

# **ManyBabies 5: A large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time**

Pre-data collection manuscript for peer-review

The ManyBabies 5 Team

Jessica E. Kosie; Arizona State University; jkosie@asu.edu  
Martin Zettersten; Princeton University; martincz@princeton.edu  
Rana Abu-Zhaya; University College London  
Dima Amso; Columbia University  
Mireille Babineau; University of Toronto  
Heidi A. Baumgartner; Stanford University  
Marina Bazhydai; Lancaster University  
Margherita Belia; University of York  
Silvia Benavides-Varela; University of Padova  
Christina Bergmann; Max Planck Institute for Psycholinguistics and Osnabrück University of Applied Sciences  
Ilaria Berteletti; Gallaudet University  
Alexis K. Black; University of British Columbia  
Priscila Borges; University of Vienna  
Arielle Borovsky; Purdue University  
Krista Byers-Heinlein; Concordia University  
Laurianne Cabrera; Université de Paris Cité- CNRS  
Giulia Calignano; University of Padova  
Anjie Cao; Stanford University  
Hitomi Chijiwa; Osaka University  
Christopher M. M. Cox; Aarhus University  
Rodrigo Dal Ben; Ambrose University  
Isabelle Dautriche; CNRS, Aix-Marseille University  
Michaela C. DeBolt; University of California, Davis  
Anna Exner; Ruhr University Bochum  
Donna Fisher-Thompson; Niagara University  
Samuel H. Forbes; Durham University  
Laura Franchin; University of Trento  
Michael C. Frank; Stanford University  
Gökhan Gönül; University of Neuchâtel  
Nayeli Gonzalez-Gomez; Oxford Brookes University  
Charlotte Grosse Wiesmann; Max Planck Institute for Human Cognitive and Brain Sciences

J. Kiley Hamlin; University of British Columbia  
 Erin E. Hannon; University of Nevada Las Vegas  
 Naomi Havron; University of Haifa  
 Jean-Remy Hochmann; CNRS, Institut des Sciences Cognitives Marc Jeannerod  
 Stefanie Hoehl; University of Vienna  
 Carmel Houston-Price; University of Reading  
 George Kachergis; Stanford University  
 Zsuzsa Kaldy; University of Massachusetts Boston  
 Osman S. Kingo; Aarhus University  
 Simon Kizito; Makerere University  
 Eon-Suk Ko; Chosun University  
 Nina-Alisa Kollakowski; LMU Munich  
 Shannon P. Kong; Oxford Brookes University  
 Vanja Kovic; University of Belgrade  
 Peter Krøjgaard; Aarhus University  
 Shari Liu; Johns Hopkins University  
 Belén López Assef; University of Ottawa  
 Helen S. Lu; University of Southern California  
 Madhaviatha Maganti; Ashoka University  
 Olivier Mascaro; Université de Paris Cité- CNRS  
 Emily Mather; University of Hull  
 Julien Mayor; University of Oslo  
 Brianna T. M. McMillan; Smith College  
 Marek Meristo; University of Gothenburg  
 Toben H. Mintz; University of Southern California  
 Monika Molnar; University of Toronto  
 David Moreau; University of Auckland  
 Yusuke Moriguchi; Kyoto University  
 Margaret C. Moulson; Toronto Metropolitan University  
 Jutta L. Mueller; University of Vienna  
 Lisa M. Oakes; University of California, Davis  
 Sharon Peperkamp; CNRS, École Normale Supérieure  
 Stefanie Peykarjou; Heidelberg University  
 Mónica Taveira Pires; Universidade Autónoma de Lisboa  
 Gal Raz; MIT  
 Jennifer L. Rennels; University of Nevada, Las Vegas  
 Pablo E. Requena; University of Texas at San Antonio  
 Joscelin Rocha-Hidalgo; Pennsylvania State University  
 Jenny Saffran; University of Wisconsin - Madison  
 Christina Schaetz; University of Vienna

Tobias Schuwerk; Ludwig-Maximilians-Universität München  
Kimberly Scott; MIT  
Jeanne L. Shinskey; Royal Holloway, University of London  
Elizabeth A. Simpson; University of Miami  
Leher Singh; National University of Singapore  
Sylvain Sirois; University of Quebec at Trois-Rivieres  
Erin Smolak; University of South Carolina  
Melanie Soderstrom; University of Manitoba  
Trine Sonne; Aarhus University  
Céline Spriet; Université de Paris Cité- CNRS  
Andrew Sentoogo Ssemata; Makerere University  
Ingmar Visser; University of Amsterdam  
Katie Von Holzen; TU Braunschweig  
Sandra R. Waxman; Northwestern University  
Gert Westermann; Lancaster University  
Katherine S. White; University of Waterloo  
Kali Woodruff Carr; Northwestern University  
Naiqi G. Xiao; McMaster University  
Linlin Yan; Zhejiang Sci-tech University  
Katharina Zahner-Ritter; University of Trier  
Tania S. Zamuner; University of Ottawa  
Henriette Zeidler; Aston University  
Xi Jia Zhou; Stanford University  
Lucie Zimmer; Ludwig Maximilian University of Munich  
Zorana Zupan; University of Belgrade  
Casey Lew-Williams; Princeton University

## Abstract

Much of our basic understanding of cognitive and social processes in infancy relies on measures of looking time, and specifically on infants' visual preference for a novel or familiar stimulus. However, despite being the foundation of many behavioral tasks in infant research, the determinants of infants' visual preferences are poorly understood, and differences in the expression of preferences can be difficult to interpret. In this large-scale study, we test predictions from the Hunter and Ames model of infants' visual preferences. We investigate the effects of three factors predicted by this model to determine infants' preference for novel versus familiar stimuli: age, stimulus familiarity, and stimulus complexity. Drawing from a large and diverse sample of infant participants (minimum expected sample size  $N = 1,280$ ), this study aims to provide empirical evidence for a robust and generalizable model of infant visual preferences, leading to a more solid theoretical foundation for understanding the mechanisms that underlie infants' responses in common behavioral paradigms. Moreover, we hope that our findings will guide future studies that rely on infants' visual preferences to measure cognitive and social processes.



## Introduction

Scholars and parents alike have long been fascinated by questions concerning the infant mind. What are infants thinking, seeing, hearing, and feeling, and from where and when do these abilities arise? Because it is not possible to query infants directly, scientists must rely on indirect techniques to study early development. However, many of the measures and procedures used with older children and adults, such as button presses or responding to questions, are difficult or impossible to use with infants. This has led researchers to adopt an indirect behavioral measure – infants' looking toward visual stimuli – as one of the primary measures for uncovering the development of psychological processes in infancy.<sup>1</sup>

Much of the prior research using infants' looking time is based on the seminal finding that infants, following repeated exposure to a particular stimulus, typically demonstrate a preference to look at a relatively new stimulus.<sup>2</sup> This novelty preference has been well-documented for both visual and auditory stimuli.<sup>3–7</sup> The approach in these looking time studies is often to present one or more stimuli until infants' interest in the familiar stimuli is reduced, thereby inducing preferences for novelty.<sup>8</sup> However, infants' responses, as well as scientists' interpretations of these responses, vary substantially.<sup>1</sup> Although many studies have found that infants show a robust novelty preference,<sup>9–11</sup> there are also conditions under which infants prefer familiar stimuli over novel stimuli.<sup>3,12–16</sup> Even within specific research areas, there are usually examples of both directions of preference, and there are no clear reasons as to why.<sup>5,17</sup> Furthermore, in some cases, the same type of stimulus that draws infants' attention in one paradigm can fail to draw attention in another.<sup>18,19</sup>

Despite this variability, scientists have generally focused on novelty preferences as a basis for making inferences about infants' cognitive processing. This idea is consistent with

