (Hunter & Ames,

1988). This results in decreasing looking times with subsequent test trials with larger decreases in older infants. In sum, hence, the hypotheses related to our second research question are:

Hypothesis 2a: Rule learning is robust across age and increases with age.

Hypothesis 2b: Looking times decrease with test trial; this effect will be stronger for older infants than younger infants.

Hypothesis 2c: The relationship between multilingualism and rule learning will be explored. Both absence or presence of this effect or its reverse can inform the debate about potential cognitive advantages related to multilingualism.

Hypothesis 2d: Rule learning is robust across variation in experimental paradigms and is facilitated in the Headturn Preference Procedure relative to other paradigms.

Question 3: Is rule learning facilitated by direct repetition?

From the literature discussed above it appears that under more difficult circumstances, i.e., when rules have to be learned from abstract visual stimuli (Johnson et al., 2009), from more complex stimuli (Kovács, 2014) or when two rules have to be learned simultaneously (Kovács & Mehler, 2009b), infants are not able to learn repetition patterns when they involve non-adjacent items (ABA). Furthermore, neonates only seem to be sensitive to syllable sequences with adjacent repetitions, that is AAB and ABB (Gervain et al., 2008). Finally, in a study by Geambasu (2018) 7-month-old infants were more interested in AAB test items than in ABA test items, independent of their training sequence (AAB or ABA). By studying the rule learning abilities of a large number of infants, in a wide age-range and with different linguistic backgrounds, we can seize the opportunity to analyze the effect of this factor. The hypotheses related to our third research question are:

Hypothesis 3a: There will be an overall preference, i.e. longer looking times, for adjacent repetition items (i.e. ABB items) in the test phase, regardless of training condition.

Hypothesis 3b: There will be a larger effect of learning overall in the participants trained on adjacent repetition items (i.e. ABB) then the participants trained on other items (i.e. ABA).

Methods

Participants

Participating laboratories are recruited via a call for participation through social media, relevant listservs, and by word of mouth. To confirm their intention to participate, laboratories will complete a registration form prior to data collection that includes information about laboratory characteristics, population characteristics, laboratory-specific protocol details and information relevant to preregistration (e.g., sample size commitment), and the research paradigm they will use. Laboratories are responsible for obtaining any necessary ethics approval via the ethics oversight committees at their home institutions. In locations where no such oversight exists, laboratories can obtain ethics approval either as a rider to an existing approval from another laboratory or submit a signed statement committing to conducting their research in accordance with established research ethical norms, including but not limited to the Declaration of Helsinki. Laboratories commit to a minimum sample of $n = 16$ participants with usable data, but data from laboratories that are unable to meet this minimum sample will nonetheless be

included. Data collection will start as soon as this manuscript receives in-principle-acceptance status, and will end one year later. Table 1 will provide details regarding the participating laboratories, with VVV laboratories using the HPP (of which XX used online and YY used offline coding), XXX laboratories using an eye-tracker (all automatic coding), ABC laboratories using CP (of which XX used online and YY used offline coding).

Table 1

Participating Laboratories

Laboratory	**Country**	***N***	**Testing method**	**Coding method**	**Mean (SD) infant age**
(dummy info)					

Note. N refers to the number of infants included in the final analysis. Testing method refers to HPP with lights, HPP with screens, vs. central fixation (CF). Coding method refers to automatic eye-tracking vs. offline hand-coding vs. online coding.

Table 2 will describe participant demographic information including linguistic background and information on mono-/multi-lingualism; note that laboratories can have multiple primary language groups.

Table 2

Demographic Information

Lab	***N* included**	***N* excluded**	**Mean Age in months (*SD*)**	**Age range in months**	**Gender (M/F)**	**Primary Language**	**Mean (SD) exposure (%) to primary language***
XYZ							
XYZ							
XYZ							
Totals							

Note. Age, gender, and language refer only to included participants. Some labs are listed more than once if they tested participants with different primary languages. *In the analyses, we will use a continuous variable indicating the proportion of non-primary language input that the infants get.

Participant inclusion and exclusion

Infant participant age will range from 5;0 months (152 days) to 12;0 months (365 days). There are no specific guidelines for individual laboratories regarding distribution of infants within this range, although laboratories will be encouraged to test across as wide an age range as possible. Inclusion criteria for participation are: full term (born at 37 or more weeks gestation), with no parent-reported developmental disorders or sensory impairments. Physical

developmental disorders that do not affect cognition or the infant's ability to respond to the stimuli will not count as exclusion criteria. Each lab will provide all data (in anonymized form) collected for the study to the analysis team. Participants' data will be excluded from analyses based on the following criteria, applied sequentially:

1. Age. Infants younger than 5;0 months or older than 12;0 months will be excluded from analyses.
2. Pre-term. Pre-term infants, here defined as fewer than 37 weeks gestation, will be excluded from analyses.
3. Diagnosed developmental disorders. Infants will be excluded based on parental reports of developmental disorders, vision, or hearing impairments.
4. Session-level errors. Session-level errors are errors that impact the entire session, and result in excluding a participant's data from analyses completely. Session level errors include:
    a. equipment error (e.g., an eye tracker failing to record data),
    b. experimenter error (e.g., an experimenter becoming unblinded when measuring infants' looking by live button press),
    c. evidence of parent/outside interference (e.g., parents talking or pointing at the screen, loud construction noise, siblings entering the room or pounding on the door).
5. Insufficient usable data, i.e., less than one pair of test trials. To be included in the study, an infant must have contributed non-zero looking times on at least one ABA trial and one

ABB trial, after trial level exclusions are applied (see Data Analytic Approach section for details). Hence exclusion criteria are:

a. fussiness causing the infant not to reach the test phase,

b. infant fussiness leading to absence of valid test trials,

c. equipment error (e.g., an eye tracker failing to record data).