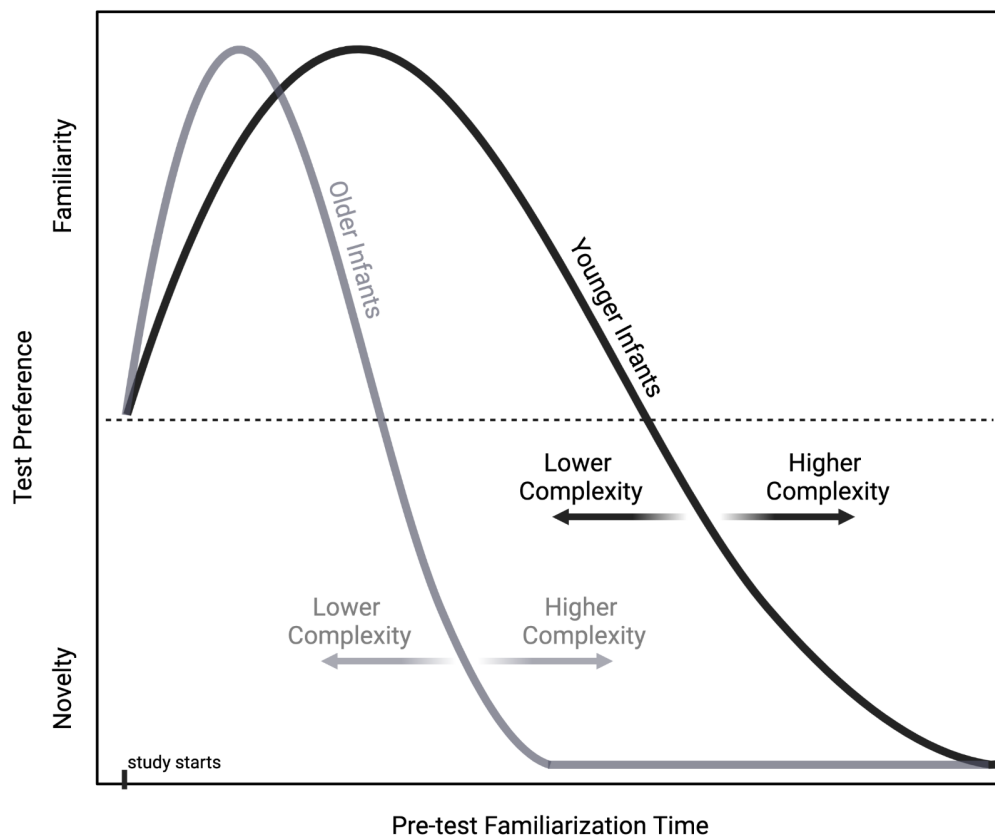
arguments from multiple disciplines suggesting that stimulus novelty may serve as a key signal for exploration and learning.<sup>20</sup> That is, the novelty of a stimulus may engage intrinsic motivation mechanisms that, in turn, drive organisms to invest cognitive resources in learning under conditions of uncertainty.<sup>21</sup> Preferences for novelty have been identified in numerous diverse species<sup>22</sup> and have been linked with both basic survival processes like locomotor adaptation<sup>23</sup> and high-level constructs like curiosity.<sup>24–27</sup> Understanding the determinants of infants' familiarity and novelty preferences, therefore, is important not only for methodological reasons but also for clarifying mechanisms of cognitive development.

The most influential account of looking preferences in the human infant literature is the conceptual model put forth by Hunter and Ames (henceforth, the Hunter and Ames model<sup>28</sup>), which was based on several studies.<sup>3,29,30</sup> Hunter and Ames aimed to account for the circumstances under which infants show familiarity vs. novelty preferences while processing perceptual stimuli. The model assumes that when one stimulus is repeatedly presented, infants gradually develop a preference for this familiar stimulus compared to a second, novel stimulus. After continued exposure to the familiar stimulus, infants' preferences eventually switch, and they begin to show a preference for the novel stimulus (see Figure 1). The key premise of the model is that the degree to which infants prefer familiar vs. novel stimuli depends on three main factors and their interactions (see Figure 1):

- (1) Familiarization time.** With increasing familiarization to a stimulus, infants transition from preferring a familiar stimulus to preferring a novel stimulus.
- (2) Infant age.** The transition from preferring a familiar stimulus to preferring a novel stimulus is faster for older infants than younger infants.

**(3) Task difficulty.** As task difficulty increases (operationalized in the current study as **stimulus complexity**), infants make a slower transition from preferring a familiar stimulus to preferring a novel stimulus.

At the time of writing, this influential model has been used to explain findings within a broad range of papers examining visual attention, speech segmentation, object categorization, face recognition, quantification, memory, music experience, knowledge of the physical world, and social cognition, among others.<sup>4,12,31–34</sup>



*Figure 1.* This figure (adapted from Bergmann and Cristia<sup>5</sup>) depicts Hunter & Ames' model<sup>28</sup> of infant looking to familiar (plotted up) and novel (plotted down) stimuli at different levels of pre-test familiarization time (shown on the x-axis). The dashed line indicates equal preference for familiar and novel stimuli. The grey line represents the model's prediction for older infants, and the black line for younger infants. The "lower complexity" and "higher complexity" arrows

indicate how the proposed relation between age and familiarization time might shift with variation in stimulus complexity (i.e., the current operationalization of task difficulty).

However, this model has not undergone a rigorous and comprehensive test, and the predictions from the model have not been supported unequivocally by the existing literature. Whereas several studies have demonstrated the proposed shift from familiarity to novelty preferences,<sup>35–38</sup> other studies have found limited evidence for this shift.<sup>5,39–42</sup> More generally, there is minimal consensus on the factors that drive a familiarity preference in some conditions and a novelty preference in others. This gap in understanding about a core component of visual preferences in infancy has raised numerous interpretive challenges over the decades, including when to predict a preference for novelty versus familiarity and how to interpret unexpected, unstable, or null preferences across studies.<sup>36,37,43,44</sup>

A large-scale investigation into the fundamental tenets of the Hunter and Ames model is an important step toward quantifying relations between infant age, familiarization time, task difficulty, and preference for familiar vs. novel stimuli. In the present study, we bring together researchers from around the world to create and implement a best-test of key predictions of the Hunter and Ames model. This investigation will enable a broad community of researchers to improve their interpretations of infants' visual preferences and thereby enable new and more robust insights into the infant mind.

There are two main challenges in evaluating the Hunter and Ames model. First, most behavioral studies with infants have lacked sufficient statistical power due to the inherent limitations within any given lab in testing large numbers of participants.<sup>45,46</sup> Recruiting participants and obtaining a high data 'yield' are notoriously challenging in infant research. Moving forward, infant research is in need of solutions to achieve higher statistical power in

order to improve its reliability and replicability.<sup>47</sup> A second, related challenge is that most behavioral studies with infants have reported findings based on a geographically and culturally restricted sample, causing limited generalizability across populations. This is in part a consequence of the over-representation of researchers in North America and Western Europe, who tend to recruit primarily local families who have the means and the time for a lab visit. The result of this tradition of convenience has been an overreliance on samples of White infants from middle-class families – a biased sample of the world’s population when making conclusions about both human development and human nature.<sup>48,49</sup> Overcoming these challenges is essential to the development of a robust and generalizable model of infants’ visual preferences.

We address these challenges by providing a large-scale, high-powered test of the Hunter and Ames model through an international collaboration, harnessing the infrastructure of ManyBabies.<sup>50</sup> ManyBabies is a six-continent network of scientists who are interested in understanding key theoretical questions about early child development, promoting best practices in developmental research, and broadening participation to include scientists and families from a broad range of communities and cultures. Drawing on this network, we designed an experiment testing visual preferences for familiar and novel stimuli that will include a large and diverse infant sample (minimum expected sample size  $N = 1,280$ ), enabling an evaluation of the extent to which the Hunter and Ames model generalizes across communities and cultures. We test three specific hypotheses (see Table 1): Infants’ preferences for novel stimuli will be stronger (1) following longer familiarization time, (2) as infants grow older, and (3) for simpler vs. more complex stimuli.

## Method

### Ethics

Prior to beginning data collection, all labs contributing data will be asked to complete a survey that asks for information about, among other things, their lab's ethics approval procedures. Labs will indicate whether they have new or existing approval from their local ethics committee or institutional review board to conduct this study or whether the current study does not require ethics approval at their institution. The experiment is currently approved by the Office of Research Integrity and Assurance at Princeton University (Protocol #7117).

For all labs, we will require the following consent procedure: A parent or legal guardian will provide informed consent for each participating infant, though the specifics of consenting procedures will vary across labs to comply with local legal requirements and cultural norms. Families will be compensated according to each lab's standard practices as approved by their local ethics board. All de-identified data will be stored on the ManyBabies 5 Open Science Foundation repository ([https://osf.io/g3udp/?view\\_only=0f4f6e3b48d8456999d6459c0bfe5510](https://osf.io/g3udp/?view_only=0f4f6e3b48d8456999d6459c0bfe5510)).

### Participation Details

An initial open call to participate in “ManyBabies 5: Hunter and Ames” was issued on July 23, 2020 via social media and developmental science listservs. Over the next three years, scientists from around the world contributed in various ways to study design (see Figure S1 in Supplementary Materials). Each participating lab will be asked to contribute a sample of at least 32 infants between the ages of 3 and 15 months; effort will be made across labs to ensure that age is distributed evenly across this age span, to the extent possible. Because many of our analyses will examine effects across labs rather than within a single lab, we will also allow labs to contribute a “half sample” of 16 infants. As in the ManyBabies 1 project,<sup>51</sup> allowing “half

samples” will increase the number of labs capable of contributing to data collection and will enable participation from more labs, especially those from under-resourced and/or underrepresented communities. We will ask that the sample size per lab (a full sample of 32 or half-sample of 16) includes any infant that enters the lab and not the number of infants retained after exclusion criteria are implemented because final decisions about participant exclusion will take place centrally.

## **Participants**

We will recruit a minimum final sample of at least 1,280 infants between the ages of 3 and 15 months. The minimum sample size was determined based on initial surveys of potential contributors, indicating that at least 40 labs plan to contribute data from approximately 32 infants each. This minimum sample size will allow us to detect an effect at least as large as  $d = 0.2$  for all main confirmatory analyses with at least 95% power (Table 1; see the Power Analysis section for a justification of the smallest effect size of interest). The final sample will be determined based on all infants collected as part of ManyBabies 5 by participating labs. The 3- to 15-month age span was chosen because it covers the majority of ages for which the Hunter and Ames model has been used to describe infants’ responses to stimuli in previous research. Because the ManyBabies model makes it possible to collect a sizable sample of infants with a variety of developmental histories, we will adopt an inclusive recruitment strategy. Labs will be encouraged to recruit (using lab-standard practices) any infant who meets just two criteria: (1) the infant’s age falls within the 3- to 15-month age range, and (2) the infant has no known issues that would directly impede their ability to process visual stimuli (i.e., visual impairments). However, we will additionally collect data about the criteria on which infants are typically excluded in developmental studies, such as prematurity, developmental delays or disorders,

family history of colorblindness, and hearing status. For the analyses of interest in the current paper, we will exclude infants when there is a reason to expect that the predictions of our model would differ based on their developmental history (such as prematurity; see inclusion criteria below). However, all data will be retained for potential future analyses beyond the scope of the current paper (e.g., investigating the effects of prematurity on infants' preferences for novel versus familiar stimuli<sup>52</sup>).

In some labs, we expect that infants will also be tested in other experiments during their visit. In these cases, we will strongly encourage labs to administer the current study first because it is commonly observed in infant studies that “second session” experiments result in greater dropout rates. If labs absolutely must administer this study second, we will ask them to report whether infants participated in the current study first or second and, when the infant participated in the current study second, we will ask for information about the first study. All infants will be included in the primary analyses (following previous ManyBabies studies<sup>51</sup>), but information about first versus second session will be retained for exploratory analyses examining the effect of first versus second session testing.

### ***Participant Demographics***

Basic demographic data will be collected for all participants using the ManyBabies Demographics questionnaire.<sup>53</sup> This questionnaire asks about participants' biographical information, race and ethnicity, gestational age, caregiver information, family socio-economic status, language exposure, and relevant developmental concerns. We will supplement the questionnaire with several experiential variables relevant to visual processing (e.g., screen exposure, object experience, and family history of color blindness). Researchers from different communities, nations, and cultures will be able to modify the content of the questionnaire in




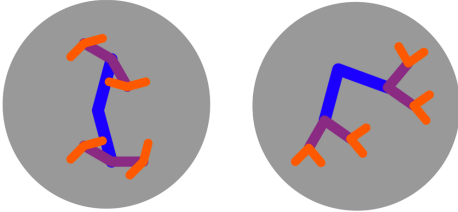

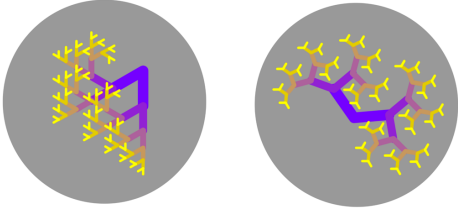
structured ways to ensure appropriateness of the questions for local contexts while preserving, to the extent possible, comparability across test sites. Demographic information will be used for exploratory analyses, as described in the Supplementary Materials (S6.5.).

## Materials

**Visual Stimuli.** In the current experiment, we chose to operationalize *task difficulty* in terms of stimulus complexity: we assume that task difficulty increases as stimuli become more complex. Because the sample size possible in ManyBabies projects provides a unique opportunity to test generalizability across multiple stimulus types, we chose two different categories of stimuli: *fribbles*<sup>54–58</sup> and *fractals*.<sup>59,60</sup> Fribbles are colorful, computer-generated models of 3D figures. All fribbles are made up of one of 12 possible body shapes and up to four appendages that can be chosen from a set of possible options. Fractals are geometric shapes in which the same structure is repeated iteratively at different levels of scale. These two classes of stimuli were chosen because: (1) they allow for variation in stimulus complexity within the same stimulus category (i.e., by manipulating the number of features); (2) the inclusion of two stimulus sets allows us to begin examining the generalizability of any observed effect; and (3) infants are unlikely to have previous experience with either set of stimuli, which increases the likelihood that they are appropriate across communities and cultures.

We created 12 unique fribbles and 12 unique fractals, operationalizing complexity as the number of features present (these features are appendages in the case of fribbles and iterations of the same structure for fractals; see Figure 2). Within each stimulus type, 6 items were designed to be low-complexity and 6 were designed to be high-complexity. Specifically, low-complexity fribbles are constructed of one body shape and one appendage while high-complexity fribbles have one body shape and four appendages. In the case of fractals, low-complexity items consist

of 3 iterations of the same fractal pattern whereas high-complexity items are made of 6 iterations of the same pattern. All stimuli will be openly available in the project repository on OSF.

	Fribbles	Fractals
<b>Low Complexity</b>		
<b>High Complexity</b>		

*Figure 2.* Examples of low- and high-complexity Fribble and Fractal stimuli. The full set of stimuli will be available in the project repository on OSF.

## Procedure

Full procedural instructions provided to each lab will be available in Supplementary Materials on the OSF. Before data collection begins, all labs will be asked to complete a pre-data-collection survey (which will be viewable on the OSF) and submit a walkthrough video depicting a standard testing session (described in further detail below). During the experiment, infants will be seated on their caregiver's lap or in a high-chair or car seat (whichever option corresponds to labs' standard procedures for testing infants, as reported in the pre-data-collection survey). Each trial will be made up of a familiarization phase, in which infants are first exposed to a stimulus, and two test phases where infants are presented with both the item they viewed

during the familiarization phase and a novel test item. These phases will be repeated on each of 12 total trials and are described in further detail below.

***Infant-controlled versus fixed-length methods.*** Familiarization can be achieved either by presenting a fixed amount of exposure, i.e. a *fixed-length design*, or by ending the familiarization phase of an experiment after infants have acquired a certain amount of looking at the target, i.e. an *infant-controlled design*. Both approaches have been used in the literature. In the fixed-length condition the familiarization stimulus is presented for a fixed duration regardless of how much the infant looks.<sup>61</sup> This procedure allows for variation in how much time infants look at the stimulus before their visual preference is assessed while controlling for the amount of time from the start of the familiarization phase to the test phase. Thus, in the fixed-length design, the length of the familiarization phase is fixed while the duration of visual attention directed towards the stimulus varies.

In the infant-controlled design, in contrast, the familiarization stimulus is presented until infants accumulate the same amount of looking.<sup>3,62</sup> This procedure ensures that all infants receive the same amount of familiarization before their preferences for novel stimuli are tested. It also means that the length of the familiarization phase is determined by how long it takes infants to accumulate the required amount of looking (e.g., if they have one or two long looks, with little looking away time, or multiple short looks interspersed with variable length looks away). Thus, in the infant-controlled design, the length of the familiarization phase varies while the duration of infants' visual attention directed towards the stimulus remains consistent.

The consensus across the ManyBabies 5 community was that an infant-controlled design, which facilitates equating infants' familiarization time, represented the best test of our main research questions. Participating labs will be encouraged to use and supported in their use of an

infant-controlled procedure. However, because infant-controlled procedures require either well-trained testers who can reliably code infant looking time in the moment or eye-tracking systems that can support gaze-contingent protocols, it may not be possible for some labs to use this procedure. In line with ManyBabies' philosophy of inclusive participation, we will attempt to support all labs in implementing infant-controlled familiarization, but allow for use of a fixed-length procedure in cases in which an infant-controlled procedure is not possible (i.e., labs lacking the required personnel or equipment to implement an infant-controlled procedure). This will allow labs to flexibly choose the option that works best for their resources and maximizes overall data collection, while also working towards a high proportion of labs implementing the desired procedure. To ensure that any observed differences across labs using infant-controlled and fixed-length procedures will not be due to selection effects (as labs self-selected into one paradigm or the other), a subset of labs who use the infant-controlled design will be asked to collect a second sample of infants using a fixed-length procedure. Potential effects of procedure type (fixed-length versus infant-controlled) will be examined in exploratory analyses (see S6.1 in the [Supplementary Materials](#)).