The number of exclusions resulting from each of these criteria will be listed in Table 3. After sequential application of these criteria, the final sample size for analysis was XX infants (YX male and XX female).

Table 3

Exclusion Numbers

Criterion	Number excluded
1. Wrong age	
2. Pre-term	
3. Diagnosed developmental disorders	
4. Session level exclusions a. Equipment error b. Experimenter error c. Interference	
5. Insufficient useable data a. Fussiness causing the infant not to reach the test phase b. Fussiness leading to absence of valid test trials c. Trial-level equipment error	

Note. Number of infants excluded following the exclusion criteria in the order listed.

Stimuli

The stimuli in the experiment we are replicating, Experiment 2 of Marcus et al. (1999), consisted of synthesized speech syllables in a male voice. In the original article, the syllables used in the familiarization sequences were rendered as “le” “je” “we” “de” “li” “ji” “wi” and “di” and those in the test as “ba” “po” “ko” and “ga”. Listening to the original stimuli, provided to us by the original first author through Scott Johnson, these syllables can be provided with the International Phonetic Alphabet (IPA) transcripts in Table 4.

Table 4

International Phonetic Alphabet Transcripts

“le”	[li]	“wi”	[waɪ]
“je”	[dʒi]	“di”	[daɪ]
“we”	[wi]	“ba”	[bʌ]
“de”	[di]	“po”	[pʊ]
“li”	[li:]	“ko”	[kʊ]
“ji”	[dʒaɪ]	“ga”	[gʌ]

Note. IPA transcripts of the stimuli in Experiment 2 of Marcus et al. (1999)

The original syllables contain several English-specific sounds, like the affricate [dʒ] and the vowels [ʌ] and [aɪ], and the syllable “li” and “le” sound very similar. The syllables also have different lengths, varying between 216.8 ms and 415.2 ms, with varying intonation contours.

In the current study, we chose to deviate from the original stimuli in several ways, for a number of reasons. First, because our language stimuli will be presented to infant participants from many different language backgrounds, we opted for a set of less English-specific sounds that would be perceived as distinct from each other across a diverse set of languages. While the specific sounds were chosen to ensure correspondence with distinct phonemes within the expected set of languages across participating labs, they are not necessarily perceived in exactly the same way across languages. For example, we have only included stop consonants with a voice onset time (VOT) of ~0 ms, which is interpreted as unvoiced in languages like Dutch and French but as voiced in languages like English and German. However, the resulting syllables that make up the ABA and ABB patterns in the experimental sequences should be interpretable and perceived as different from each other by all participants.

Second, our syllables are based on natural speech, spoken by a male voice, and manipulated to be of similar length and with a flattened intonation. If the linguistic nature of stimuli would indeed be a facilitating factor in this type of rule learning task, then more natural-sounding stimuli should help the participants. However, we wanted to avoid confounds triggered by patterns in intonation contours or differences in syllable length. Hence we have recorded spoken syllable tokens that were subsequently manipulated in Praat (Boersma & Weenink, 2019) to create a set of individual syllables that are matched on various prosodic measures (see Syllable Preparation and Validation section below).

The final list of syllables is composed of specific combinations of the consonants /p/, /t/, /k/, /f/, /s/, /l/, /n/, /m/ and the vowels /a/, /o/, /u/, /e/, /i/, and individual syllables are expected to be neither odd nor unattested in any of the tested populations' languages.

Syllable preparation and validation

Stimuli were recorded in a quiet space using a Behringer Ultravoice Xm8500 Dynamic Vocal Cardioid Microphone with a pop shield, onto a Marantz solid state recorder in battery mode. Syllables were produced using as even a rate and as flat a prosody as possible. All subsequent handling and analyses were in Praat.

Individual syllables were manually segmented out, and clipped to nearest zero-crossings. All syllables were then concatenated with 250 ms silence between each. The entire concatenated sound file was pitch flattened by setting the pitch to a constant value of 132 Hz (the mean pitch across the recorded syllables). Individual syllables were then segmented out again and adjusted to have a duration of 250 ms (actual average 248 ms ± 6.4 ms), intensity of 70 dB (actual average 69.83 dB ± 0.3 dB), and mean pitch of 132 Hz (actual average 132.38 Hz ± 0.55 Hz) each. Pilot transcriptions by members of this author group led to some individual syllable tokens being re-recorded.