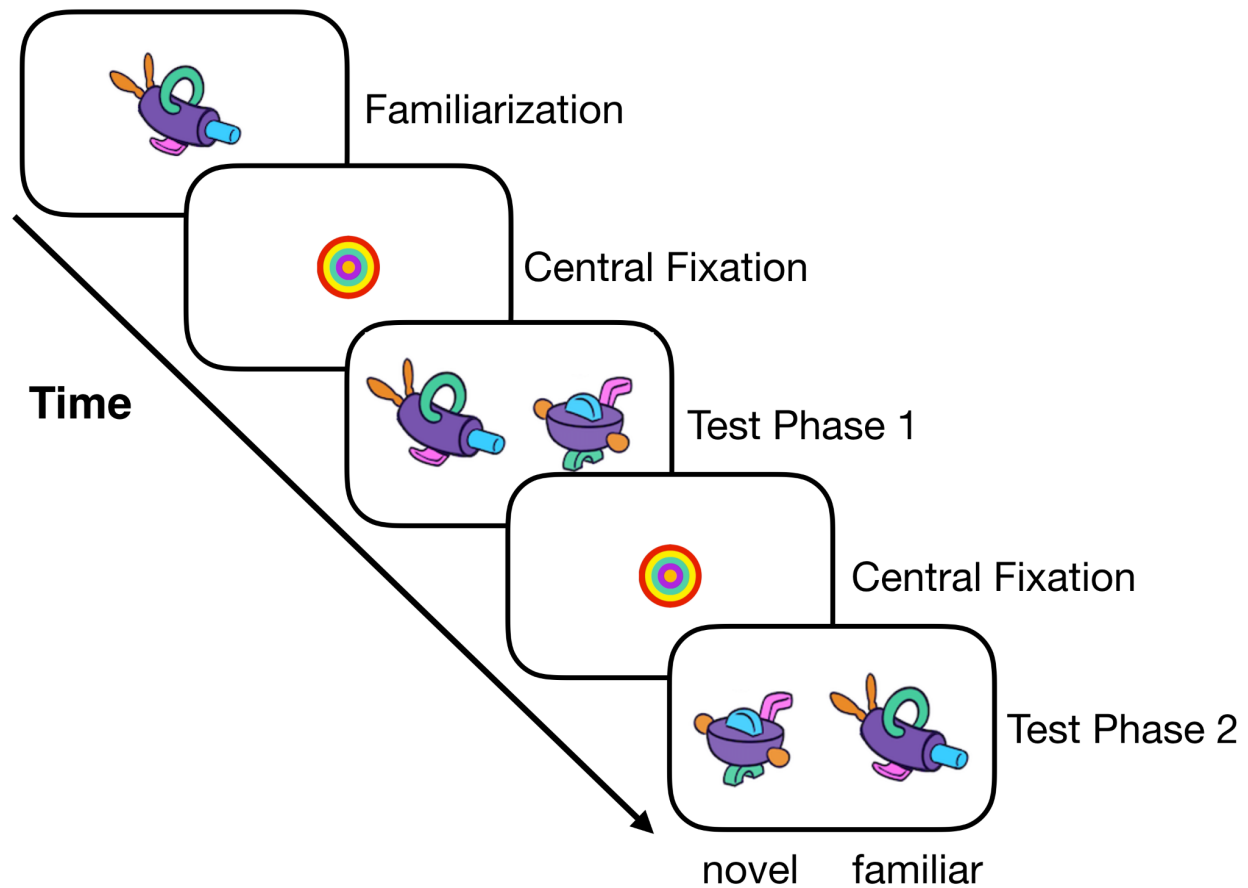
***Trial structure.*** The trial sequence is patterned after classic work by Rose and colleagues.<sup>3</sup> On each trial, a single stimulus is presented centrally on the screen for a familiarization phase, followed by two test phases in which the now-familiar stimulus is paired with a novel stimulus (the novel stimulus will be the same across the two test phases). Following Rose and colleagues,<sup>3</sup> each test phase will be 5 s in duration, and they will differ only in the left-right placement of the novel and familiar stimulus, i.e. the location of the paired test item will be switched between the two test phases.

In order to test the effect of *familiarization time*, we will vary the duration of the familiarization phase, yielding three types of trials. In the infant-controlled design, following Rose and colleagues,<sup>3</sup> the familiarized stimulus will be presented until infants accumulate 5, 10, or 15 s of looking to the familiarization stimulus (thus, the actual trial length will vary across infants depending on how long it takes each infant to accumulate the required looking time). We will additionally implement a maximum trial length criterion; if infants do not accumulate the required looking time within twice the target familiarization time (i.e., 10, 20, or 30 s), the familiarization phase will end and the next trial will begin. In the fixed-length design, the familiarized stimulus will be presented for 5, 10, or 15 s total, regardless of the duration of infants' looking.

Immediately following familiarization, a central-fixation stimulus – looming circles accompanied by chimes (as used in ManyBabies 1<sup>51</sup>) – will be presented in the center of the screen to reorient infants' gaze. In the infant-controlled design, the stimulus will play until the infant looks at the center of the screen for 500 ms; in the fixed-length design, the stimulus will play for a fixed duration of 750 ms (which is slightly longer than the infant-controlled design to allow infants time to fixate the central stimulus). The central-fixation stimulus will be immediately followed by the two test phases, with each at a fixed length of 5 s, for a total of 10 s of test.<sup>3</sup> The test phases will be a fixed duration, even if familiarization was infant-controlled. In each test phase, infants will see a familiar and novel item side by side; the left-right position of the two stimuli will be reversed in the second phase. The novel item for each test will be from a different family within the same stimulus set (i.e., fribbles or fractals) that is the same level of complexity as the one infants were familiarized with. Matching the stimuli on complexity reduces the likelihood of ceiling/floor effects and rules out that familiarity or novelty preferences

are driven by complexity differences in the paired stimuli at test, as opposed to the manipulated difference in familiarization. The central-fixation stimulus will additionally be presented between the two test phases. Figure 3 depicts the structure of an example trial.

To retain infants' attention between trials, an attention-getter will be played before the first trial and between trials. Labs will have the option of using a default attention-getter video of a laughing baby, which is effective in drawing infants' attention to the screen, but will be encouraged to use any attention-getter that is a better match for their lab or community.



*Figure 3.* This figure depicts the design of an example trial. At familiarization, a single image is shown for 5, 10, or 15 s. After the desired familiarization time is reached (based on the infant-controlled or fixed-length design criteria), infants are presented with a central fixation stimulus. During the two test phases (5 s each), infants view the same image to which they were familiarized paired with a new stimulus from the same set (e.g., fribbles or fractals) and the same

level of complexity. The side on which the familiar and novel images are presented is counterbalanced across test phases, but the images remain the same.

**Counterbalancing.** Each infant will view both stimulus sets (fribbles and fractals), with presentation blocked by stimulus set and presentation order (fribbles first or fractals first) and counterbalanced across infants. Additionally, within each stimulus set, infants will experience all combinations of complexity level and familiarization time. As a result, infants will view a total of 6 trials per stimulus set, or 12 trials total. If infants successfully view all 12 trials, this represents approximately 4.5 minutes of looking to the screen (note that the total study time will increase when the amount of time it takes for infants to accumulate each familiarization time is taken into account in the infant-controlled version).

We will additionally impose a number of constraints during counterbalancing. First, we will ensure that: (1) the familiar item does not occur on the same side in Test Phase 1 on more than two trials in a row, (2) no more than two trials in a row feature stimuli from the same level of complexity, (3) the same familiarization time is not used for more than two trials in a row. Additionally, each familiar and novel stimulus will only be used in one trial, as repeating stimuli will impact how familiar or novel a given stimulus is for the infant. Across infants, we will counterbalance which stimulus is used as the familiar or novel stimulus.

**Apparatus.** To facilitate standardization across labs, we will create infant-controlled and fixed-length versions of the experiment on a variety of software platforms (e.g., Habit, PyHab, EyeLink Experiment Builder, Tobii Studio; or, in the case of the fixed-length design, experimenters will be able to use video-presentation software<sup>63,64</sup>). Individual labs will be instructed to present the program using the setup with which they are most familiar (e.g., TV

screen, projection screen, or computer monitor). Labs will additionally choose to collect data using central fixation methods or an eye-tracker.

***Coding.*** In labs that assess looking time by coding all or part of their videos offline, coding of infants' looking to the familiar versus novel stimulus at test (for both the infant-controlled and fixed-length procedures) and during the familiarization phase (during the fixed-length procedure only) will be conducted via standard procedures in each lab. We will additionally provide a “best practices” manual that includes information about training and reliability standards that can be used by labs without existing standard procedures or those interested in revising their current procedures.

While labs will be allowed to code videos according to their existing lab practices, labs will be asked to report inter-coder reliability by independently re-coding a minimum of 20% of videos and reporting the average frame agreement, i.e., the proportion of frames on which both coders agree on the gaze location. The target reliability for this coding is 90% frame agreement.

***Minimizing bias.*** To minimize caregiver bias, each lab will be asked to ensure (using lab-standard methods) that caregivers remain unaware of the visual stimuli being presented (e.g., by asking them to wear darkened glasses, close their eyes, or keep their gaze directed toward the infant's head). To minimize experimenter bias, experimenters making online decisions about infants' looking to or away from the screen will not be aware of which visual stimulus is being presented nor the side of the familiar stimulus during the test phase. Off-line coding, including reliability coding, will be conducted with these same efforts to minimize bias. Reliability coders will also not have access to the original coder's decisions.



## General Lab Practices

Each lab will be responsible for conducting this study using the same rigorous standards applied to all other studies in their lab. Labs will be asked to provide information about their standard protocols for testing, such as researcher training practices and basic practices for greeting families, obtaining consent, and debriefing. Labs that are new to infant testing will be paired with more established laboratories for support and guidance throughout data collection.

Prior to data collection, labs will be asked to submit a walkthrough video depicting a standard testing session. This video can be of an infant participant or of a placeholder toy (e.g., a teddy bear). Before labs are invited to move on to data collection, this video will be reviewed by the ManyBabies 5 study implementation team to address any questions or inconsistencies that arise in study implementation. Labs will also be asked to upload a practice data file to ensure that each lab's data format is in compliance with the ManyBabies 5 data structure. After data collection is complete, participating investigators will submit their data to the analysis team. We will ask labs to refrain from analyzing the data independently prior to submission and analysis by the ManyBabies 5 team (with exceptions for trainees' educational timelines). Labs will be asked to provide experimental and participant data for any participant who enters the lab to participate regardless of whether the participant began or completed the study. The purpose of this request is to ensure that researchers do not selectively submit data based on their own impressions of data quality. For each participant, labs will be asked to report on any unusual events, such as equipment failure, parental interference, infant fussiness, or infants taking a break during the session. Whenever possible (i.e., the lab has appropriate permissions from an ethics board), sessions will be video recorded, and video recordings will be submitted together with coded data or uploaded to repositories such as Databrary.<sup>65</sup>

## Inclusion Criteria

All data collected for the study will be handed to the analysis team for confirmatory analyses (i.e., demographic and experimental data for every infant who entered a participating lab, regardless of how many trials they complete). While we are opting to use a maximally inclusive recruitment strategy (the only two criteria being that the infant's age falls within the 3- to 15-month age range and the infant has no known issues that would directly impede their ability to process visual stimuli), infants will be included in the primary analyses only if they meet the following additional criteria: (1) they were born at 37 or more weeks gestation (if known) and/or do not meet local definitions of preterm birth, and (2) they do not have known developmental delays. XX infants (XX% of the total sample) will be excluded for failing to meet these criteria. However, data from all infants will be retained and effects of typical exclusion criteria (e.g., preterm birth, developmental delays) will be investigated in exploratory analyses.

Additionally, infants will be excluded for session-level errors, such as experimenter error, equipment failure, or other forms of interference that impede infants' ability to attend to at least one trial in the experiment or that affect all trials ( $N = XX$ , XX% of infants remaining after recruitment-based exclusions have been implemented).

Finally, we will exclude individual trials for which issues are reported (e.g., infant fussiness, incorrect stimulus, single instance of parental interference, failure to accumulate the required familiarization time). A total of XX (XX%) trials will be excluded due to trial-level errors. To be included in the final analyses, eligible infants (as defined above) will be required to contribute nonzero looking times for at least 1 trial after trial-level exclusions are applied. XX infants (XX%) will be excluded for failing to meet this criterion. After all exclusion criteria are

applied, the study will include a final sample of XX infants (minimum sample size post-exclusions:  $N = 1,280$ ).

**Table 1.**

*Design Table.*

Question	Hypothesis	Sampling plan (e.g. power analysis)	Analysis plan	Interpretation given to different outcomes
How does familiarization time influence infants' novelty preference?	<b>(H1A)</b> As familiarization time increases, infants' novelty preference increases.	The design ( $n=1,280$ ) provides 96.6% power to detect an effect of $d=0.2$ or larger (S3 in Supplementary Materials).	We will use a linear mixed-effects model to predict infants' novelty preference (proportion looking to the novel stimulus) from familiarization time ( <b>H1A</b> ), age ( <b>H1B</b> ) and stimulus complexity ( <b>H1C</b> ). The model will include all two- and three-way interactions between the three predictors (these correspond to exploratory tests for interactions between the main predictors; S2 in Supplementary Materials). The model will include random effects for participant, item, and lab (see main text for full model specification). The three simple effects of familiarization time, age, and stimulus complexity correspond to the three main hypotheses of interest.	A significant coefficient in the predicted direction ( <b>positive</b> : familiarization time, age; <b>negative</b> : stimulus complexity) will be interpreted as evidence consistent with the corresponding hypothesis. A significant coefficient in the opposite direction will be interpreted as providing evidence in the opposing direction from that predicted by the Hunter & Ames model. In the case of non-significant coefficients, we will use equivalence tests to determine whether each (non-significant) coefficient is statistically equivalent to the absence of an effect (equivalence bounds: $[-0.2, 0.2]$ ).
How does novelty preference change as infants age?	<b>(H1B)</b> As infant age increases, infants exhibit a stronger novelty preference.	The design ( $n=1,280$ ) provides >99% power to detect an effect of $d=0.2$ or larger (S3 in Supplementary Materials).		
How is novelty preference affected by the complexity of stimuli?	<b>(H1C)</b> Novelty preference is larger when the familiarized stimulus is simpler than when it is more complex.	The design ( $n=1,280$ ) provides 98.8% power to detect an effect of $d=0.2$ or larger (S3 in Supplementary Materials).		

## Data Analysis Plan

### Dependent and Independent Variables

The analyses will include the following variables:

- **Novelty preference (model term: novelty\_preference).** Infants' preference for the novel stimulus over the familiar stimulus during the entire test phase for each trial (summed across the two test phases) will be measured as the proportion looking to the novel stimulus.<sup>3</sup>

$$\text{novelty preference} = \frac{\text{novel looking time}}{\text{novel looking time} + \text{familiar looking time}}$$

Based on past studies, we expect infants' looking times to vary widely across the full range of this dependent measure (i.e., we do not expect overall ceiling/floor effects).<sup>4,66</sup>

- **Infant age (age).** Infant age (mean-centered and scaled), will be a continuous predictor.
- **Familiarization time (familiarization\_time).** Familiarization time will be treated as a continuous predictor and centered. Specifically, the three familiarization times 5, 10, and 15 s will be coded as -0.5, 0, 0.5.
- **Stimulus complexity (stimulus\_complexity).** The two levels of stimulus complexity will be centered and coded as -0.5 (low) and 0.5 (high).
- **Participant.** The unique identifier for each individual infant participating in the study.
- **Item.** The unique familiar/novel stimulus pair in a given trial.
- **Lab.** The unique identifier for each individual lab that collected data for the study.

## Statistical Modeling Approach

**Modeling Approach.** In the main model, we will use a linear mixed-effects model to predict infants' novelty preference from age, familiarization time, and stimulus complexity, as well as all two- and three-way interactions between the three predictors. We will include random effects for participant, item, and lab. The planned model will include the maximal random effects structure,<sup>67</sup> which we will prune as necessary to allow the model to converge (see [Supplementary Materials](#) S4 for details on our planned approach for handling model non-convergence). Models

will be fit using the lme4 package in R,<sup>68,69</sup> and  $p$ -values will be estimated using the Satterthwaite approximation implemented in the R package lmerTest.<sup>70</sup>

**Model.** The main model will be specified as:

```
novelty_preference ~ 1 + age + familiarization_time + stimulus_complexity +
  age : familiarization_time +
  age : stimulus_complexity +
  familiarization_time : stimulus_complexity +
  age : familiarization_time : stimulus_complexity +
  (1 + familiarization_time * stimulus_complexity | participant) +
  (1 + age * familiarization_time * stimulus_complexity | lab) +
  (1 + age * familiarization_time | item)
```

## Hypothesis Tests

Our main hypotheses are that familiarization time, age, and stimulus complexity will each systematically predict infants' novelty preference (Table 1). In addition to the main effects of interest, the model will also allow us to test for two- and three-way interactions between the main effects of interest. These possible interaction effects are not the main focus of our confirmatory analyses, and thus we will interpret these tests as exploratory (rather than confirmatory). Table S1 in Supplementary Materials lists the question tested by the two- and three-way interactions in the model.

If any of the main hypotheses do not reach statistical significance, we will use equivalence tests with an equivalence bound determined by our smallest effect size of interest (equivalence bounds:  $[-0.2, 0.2]$ ).<sup>71,72</sup> Statistical equivalence will indicate that we can reject the hypothesis that effect is at least as large as our smallest effect size of interest ( $d = 0.2$ ) and allow

us to conclude that the effect falls within a region of practical equivalence to the absence of a meaningful effect.

### Power Analysis

We used a simulation-based approach to evaluate whether the expected sample size (i.e., 1,280 infants) can provide sufficient power to detect effect sizes of interest. All code associated with these simulations is publicly available on the project's OSF page ([https://osf.io/g3udp/?view\\_only=0f4f6e3b48d8456999d6459c0bfe5510](https://osf.io/g3udp/?view_only=0f4f6e3b48d8456999d6459c0bfe5510)), and the main simulation-based results are summarized in our Data Simulation and Power Analysis [Supplement](#).