We then gathered syllable identification data for the final set of syllables using Amazon Mechanical Turk. Participants judged the identity of each syllable and provided a rating of the prototypicality/quality of the syllable they identified on a seven-point scale from 0 (Poor example) to 7 (Great example). Modal responses matched the actual syllable for all tokens, and mean prototypicality/quality ratings were all >4. The stimuli, further details about their creation and the full data of the syllable ratings can be found at the MB3-OSF repository.

Training and test sequences

The final selection of syllables is divided into training and test sets, ensuring that phonetic features (stops, nasals, open vowels, etc.) are as evenly distributed across training and test sets as possible. The syllables in the training set are constructed with the consonants /p/, /k/, /f/, /m/ and the vowels /a/, /u/, /e/, while the syllables in the test set are constructed with the consonants /t/, /s/, /l/, /n/ and the vowels /o/, /i/. Further, within each group, syllables are assigned to either A or to B classes to enable construction of the ABB/ABA sequences. The full syllable set is shown in Table 5 below.

Table 5

Syllable Set

Training	A	/ku/, /ke/, /mu/, /ma/, /fe/, /pa/
	B	/ka/, /me/, /fa/, /fu/, /pe/, /pu/
Test	A	/ti/, /lo/
	B	/so/, /ni/

Note. Syllables assigned to A and B classes in training and test.

Individual “sentences" are constructed by concatenating syllables from A and B classes within training- and test groups in all possible combinations, for both ABA and ABB structures, with the restriction that the A and B syllables within a given sequence do not share the vowel or

the consonant. This restriction prevents the possibility of presenting "XXX" patterns over three identical consonants (e.g., fefafa) or vowels (e.g., pafafa); hence, both the consonants and the vowels within the sequence highlight the ABB or ABA pattern (Thiessen, 2012). This contrasts with the original Marcus et al (2007) stimuli, in which 7 out of the 16 training sequences contained identical vowels ("le je le" [li dʒi li], "le we le" [li wi li], "wi di wi" [waɪ daɪ waɪ], "ji di ji" [dʒaɪ daɪ dʒaɪ], "de je de" [di dʒi di], "de li de" [di dʒi di] "de we de" [di wi di], 3 contained identical consonants ("wi we wi" [waɪ wi waɪ], "je ji je" [dʒaɪ dʒi dʒaɪ], "de di de" [di daɪ di], and 1 contained both identical consonants and vowels ("le li le" [li li: li]).

As in the original stimuli, each trisyllabic sequence has a 250 ms silence between the syllables. Training and test sequences are constructed by concatenating "sentences" of either the pattern ABB or the pattern ABA into pseudo-random sequences (see Procedure section for details) with a 1.2 s pause inserted between sentences.Table 6 shows the familiarization and test items for both the ABA and ABB patterns.

Table 6

Familiarization Items

Familiarization items ABA	Familiarization items ABB	Test items
fekafe	fekaka	lonilo
fepufe	fepupu	lonini
kefake	kefafa	tisoso
kefuke	kefufu	tisoti
kepuke	kepupu	
kufaku	kufafa	
kumeku	kumeme	
kupeku	kupepe	
mafuma	mafufu	
mapema	mapepe	
mapuma	mapupu	
mufamu	mufafa	
mukamu	mukaka	
mupemu	mupepe	
pafupa	pafufu	
pamepa	pameme	

Note. Familiarization and test items following both the ABB and ABA pattern.

Procedure

The experimental procedure will consist of a familiarization phase followed by a test phase. During the familiarization phase, infants will be exposed to a speech stream following either an ABA or an ABB pattern. Either stream is composed of 16 unique sequences which will each be presented three times, randomised within blocks of 16 sequences. The streams will thus have a duration of approximately two minutes (116.3 s).

The test phase will consist of three blocks of two different ABA and two different ABB sequences, for a total of 12 test trials, composed of syllables that are not presented in the familiarization phase. Test items are therefore considered either consistent (familiar) or inconsistent (novel) with the pattern presented in the familiarization stream. Per test trial, each sequence will be repeated for a maximum of six times, with a 1.2 s pause between repetitions.

Each laboratory will be asked to test approximately equal numbers of participants in each of four counterbalancing conditions, resulting from the ABA/ABB familiarisation pattern (2) by familiar versus novel first test trial (2).

Headturn Preference Procedure (HPP)

In the HPP testing paradigm, infants will be positioned on a caregiver's lap, or near the caregiver in an infant seat, in a quiet room, with the caregiver wearing headphones playing masking auditory stimuli. The HPP setup will include a central fixation point and one or two fixation points on the left and/or on the right side of the infant. The fixation points will be either

blinking lights or screens, depending on each laboratory's hardware. In the original study, lights were used. Speakers will be located in close proximity to the side fixation point(s).

The familiarization phase will be initiated with an attention grabber - visual only, no sounds - played from the central fixation point. Once the infant fixates on the central attention grabber, the familiarization stream will begin to play, continuously and independently of the infant's attention.

After the familiarization phase, an attention grabber will be played to refixate the infant's attention to a neutral central position. Once the infant fixates to the center, the testing phase will start.