**Power for the main hypotheses of interest.** In order to calculate the statistical power for our three main effects of interest (i.e., *stimulus type*, *age*, and *familiarization time*; Table 1), we simulated 500 datasets with a Cohen's  $d$  effect size of 0.2 as our smallest effect size of interest. We chose our smallest effect size of interest as follows. First, we selected the smallest average effect size ( $d = 0.24$ ) for single-screen preference methods in a previous large-scale preferential-looking study, ManyBabies 1<sup>51</sup>, rounding down in the interest of selecting a more conservative estimate ( $d = 0.2$ ). Given that infant studies with more than 30 infants per condition are rare<sup>47</sup>, we determined that the vast majority of infant studies would have less than 20% power to discover an effect size smaller than  $d = 0.2$  (paired-samples  $t$ -test). We therefore reasoned that it would not be practical for infant researchers to detect effect sizes smaller than  $d = 0.2$ , given current practices and resource constraints.<sup>73,74</sup> The current study represents a unique opportunity to conduct a test with a sample size large enough to estimate effect sizes that infant cognition research typically lacks the power to detect reliably.

For each of these simulated datasets, we ran a linear mixed-effects regression model with the following structure:

$$\begin{aligned} \text{novelty\_preference} \sim & 1 + \text{stimulus\_complexity} * \text{age} * \text{familiarization\_time} + \\ & (1 + \text{stimulus\_complexity} * \text{familiarization\_time} \mid \text{participant}) + \\ & (1 \mid \text{lab}) + \\ & (1 \mid \text{item}) \end{aligned}$$

Based on preliminary simulations, we chose to focus on this model because it was the maximal random effects structure that reliably converged on the vast majority of simulation runs (i.e., >90%). In our Data Simulation and Power Analysis Supplement, we also include the results from simulations for a model with a simpler random effects structure that included only random intercepts for participants, labs, and items, given past evidence that models including random slopes may not converge<sup>51</sup>. We evaluated the statistical power by counting the number of times the effect displayed a significantly non-zero coefficient (i.e.,  $p < .05$ ). The goal was to establish that we had sufficient power to detect each of the three main effects of interest (age, familiarization time, and stimulus type), i.e., to test the three main predictions of the Hunter and Ames model. The analysis indicates ample power to detect each of these three main effects for an effect size at least as large as  $d = 0.2$  (>95% power for all main effects; see Supplementary Materials S3 for further details).

In order to obtain a broader overview of the *a-priori* power of our design and to examine the sensitivity of our power analyses<sup>73</sup>, we also simulated power across a range of effects (from  $d = 0.2$  to  $d = 0.5$ ) that have been observed in previous large-scale preferential-looking studies (e.g., in ManyBabies 1, the main meta-analytic effect size was  $d = 0.35$ , with effect sizes ranging from  $d = 0.24$  to  $d = 0.51$  depending on the experimental method) as well as median effect size

estimates in meta-analyses within infant populations.<sup>45,51</sup> Overall, the power analysis indicates that the current design has sufficient power to detect each of the three main predictions of the Hunter and Ames model (see Table S2 in Supplementary Materials S3 for a full overview of power results across a range of effect sizes).

**Power for interactions.** Though the two- and three-way interactions between the three main effects are not a focal question in our main confirmatory analyses and are exploratory in nature, we also estimated our power to detect these effects given our expected sample size. The power analysis demonstrates that the design has excellent power to detect two-way interactions with age (>95% for  $d = 0.2$  or larger) and has reasonable power to detect a three-way interaction between the main effects (87.8% power for  $d = 0.2$ ; 97.2% power for  $d = 0.3$ ). Power is weakest for the (exploratory) interaction between stimulus complexity and familiarization time, but remains robust for detecting a medium-sized effect (>95% power for  $d = 0.4$  or larger).

### **Robustness and Exploratory Analyses**

The analytic strategy described above represents our main confirmatory approach, consistent with contemporary analytic approaches used in the infant literature, and our main conclusions will be based on the results of these confirmatory models. However, the large dataset collected in the current study presents a unique opportunity to explore the relation between the underlying assumptions of these models and the distributional properties of infant-looking time data. We will therefore examine the robustness of the effects of interest across various analytic decisions and inclusion criteria. The main results from the confirmatory models will be interpreted in light of the consistency or variability of the main findings across these exploratory analyses (see S5 and S6 in the [Supplementary Materials](#) for details on specific robustness and exploratory analyses).



### **Data Availability**

Full data will be publicly available at

[https://osf.io/g3udp/?view\\_only=0f4f6e3b48d8456999d6459c0bfe5510](https://osf.io/g3udp/?view_only=0f4f6e3b48d8456999d6459c0bfe5510).

### **Code Availability**

All analysis code is publicly available at

[https://osf.io/g3udp/?view\\_only=0f4f6e3b48d8456999d6459c0bfe5510](https://osf.io/g3udp/?view_only=0f4f6e3b48d8456999d6459c0bfe5510).

### **Acknowledgements**

This work was financially supported by the Early Career Award from the Einstein Foundation (awarded to Jessica E. Kosie and Martin Zettersten), a grant from the Princeton Data-Driven Social Science Initiative (awarded to Martin Zettersten, Jessica E. Kosie, and Casey Lew-Williams), a Global Collaborative Network Grant from Princeton University (awarded to Casey Lew-Williams), an SSHRC Partnership Development Grant (awarded to J. Kiley Hamlin and members of the ManyBabies Governing Board), National Science Foundation SBE 2004983 to Jessica E. Kosie, and by grants from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Numbers F32HD110174 (awarded to Martin Zettersten) and F32HD103439 (awarded to Jessica E. Kosie). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### **Author Contributions**

Authorship order was determined as follows: The first two authors are the primary co-leads of the project and share first authorship. The last author was the primary senior researcher supervising the project. All other authors are listed in alphabetical order. An overview of

authorship contributions following the CRediT taxonomy can be viewed here:

[https://docs.google.com/spreadsheets/d/e/2PACX-1vRbYFrQYIUeGdXLEOw7miYjzfNSBLR3QKKAQqU52F-J6nGCupIzgQZywez55NzaN9Vbm3cMh-XK-Fr\\_/pubhtml](https://docs.google.com/spreadsheets/d/e/2PACX-1vRbYFrQYIUeGdXLEOw7miYjzfNSBLR3QKKAQqU52F-J6nGCupIzgQZywez55NzaN9Vbm3cMh-XK-Fr_/pubhtml)

### **Competing Interests**

All authors declare no competing interests.

## References

1. Aslin, R. N. What's in a look? *Dev. Sci.* **10**, 48–53 (2007).
2. Fantz, R. L. Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science* **146**, 668–670 (1964).
3. Rose, S. A., Melloy-Carminar, P. & Bridget, W. H. Familiarity and Novelty Preferences in Infant Recognition Memory: Implications for Information Processing. *Dev. Psychol.* **18**, 704–713 (1982).
4. Rose, S. A., Feldman, J. F. & Jankowski, J. J. Infant visual recognition memory. *Dev. Rev.* **24**, 74–100 (2004).
5. Bergmann, C. & Cristia, A. Development of infants' segmentation of words from native speech: a meta-analytic approach. *Dev. Sci.* **19**, 901–917 (2016).
6. Tsuji, S. & Cristia, A. Perceptual attunement in vowels: A meta-analysis: Infant Vowel Attunement: A Meta-Analysis. *Dev. Psychobiol.* **56**, 179–191 (2014).
7. Slater, A. Visual perception and memory at birth. *Adv. Infancy Res.* **9**, 107–162 (1995).
8. Oakes, L. M. Using Habituation of Looking Time to Assess Mental Processes in Infancy. *J. Cogn. Dev.* **11**, 255–268 (2010).
9. Colombo, J. & Bundy, R. S. Infant response to auditory familiarity and novelty. *Infant Behav. Dev.* **6**, 305–311 (1983).
10. Kirkham, N. Z., Slemmer, J. A., Richardson, D. C. & Johnson, S. P. Location, Location, Location: Development of Spatiotemporal Sequence Learning in Infancy. *Child Dev.* **78**, 1559–1571 (2007).
11. Rose, S. A. & Feldman, J. F. Infant visual attention: Stability of individual differences from 6 to 8 months. *Dev. Psychol.* **23**, 490–498 (1987).