Each test trial will begin with the initiation of one of the visual stimuli on either side of the infant. Once the infant turns his or her head to the visual stimulus, the accompanying auditory stimulus will begin to play. Each test trial will play for as long as the infant maintains attention towards the visual stimulus, or until the maximum of 15 s per trial is reached. If the infant looks away for less than 2 s, both the auditory and the visual stimuli will continue to play. If the infant looks away for more than 2 s, both the auditory and the visual stimulus will end and thereby the trial will end. Once the test trial has ended, the attention grabber will be played again until the infant fixates on the neutral central point, after which a new test trial will start. This proceeds until all 12 test trials will have been played.

Central Fixation (CF) and Eye Tracking (ET)

The CF and ET paradigms will proceed in a similar manner to the HPP. The key difference is that in these paradigms, a single screen will be used instead of multiple lights or monitors. In these

setups, infants will also be positioned on a caregiver's lap or in an infant seat in a quiet room, while parents wear headphones playing masking auditory stimuli. In addition, in the eye-tracking set-up the look-coding will typically be done using an eye-tracking software program.

During the familiarization phase, one of the familiarization streams, following either the ABA or ABB pattern, will be played uninterrupted, while the screen will show a static checkerboard pattern over its entirety. This will proceed independent of whether the infant fixates on the visual stimulus or not. Once the familiarization stream will be finished, a dynamic attention grabber will be played to refixate the infant's gaze. The attention grabber will be played until the experimenter or the eye-tracker software has determined that the infant has looked for 500 ms. If the maximum duration (15 s) is reached without a fixation, the attention grabber will be restarted. Once the infant fixates on the attention grabber, the first test trial will be initiated.

For test trials, as in the familiarization phase, the static checkerboard pattern will be shown on the entirety of the screen, while the auditory stimulus is playing. If the infant looks away from the screen for less than 2 s, the auditory and visual stimuli will continue to play. If the infant looks away from the screen for more than 2 s, both the auditory and the visual stimuli will be stopped and the trial will end. Once the test trial has ended, the attention grabber stimulus will play again until the infant fixates on it, after which a new test trial will begin. This will proceed until all 12 test trials are played.

Pilot

A pilot study was conducted by one laboratory in March, 2021 to verify the feasibility of the experiment's design, to test the planned procedures, and to test-run the planned analysis.

Given the small size of the pilot, no significant effects were observed, but nonetheless the planned analysis was carried out; the full pilot report can be found at the MB3-OSF repository.

The pilot data gathering and the pilot data results resulted in one minor change in the procedure where the length of central fixation for eye-tracking set-ups was lowered to 500 ms - that is time prior to the start of the trial that the infant fixates the central stimulus.

Power analysis

In March 2021, a call for participation of laboratories was posted on several mailing lists for developmental and infant researcher communities. As of June 2021, 25 laboratories expressed their interest in contributing with an average of over 20 infants planned to be tested. To be somewhat conservative, our power analysis assumed that a minimum of 20 laboratories would recruit 20 participants (after exclusions), for a total of 400 infants (*age range*: 5-12 months of age). As traditional power analysis is not possible for mixed-effects regressions, we instead used a simulation approach, in which we repeatedly generated random datasets (N = 20 laboratories * 20 children * 12 trials = 4800), specifying a chosen effect size (e.g., 0.3 SD, the average effect size across all published developmental experiments; Rabagliati et al. 2019) for any main/interaction effects of interest. For simplicity, we generated datasets using the same specified effect size for trial type, age, and their interaction, and used a slightly simpler regression model than will be used in the full analysis: log(looking time) ~ 1 + trial_type * trial_number + trial_type * age + (1|subject) + (1|lab). Power was evaluated by examining in how many of 1,000 such generated datasets the mixed-effects regression estimated a significantly non-zero ($p < .05$) coefficient. We conducted three such power simulations, assuming effect sizes

(Cohen's *d*) of 0.1, 0.2, or 0.3[2] for the three main effects of interest (trial type, age, and trial type * age). For effects of size *d* = 0.1, 92.8% (928 of the 1,000) of the simulations recovered a significant trial type coefficient, and 100% of simulations showed a significant age * trial type interaction. For effects of size 0.2 or 0.3, 100% of simulations showed significant trial type effects, as well as age * trial type interactions. Thus, our power analysis suggests ample power to detect the expected trial type effect and expected interactions with age, even if the effect size is quite small (0.1-0.2). A more elaborate description of the power analysis including the R-code used can be found at the MB3-OSF repository.

Data analytic approach

Laboratories will provide anonymized data for all infants to the analysis team. Exclusions for data analysis on the infant level are detailed in the Participant section. Contributing laboratories will be asked to include first-session infants only; second-session infants (i.e. infants run in a different study on the same visit prior to their participation in MB3) will be included in the data, but labeled as such and the effect of this inclusion was studied in the robustness analyses (see Robustness Analyses section below).

Data from individual trials will be discarded in the case of trial-level errors. Trial-level errors are errors that affect a single trial, but do not impact the entire session. For example, the experimenter might accidentally end a trial before the infant has finished looking, or a parent might briefly interfere during one trial. These errors would render the specific trial unusable, but

[2] For context, *d* = 0.25 is the average effect size from a meta-analysis of published rule learning studies (Rabagliati et al 2019), while *d* = 0.3 is the average effect size across all published developmental experiments. *d* = 0.2 is generally considered a 'small' effect size, and thus *d* = 0.1 is quite small.

are unlikely to affect other trials. In these cases, only the specific trial will be classified as an error and excluded from analyses. As looking times will be log-transformed, trials with zero looking time will also be excluded from the analyses (Csibra et al. 2016). In total there were XXX trials that were excluded for trial-level reasons.

The final sample consisted of XXX infants included in the analysis, and, after exclusion of the trial-level errors, they completed an average of XXX trials, out of a total possible of 12. The total number of test trials available for analysis was XXX.

Dependent and independent variables

The following variables will be included in the analyses (variable names are bolded and, for categorical variables, levels are italicized):

- Infants' looking time (**LT**) is our dependent variable. Looking time is defined as the time spent looking at the screen (for CF and ET methods and some HPP set-ups) or at the light (HPP) during test trials. As looking times are non-normally distributed, they will be log-transformed prior to statistical analyses, following Csibra et al. (2016).
- **Familiarization_rule** indicates the sequence to which infants will be exposed during familiarization (either *ABA* or *ABB*). Infants were exposed to only one sequence, with the sequence determined by random assignment. This variable is a binary variable in which trials presenting infants with the ABA rule will be coded as 0 (baseline) and trials presenting infants with the ABB will be coded as 1.
- Trial type (**trial_type**) indicates for each test sequence whether it follows the same rule to which the infant is familiarized or a different rule. For example, if an infant will hear an

ABA rule during familiarization, ABA tes trials will be the *same* trial type and ABB test trials will be the *different* trial type). This variable is a binary variable in which the same trial type is coded as 0 (baseline) and a different trial type is coded as 1.

- Test with **repetition** sequence; whether a test trial was ABB or not.
- Trial number (**trial_num**) indicates the sequential order in which test trials are presented. Trial number thus ranges from *1* to *12*.
- Infant **age** indicates the infants' age in days. Age will be centered before entering the models.
- Experimental **method** indicates the method that was used to record infants' responses to the stimuli, one of: headturn preference procedure (*HPP*), central fixation (*CF*), or eye tracking (*ET*); CF will be used as baseline in the analyses.
- **Lab** indicates the laboratory in which data were collected. Each laboratory has a unique laboratory identifier.
- **Subject** indicates the unique subject identifier for each infant tested.
- Multilingual exposure (**multilingual_exposure**) indicates the infants' exposure to other languages than the primary language, with a minimum of 0%, meaning no exposure to any but the primary language. Infants with higher scores have more bilingual/multilingual exposure.
- **Primary language**. The language with the largest exposure percentage.

Besides the above variables that are part of our planned analyses, a number of other variables will be collected that may be useful in later secondary data-analyses, e.g. participant gender and demographic background variables.

Statistical modelling

To test our hypotheses, we will use mixed-effect models to examine effects of the independent variables (IV) on the dependent variable (DV). Mixed-effect models allow us to control for random effects in both the DV ("random intercepts") and the relationship of the IV to the DV ("random slopes") based on relevant grouping units (subjects and laboratories). We will begin with a fairly complex model including a maximal random effects structure (Barr et al., Levy, Scheepers, & Tily, 2013)(listed below) that we plan to prune in the case of non-convergence. We will prune the model by first removing random effects that are theoretically less important (see Pruning the model Section below), and then repeat this procedure by gradually removing other random effects until the model converges. All models will be fitted using the lme4 package (Bates et al., 2015) and *p*values will be computed using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

The mixed-effect model is specified below in lme4 R codes. Using R codes, lower-order effects are subsumed by interactions in the model, even though they are not explicitly written. For example, the interaction between age and trial_num (age * trial_type) enters a main effect of age and a main effect of trial type into the model. Our initial maximal random effects model is:

Infants' looking time (DV) ~ 1

+ familiarization_rule*trial_type

+ age * trial_type

+ experimental_method * trial_type

+ multilingual_exposure * trial_type

+ trial_num*trial_type

+ trial_num*age

+ repetition*trial_type

+ (trial_num*trial_type | subject)

+ (trial_type + test_order | lab)

The model will include fixed main effects for familiarization rule, trial type, trial number, infants' age, experimental method, multilingual exposure, repetition, and primary language. Main effects will test how the IV affects infants' looking time. For example, trial_number models the potential decline of infants' looking time over subsequent test trials, and trial_type examines whether infants look at the same and different rules at test differently. Importantly, we enter a number of two-way interactions as fixed effects to test our research questions (see details below where we outline which statistical tests relate to which hypothesis).

We control the subject-level and laboratory-level groupings with the random effects structure of the model. For the subject-level grouping, the maximal model includes a random intercept and controls the interaction between trial number and trial type using random slopes. Controlling the interaction between trial number and trial type allows us to model the differences

in habituation rate across individuals and individuals' preference to the same/different rules at test. For the laboratory-level grouping, in the maximal model we also enter a random intercept. In addition to that, we allow for random slopes by test order per lab as part of the maximal random effects structure that is justified by the design (Barr et al., 2013).

As some of the tested effects of interest are also informative when they show a null effect we will also report Bayes factors next to the frequentist statistics. Bayesian models will be estimated using the brms package in R (Bürkner, 2017).