12. Damon, F., Quinn, P. C. & Pascalis, O. When novelty prevails on familiarity: Visual biases for child versus infant faces in 3.5- to 12-month-olds. *J. Exp. Child Psychol.* **210**, 105174 (2021).
13. Johnson, S. P. *et al.* Abstract Rule Learning for Visual Sequences in 8- and 11-Month-Olds. *Infancy* **14**, 2–18 (2009).
14. Lew-Williams, C. & Saffran, J. R. All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition* **122**, 241–6 (2012).
15. Marquis, A. R. & Sugden, N. A. Meta-analytic review of infants’ preferential attention to familiar and unfamiliar face types based on gender and race. *Dev. Rev.* **53**, 100868 (2019).
16. Sonne, T., Kingo, O. S. & Krøjgaard, P. 6-, 10-, and 12-month-olds remember complex dynamic events across 2 weeks. *J. Exp. Child Psychol.* **229**, 105627 (2023).
17. Bergmann, C., Rabagliati, H. & Tsuji, S. What’s in a looking time preference? Preprint at <https://doi.org/10.31234/osf.io/6u453> (2019).
18. Hamlin, J. K., Wynn, K. & Bloom, P. Social evaluation by preverbal infants. *Nature* **450**, 557–559 (2007).
19. Sommerville, J. A., Schmidt, M. F. H., Yun, J. & Burns, M. The Development of Fairness Expectations and Prosocial Behavior in the Second Year of Life. *Infancy* **18**, 40–66 (2013).
20. Ahmadlou, M. *et al.* A cell type–specific cortico-subcortical brain circuit for investigatory and novelty-seeking behavior. *Science* **372**, eabe9681 (2021).
21. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).
22. Hall, B. A., Melfi, V., Burns, A., McGill, D. M. & Doyle, R. E. Curious creatures: a multi-taxa investigation of responses to novelty in a zoo environment. *PeerJ* **6**, e4454 (2018).

23. Ruitenberg, M. F. L., Koppelmans, V., Seidler, R. D. & Schomaker, J. Novelty exposure induces stronger sensorimotor representations during a manual adaptation task. *Ann. N. Y. Acad. Sci.* **1510**, 68–78 (2022).
24. Berlyne, D. E. Curiosity and exploration. *Science* **153**, 25–33 (1966).
25. Gottlieb, J. & Oudeyer, P. Y. Towards a neuroscience of active sampling and curiosity. *Nat. Rev. Neurosci.* **19**, 758–770 (2018).
26. Kidd, C. & Hayden, B. Y. The psychology and neuroscience of curiosity. *Neuron* **88**, 449–460 (2015).
27. Twomey, K. E. & Westermann, G. Curiosity-based learning in infants: A neurocomputational approach. *Dev. Sci.* 1–13 (2017) doi:10.1111/desc.12629.
28. Hunter, M. A. & Ames, E. W. A multifactor model of infant preferences for novel and familiar stimuli. in *Advances in infancy research* vol. 5 69–95 (Ablex Publishing, 1988).
29. Hunter, M. A., Ames, E. W. & Koopman, R. Effects of Stimulus Complexity and Familiarization Time on Infant Preferences for Novel and Familiar Stimuli. *Dev. Psychol.* **19**, 338–352 (1983).
30. Wetherford, M. Developmental Changes in Infant Visual Preferences for Novelty and Familiarity. *Child Dev.* **44**, 416–424 (1973).
31. Balaban, M. T. & Waxman, S. R. Do words facilitate object categorization in 9-month-old infants? *J. Exp. Child Psychol.* **64**, 3–26 (1997).
32. Kidd, C., Piantadosi, S. T. & Aslin, R. N. The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One* **7**, e36399 (2012).
33. Thiessen, E. D., Hill, E. A. & Saffran, J. R. Infant-directed speech facilitates word

- segmentation. *Infancy* **7**, 53–71 (2005).
34. Ko, E.-S. & McDonald, M. Korean infants' perceptual responses to Korean and Western music based on musical experience. *Dev. Sci.* **26**, e13378 (2023).
  35. Bahrick, L. E. & Pickens, J. N. Infant Memory for Object Motion across a Period of Three Months: Implications for a Four-Phase Attention Function. *J. Exp. Child Psychol.* **59**, 343–371 (1995).
  36. Courage, M. L. & Howe, M. L. The Ebb and Flow of Infant Attentional Preferences: Evidence for Long-term Recognition Memory in 3-Month-Olds. *J. Exp. Child Psychol.* **70**, 26–53 (1998).
  37. Roder, B. J., Bushnell, E. W. & Sasseville, A. M. Infants' Preferences for Familiarity and Novelty during the Course of Visual Processing. *Infancy* **1**, 491–507 (2000).
  38. Thiessen, E. D. Effects of Inter- and Intra-modal Redundancy on Infants' Rule Learning. *Lang. Learn. Dev.* **8**, 197–214 (2012).
  39. Fisher-Thompson, D. Exploring the emergence of side biases and familiarity-novelty preferences from the real-time dynamics of infant looking. *Infancy* **19**, 227–261 (2014).
  40. Fisher-Thompson, D. & Peterson, J. A. Infant side biases and familiarity - Novelty preferences during a serial paired-comparison task. *Infancy* **5**, 309–340 (2004).
  41. Rabagliati, H., Ferguson, B. & Lew-Williams, C. The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Dev. Sci.* **22**, e12704 (2019).
  42. Raz, G., Cao, A., Bui, M. K., Frank, M. C. & Saxe, R. No evidence for familiarity preferences after limited exposure to visual concepts in preschoolers and infants. *Proc. Annu. Meet. Cogn. Sci. Soc.* **45**, (2023).
  43. Houston-Price, C. & Nakai, S. Distinguishing novelty and familiarity effects in infant

- preference procedures. *Infant Child Dev.* **13**, 341–348 (2004).
44. Shinskey, J. L. & Munakata, Y. Familiarity Breeds Searching: Infants Reverse Their Novelty Preferences When Reaching for Hidden Objects. *Psychol. Sci.* **16**, 596–600 (2005).
  45. Bergmann, C. *et al.* Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Dev.* **89**, 1996–2009 (2018).
  46. Byers-Heinlein, K. *et al.* Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Can. Psychol. Psychol. Can.* **61**, 349–363 (2020).
  47. Oakes, L. M. Sample Size, Statistical Power, and False Conclusions in Infant Looking-Time Research. *Infancy* **22**, 436–469 (2017).
  48. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83; discussion 83–135 (2010).
  49. Singh, L., Cristia, A., Karasik, L. B. & Oakes, L. *Diversity and Representation in Infant Research: Barriers and Bridges towards a Globalized Science*. <https://osf.io/hgukc> (2021) doi:10.31234/osf.io/hgukc.
  50. Frank, M. C. *et al.* A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy* **22**, 421–435 (2017).
  51. The ManyBabies Consortium. Quantifying sources of variability in infancy research using the infant-directed speech preference. *Adv. Methods Pract. Psychol. Sci.* **3**, 24–52 (2020).
  52. Rose, S. A., Feldman, J. F., Jankowski, J. J. & Rossem, R. Pathways From Prematurity and Infant Abilities to Later Cognition: Pathways to Later Cognition. *Child Dev.* **76**, 1172–1184 (2005).
  53. Singh, L. *et al.* A Unified Approach to Demographic Data Collection for Research with Young Children Across Diverse Cultures. *Dev. Psychol.* (2023).

54. Baumgartner, H. A. Understanding the Role of Non-Contrastive Variability in Word Learning and Visual Attention in Infancy. (University of California, Davis, 2014).
55. Deligianni, F., Senju, A., Gergely, G. & Csibra, G. Automated gaze-contingent objects elicit orientation following in 8-month-old infants. *Dev. Psychol.* **47**, 1499–503 (2011).
56. Ellis, E. M., Borovsky, A., Elman, J. L. & Evans, J. L. Novel word learning: An eye-tracking study. Are 18-month-old late talkers really different from their typical peers? *J. Commun. Disord.* **58**, 143–157 (2015).
57. Ellis, E. M., Borovsky, A., Elman, J. L. & Evans, J. L. Toddlers' Ability to Leverage Statistical Information to Support Word Learning. *Front. Psychol.* **12**, 600694 (2021).
58. Williams, P. & Simons, D. J. Detecting Changes in Novel, Complex Three-dimensional Objects. *Vis. Cogn.* **7**, 297–322 (2000).
59. Ellis, C. T. *et al.* Evidence of hippocampal learning in human infants. *Curr. Biol.* **31**, 3358–3364.e4 (2021).
60. Gustafsson, E., Francoeur, C., Blanchette, I. & Sirois, S. Visual exploration in adults: Habituation, mere exposure, or optimal level of arousal? *Learn. Behav.* **50**, 233–241 (2022).
61. Fagan, J. F. Infant Recognition Memory: The Effects of Length of Familiarization and Type of Discrimination Task. *Child Dev.* **45**, 351–356 (1974).
62. Rose, S. A. Differential Rates of Visual Information Processing in Full-Term and Preterm Infants. *Child Dev.* **54**, 1189–1198 (1983).
63. Kominsky, J. F. PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behav. Dev.* **54**, 114–119 (2019).
64. Oakes, L. M., Sperka, D., DeBolt, M. C. & Cantrell, L. M. Habit2: A stand-alone software solution for presenting stimuli and recording infant looking times in order to study infant



- development. *Behav. Res. Methods* **51**, 1943–1952 (2019).
65. Gilmore, R. O., Adolph, K. E. & Millman, D. S. Curating identifiable data for sharing: The databrary project. in *2016 New York Scientific Data Summit (NYSDS)* 1–6 (2016).  
doi:10.1109/NYSDS.2016.7747817.
  66. Wang, Y., Seidl, A. & Cristia, A. Infant speech perception and cognitive skills as predictors of later vocabulary. *Infant Behav. Dev.* **62**, 101524 (2021).
  67. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).
  68. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 1–48 (2015).
  69. R Development Core Team. R: A language and environment for statistical computing. (2022).
  70. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in linear mixed effects models. *J. Stat. Softw.* **82**, (2017).
  71. Lüdtke, D., Ben-Shachar, M., Patil, I. & Makowski, D. Extracting, Computing and Exploring the Parameters of Statistical Models using R. *J. Open Source Softw.* **5**, 2445 (2020).
  72. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence Testing for Psychological Research: A Tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
  73. Lakens, D. Sample Size Justification. *Collabra Psychol.* **8**, 33267 (2022).
  74. Simonsohn, U. Small telescopes: Detectability and the evaluation of replication results. *Psychol. Sci.* **26**, 559–569 (2015).