Pruning the model

When the model with the maximal random effects structure does not reach convergence, random effects will be pruned in the following order of relevance (that is, in order of to-be-dropped terms):

1. Change Subject level random slope from interaction to two main effects: trial_type and trial_num
2. Lab level slope trial_type
3. Lab level slope test_order
4. Subject level slope trial_num
5. Subject level slop trial_type
6. Lab intercept
7. Subject intercept

Hypotheses tests

In Table 7 below we provide the model terms, their connection with specific hypotheses, and their expectation (positive, negative, null). Note that indeed some effects are expected to be

null, and their being null is informative to the literature, hence Bayes factors are required to test these models.

Table 7

Hypotheses, Statistics and Expected Direction of Effects

Hypothesis	**Model term**	**D**
Hypothesis 1a: Rule learning is evidenced by a novelty preference for inconsistent patterns at test (with same as baseline).	trial_type	+
Hypothesis 1b: After controlling for age, linguistic background, and experimental paradigm, there is no residual heterogeneity between labs.	Random variances across lab	0
Hypothesis 2a: Rule learning is robust across age and increases with age.	age:trial_type	+
Hypothesis 2b: Looking times decrease with test trial; this effect will be stronger for older infants than younger infants.	trial_num:age	+
Hypothesis 2c: The relationship between multilingualism and rule learning will be assessed.	multilingual_exposure: trial_type	+ /0 /-

Hypothesis 2d: Rule learning is robust across variation in experimental paradigms and is facilitated in the Headturn Preference Procedure relative to other paradigms (with CF serving as baseline; also implicitly meaning a null effect of ET relative to baseline).	experimental_method: trial_type	+
Hypothesis 3a: There is expected to be an overall preference, i.e. longer looking times, for direct repetition items in the test phase, i.e. ABB items.	repetition	+
Hypothesis 3b: There will be a larger effect of learning overall in the participants trained on ABB than in the participants trained on ABA (with ABA as baseline).	familiarization: trial_type	+

Note. Hypotheses are those listed in the introductory sections; model term has the R code specification of the model coefficient testing the particular hypothesis; D denotes the expected sign of the coefficient (+ or - or 0; the latter being for expected null effects).

*****Hypotheses 2b, 3c and 3d may turn out to be infeasible as they involve large numbers of factors (primary languages, 3-way interactions). There are explicit expectations for these hypotheses and hence we will try to test them. Another way to approach 2b is to look for patterns of cultural variation and/or by clustering groups of primary languages.

Additional hypotheses

If we have collected sufficient data that permits testing further effects, a number of additional hypotheses will be tested. In particular these are:

1) Rule learning is robust and does not vary across primary languages. Related to the universality of rule learning we will test whether rule learning varies with infants' primary language. As the number of primary languages is likely to be large and collinear with laboratories this will only be explored after fitting the main model. This can be tested by including the primary_language:trial_type fixed effect into the model.
2) There will be an interaction effect between age and non-adjacent repetition (i.e. ABA), with older infants showing more learning from ABA training stimuli than younger infants. This hypothesis can be tested by including familiarization:trial_type:age into the model. This being a 3-way interaction in an already complex model, we will only add this after fitting the main model.
3) The relationship between multilingualism and non-adjacent repetition learning will be explored. Either negative, positive or null result will inform the debate about potential cognitive benefits for multilingual infants. This can be tested by adding familiarization:trial_type:multilingual_exposure into the model. Similar to the previous hypothesis 2, this involves a 3-way interaction and hence we will only add this after fitting the main model.

Robustness analyses

In order to check the robustness of our results, we will run some checks. In particular, we will do this in the form of a multiverse analysis (Steegen et al. 2016). A large advantage of pre-registering a study prior to performing it and running the analyses is that we are forced to be as specific as possible in laying out the choices that are made in pre-processing the data, determining exclusion criteria, and the analyses that will be performed. Many of these choices

are well justified based on previous research. For some choices, however, it is simply unknown what their effects are. In all of these cases it is interesting to find out whether our main results are robust to making different choices. Hence, here we will re-run the main model when different choices are made. In particular, the choices that we consider are the following:

1. **Minimum looking time per trial**: currently there is no minimal looking time for excluding trials (except that zero looking time trials are discarded). In looking time research it is quite common to use some lower threshold for any look to constitute showing any attention - which is what the overall looking time is purported to measure. In the robustness analyses we will explore setting the threshold at 100-200-300-400 ms; for reference, average fixation durations in the first year of life decrease from about 400 ms to 350 ms in free viewing situations (Helo et al, 2016).
2. **Minimum number of trials**: in the main model we include infants with as few as 2 out of 12 valid trials, assuming these trials are of different type; it could be argued that a larger minimum is required to reliably assess any novelty/familiarity effect; we will include thresholds of 3, 4, 6, 9, 12 respectively to test the robustness of the results (in all cases also assuming a minimum of 1 same/1 different trial).
3. **Test-trials within a block**: in the main model we included infants with a minimum of 2 completed test trials, one ‘same’ and one ‘different’; as a result, these test-trials can be very far apart (spanning several invalid trials) which could potentially lessen the rule learning effect; here, instead, we will only include infants if they had a pair (one ‘same’, one ‘different’) of valid test-trials within a block of 4 trials.

4. **Second-session infants**: in the main model we include both first- and second-session infants; here we exclude second-session infants.

Together, these robustness checks can help determine the generalisability of our results in the face of different analytical choices. In other words, they can strengthen the confidence in our results given consistent results across these choices. In the case of inconsistent results, these checks can help determine causes of variability in infant looking time research.

Results

XYZ

Discussion

XYZ

References

Anderson, J. R. (1996). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

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, 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2021). Linear mixed-effects models using 'Eigen' and S4 (computer software). https://github.com/lme4/lme4/

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. Child Development, 89, 1996-2009. https://doi.org/10.1111/cdev.13079

Boersma, P., & Weenink, D. (2019). Praat, Version 6.0 (computer software). Amsterdam: University of Amsterdam.

Bürkner, P.-.C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. Canadian Psychology/Psychologie canadienne, 61, 349–363. https://doi.org/10.1037/cap0000216

Chen, J., van Rossum, D., & ten Cate, C. (2015). Artificial grammar learning in zebra finches and human adults: XYX versus XXY. Animal Cognition, 18, 151-164.

Chomsky, N. (1980). *Rules and representations*. New York, NY: Columbia University Press.

Comishen, K. J., Bialystok, E., & Adler, S. A. (2019). The impact of bilingual environments on selective attention in infancy. Developmental Science, 22, e12797. https://doi.org/10.1111/desc.12797

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. Developmental Psychology, 52, 521–536. https://doi.org/10.1037/dev0000083

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. Cognition, 111, 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

de la Mora, D. M., & Toro, J. M. (2013). Rule learning over consonants and vowels in a non-human animal. Cognition, 126, 307-312. https://doi.org/10.1016/j.cognition.2012.09.015

Doebel, S., & Zelazo, P. D. (2015). A meta-analysis of the Dimensional Change Card Sort: Implications for developmental theories and the measurement of executive function in children. Developmental Review, 38, 241-268. https://doi.org/10.1016/j.dr.2015.09.001

D'Souza, D., Brady, D., Haensel, J. X., & D'Souza, H. (2020). Is mere exposure enough? The effects of bilingual environments on infant cognitive development. Royal Society Open Science, 7, 180191. http://dx.doi.org/10.1098/rsos.180191

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. Trends in Cognitive Sciences, 13, 348–353. https://doi.org/10.1016/j.tics.2009.05.005

Ferguson, B., Franconeri, S. L., & Waxman, S. R. (2018) Very young infants learn abstract rules in the visual modality. PLoS ONE, 13, e0190185. https://doi.org/10.1371/journal.pone.0190185

Fodor, J. A. (1981). The mind-body problem. Scientific American, 244, 114-123. https://www.jstor.org/stable/24964264

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28, 3-71. https://doi.org/10.1016/0010-0277(88)90031-5

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. Developmental Science, 12, 504-509. https://doi.org/10.1111/j.1467-7687.2008.00794.x

Geambasu, A. (2018). *Simple rule learning is not simple. Studies on infant and adult pattern perception and production.* [Doctoral Dissertation, LOT, Netherlands Graduate School]. http://hdl.handle.net/1887/67540

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. Cognition, 98, B67–B74. https://doi.org/10.1016/j.cognition.2005.03.003

Gervain, J., Macagno, F., Cogoi, S., Peña, M., & Mehler, J. (2008). The neonate brain detects speech structure. Proceedings of the National Academy of Sciences (USA), 105, 14222–14227. https://doi.org/10.1073/pnas.0806530105

Gleason, J.B. & Ratner, N.B. (2017). *The Development of Language* (9th ed). New York, NY: Pearson.

Green, D.M. and Swets, J.A. (1966). Signal detection theory and psychophysics. New York, NY: Wiley.

Hoff, E. (2014). *Language development* (5th ed). Belmont, CA: Wadsworth.

Hauser, M. D., & Glynn, D. (2009). Can free-ranging rhesus monkeys (Macaca mulatta) extract artificially created rules comprised of natural vocalizations? Journal of Comparative Psychology, 123, 161–167. https://doi.org/10.1037/a0015584

Helo, A., Rämä, P., Pannasch, S., & Meary, D. (2016). Eye movement patterns and visual attention during scene viewing in 3-to 12-month-olds. Visual Neuroscience, 33.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N. Z., Marcus, G. F., Rabagliati, H., & Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8- and 11-month-olds. Infancy, 14, 2-18. https://doi.org/10.1080/15250000802569611

Kail, M. & Fayol, M. (2015). L'acquisition du langage: Le langage en développement. Au-delà de 3 ans. Paris, France: Presses universitaires de France.

Kalashnikova, M., Pejovic, J., & Carreiras, M. (2021). The effects of bilingualism on attentional processes in the first year of life. Developmental Science, 24. https://doi.org/10.1111/desc.13011

Kopp, B., Maldonado, N., Scheffels, J. F., Hendel, M., & Lange, F. (2019). A meta-analysis of relationships between measures of Wisconsin Card Sorting and intelligence. Brain Sciences, 9, 349. http://dx.doi.org/10.3390/brainsci9120349

Kovács, A. M. (2014). Extracting regularities from noise: Do infants encode patterns based on same and different relations? Language Learning, 64, 65–85. https://doi.org/10.1111/lang.12056

Kovács, A. M., & Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. Proceedings of the National Academy of Sciences (USA), 106, 6556-6560. https://doi.org/10.1073/pnas.0811323106

Kovács, A. M., & Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. Science, 325, 611–612. https://doi.org/10.1126/science.1173947

Kuhn, D. (2012). The development of causal reasoning. WIREs Cognitive Science, 3, 327-335. https://doi.org/10.1002/wcs.1160

Kunkel, J. H. (1997). The analysis of rule-governed behavior in social psychology. The Psychological Record, 47, 699-716. https://doi.org/10.1007/BF03395254

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. Journal of Statistical Software, 82(13). https://doi.org/10.18637/jss.v082.i13

Laird, J. E. (2012). *The Soar cognitive architecture*. Cambridge, MA: MIT Press.

Lust, B. (2006). *Child language: Acquisition and growth*. Cambridge, MA: Cambridge University Press.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. Advances in Methods and Practices in Psychological Science, 3, 24-52. https://doi.org/10.1177/2515245919900809

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant rule learning facilitated by speech. Psychological Science, 18, 387-391. https://doi.org/10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science 283, 77–80. doi: 10.1126/science.283.5398.77

O'Grady, W. (2005). *How children learn language*. Cambridge, MA: Cambridge University Press.

Pothos, E. M. (2005). The rules versus similarity distinction. Behavioral & Brain Sciences, 28, 1–49. https://doi.org/10.1017/S0140525X05000014

Pour Iliaei, S., Killam, H., Dal Ben, R., & Byers-Heinlein, K. (2021). *Bilingualism affects infant cognition: Insights from new and open data* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ex76a

Pylyshyn Z.W. (1984). Computation and cognition: Toward a foundation for cognitive science. Cambridge, MA: MIT Press.

Prior, A., & MacWhinney, B. (2010). A bilingual advantage in task switching. Bilingualism: Language and Cognition 13, 253–262. https://doi.org/10.1017/S1366728909990526

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. Developmental Science, 22, e12704. https://doi.org/10.1111/desc.12704

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Pollack, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition, 105, 669–680. https://doi.org/10.1016/j.cognition.2006.11.004

Saxton, M. (2010). *Child Language: Acquisition and Development* (2nd Ed.). Thousand Oaks, CA: Sage.

Santolin, C., Rosa-Salva, O., Regolin, L., & Vallortigara, G. (2016). Generalization of visual regularities in newly hatched chicks (Gallus gallus). Animal Cognition, 19, 1007–1017. https://doi.org/10.1007/s10071-016-1005-2

Schonberg, C., Marcus, G. F., & Johnson, S. P. (2018). The roles of item repetition and position in infants' abstract rule learning. Infant Behavior and Development, 53, 64-80. https://doi.org/10.1016/j.infbeh.2018.08.003

Siegler, R. S. (1983). Five generalizations about cognitive development. American Psychologist, 38, 263–277. https://doi.org/10.1037/0003-066X.38.3.263

Spierlings, M. J., & ten Cate, C. (2016). Budgerigars and zebra finches differ in how they generalize in an artificial grammar learning experiment. Proceedings of the National Academy of Sciences (USA), 113, E3977-E3984. https://doi.org/10.1073/pnas.1600483113

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11, 702-712. https://doi.org/10.1177/1745691616658637

Sun, R. (2008). *The Cambridge handbook of computational psychology*. Cambridge, MA: Cambridge University Press.

Ten Cate, C., & Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philosophical Transactions of the Royal Society B, 367, 1984-1994. https://doi.org/10.1098/rstb.2012.0055

Thiessen, E. D. (2012). Effects of inter- and intra-modal redundancy on infants' rule learning. Language Learning and Development, 8, 197–214. https://doi.org/10.1080/15475441.2011.583610

Tsui, A. S. M., Ma, Y. K., Ho, A., Chow, H. M., & Tseng, C. (2015). Bimodal emotion congruency is critical to preverbal infants' abstract rule learning. Developmental Science, 19, 382-393. https://doi.org/10.1111/desc.12319

Tsui, A. S. M. & Fennell, C. T. (2019). Do Bilingual Infants Possess Enhanced Cognitive Skills? In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society, 3001-3007. Montreal, QB: Cognitive Science Society.

Tseng, C., Chow, H. M., Ma, Y. K., & Ding, J. (2018). Preverbal infants utilize cross-modal semantic congruency in artificial grammar acquisition. Scientific Reports, 8, 12707. https://doi.org/10.1038/s41598-018-30927-3

van Heijningen, C. A. A., Chen, J., van Laatum, I., van der Hulst, B., & ten Cate, C. (2013). Rule learning by zebra finches in an artificial grammar learning task: Which rule? Animal Cognition, 16, 165-175. https://doi.org/10.1007/s10071-012-0559-x

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. Developmental Review, 38, 55-68. https://doi.org/10.1016/j.dr.2015.07.001