## Supplementary Materials

### S1. Overview of Contributing Labs and Researchers



*Figure S1.* Map depicting the geographic locations of the ManyBabies 5 contributors.

## S2. Supplement to the Data Analysis Plan: Interpretation of Interaction Effects

In Table S1, we list the questions tested by each of the two- and three-way interactions included in the main model.

**Table S1**

*Interactions included in the linear mixed-effects model.*

Question	Model term
To what extent does the effect of familiarization time on novelty preference change with infant age?	age * familiarization_time
To what extent does the effect of stimulus complexity on novelty preference change with infant age?	age * stimulus_complexity
To what extent does the effect of stimulus complexity on novelty preference change as familiarization time increases?	familiarization_time * stimulus_complexity
To what extent does the interaction between familiarization time and complexity of the stimulus depend on infant age?	age * familiarization_time * stimulus_complexity

## S3. Overview of Power Analysis Simulation Results

In Table S2, we provide an overview of the power simulation results for a wider set of effect sizes ( $d = 0.2, 0.3, 0.4, 0.5$ ), in order to assess the sensitivity of our power to detect a range of plausible effect sizes. For each effect size, we estimated power using a linear mixed-effects regression model with the following structure:

$$\begin{aligned}
 \text{novelty\_preference} \sim & 1 + \text{stimulus\_complexity} * \text{age} * \text{familiarization\_time} + \\
 & (1 + \text{stimulus\_complexity} * \text{familiarization\_time} \mid \text{participant}) + \\
 & (1 \mid \text{lab}) + \\
 & (1 \mid \text{item})
 \end{aligned}$$

**Table S2.**

Overview of results in power simulation for the model with the maximal converging random effects structure. The numbers in each effect size column indicate the percentage of significant results in the 500 simulated datasets. *stim\_type* refers to stimulus type, *fam\_time* refers to familiarization time, and *age* refers to infant age in months. Highlighted columns indicate the primary power analysis for the main effects of interest. Sample size is 1,280 included participants, with 12 trials. See section 3 in the Data Simulation and Power Analysis [Supplement](#) for details.

Outcome Types	Model Terms	<i>d</i> = 0.2	<i>d</i> = 0.3	<i>d</i> = 0.4	<i>d</i> = 0.5
<i>main effects</i>	<i>stim_type</i>	98.8%	100%	100%	100%
	<i>age</i>	100%	100%	100%	100%
	<i>fam_time</i>	96.6%	100%	100%	100%
<i>two-way interactions</i>	<i>stim_type*fam_time</i>	60.0%	82.4%	96.8%	100%
	<i>stim_type*age</i>	99.4%	100%	100%	100%
	<i>fam_time*age</i>	97.6%	99.8%	100%	100%
<i>three-way interaction</i>	<i>age*stim_type*fam_time</i>	87.8%	97.2%	100%	100%

#### **S4. Model Non-Convergence and Pruning Random Effects**

If the maximal model does not converge using default settings in the lme4 package, our first step will be to attempt a series of remedies to help the model including the full random effects structure achieve convergence (Brauer & Curtin, 2018). Specifically, we will (a) increase the number of iterations in the estimation procedure, (b) provide the model with improved starting values, and (c) check whether convergence is achieved using any of the other optimization functions available with the lme4 package.

If the model does not converge after attempting these remedies, we will next prune random effects until convergence is achieved. We will first remove random effects that are of lesser theoretical interest before removing random effects that are more crucial to the main hypotheses of the study (i.e., random slopes of critical main effects of interest and random intercepts). In practice, we expect that models including random slopes (although consistent with the experiment design) will be difficult to fit and will require further pruning of the random effects structure. For example, in ManyBabies 1 (ManyBabies Consortium, 2020), the maximal random effects structure that allowed the main model to converge included random intercepts only.

We will proceed as follows in pruning random effects step-by-step. In each instance, we would only proceed to the next pruning step if the model still fails to converge. Each of steps 1-3 will be repeated within each of the three types of random effects (item, lab, participant) separately until the model with the overall maximal random effects structure across the three random effect types is achieved. We will first prune random slopes (following steps 1-3) for item-based random effects, followed by random slopes for lab-based random effects. Random slopes and covariances for participant-based random effects will be pruned last.

1. We will first sequentially remove **random slopes for interaction terms** since the interactions are of lesser theoretical interest (compared to the main effects), checking for non-convergence after each interaction term is removed. Random slopes for the three-way interaction will be removed first, followed by the interaction for stimulus complexity and age, stimulus complexity and familiarization time, and finally, familiarization time and age.
2. We will next remove **any covariances between random effects that approach 0 or 1** (likely leading to a singular fit in the main linear mixed-effects model).
3. If convergence is still not achieved, we will successively remove **random slopes for each of the main effects** of interest. We will remove random slopes in the following order, checking for convergence after each random slope is removed: (a) stimulus complexity, (b) infant age, and (c) familiarization time.

Finally, if the model still fails to converge after removing all random slopes for item, lab, and participant, we will successively remove **random intercepts**, first removing random intercepts for item, followed by random intercepts for lab and (if necessary) random intercepts for participant last.

## **S5. Analysis of Model Assumptions and Robustness**

In this section, we outline a series of analyses to test for violations of model assumptions and to investigate the robustness of the results from the main model across alternative model specifications.

### **S5.1. Model Diagnostics**

First, we will conduct diagnostic tests to inspect two key properties of the data and model: missingness and heteroskedasticity.

### S5.1.1. Missingness

We expect there to be a significant amount of missing data due to a variety of infant-related (e.g., “fussiness”) and experimenter-related (e.g., technical errors) factors. A relatively large proportion of missing data is common in infant research (e.g., around 20% of the data in ManyBabies1; The ManyBabies Consortium, 2020). Taking missingness into account is particularly important in the context of the current study because the main phenomenon of interest — infants’ visual attention — is systematically connected with a possible major source of missingness, namely infant inattention or lack of engagement leading to trial exclusion or failure to measure infant looking behavior. A possible risk is that our estimates of infants’ patterns of novelty preference may be biased. Therefore, we will investigate whether and how different factors influence the presence of missingness in the data. Specifically, we will fit a logistic mixed-effects model predicting the trial-by-trial presence or absence of the data (i.e., coded as 0 = data is present; 1 = data is missing) from several key predictors: the main predictors of interest (age, complexity, familiarization time) and trial number (the strongest predictor of missingness in ManyBabies1). We will include random effects for participant, lab, and item, following the same approach for specifying and pruning random effects as in the main model. The model will be specified as:

$$\text{NAs} \sim \text{familiarization\_time} + \text{age} + \text{stimulus\_complexity} + \text{trial\_number} + \\ (\dots | \text{participant}) + (\dots | \text{lab}) + (\dots | \text{item})$$

In exploratory analyses, we will also test a more complex model including a series of additional predictors of interest (e.g., time of testing, time of last nap). If we find evidence of systematicity in missingness, we will explore routes for handling this systematicity, in particular conducting

multiple imputation, and discuss limitations in the interpretation of the results from the main model in light of missingness issues.

### **S5.1.2. Heteroskedasticity**

A particular concern in evaluating our main model is the possibility of heteroskedasticity. Specifically, it is likely that variance in looking times will increase across trials, which may in turn lead to higher variance in our measures of preferential looking at later trials in the experiment. To check for issues related to heteroskedasticity, we will use the performance package in R (Lüdtke et al., 2021) to conduct a Breusch-Pagan test (Breusch & Pagan, 1979) of non-constant error. If we find evidence of heteroskedasticity, we will explore the use of location-scale modeling and general additive modeling approaches to model and account for sources of variability (e.g., trial number).

### **S5.2. Alternative Link Functions**

We chose the primary dependent variable, infants' novelty preference score, based on its prevalence in the literature on infant preferential looking (e.g., Rose et al., 1982). However, because this dependent measure is a proportion variable, it raises several potential analytical issues. For example, one possible concern is that the residuals (i.e., the error variance) from the model may not be normally distributed, violating an assumption of linear mixed-effects models. Therefore, in the event that the observed data do violate this assumption, we will explore alternative models that incorporate specialized link functions (i.e., beta regression) that may be better suited for the properties of proportion score data. We will also investigate the use of alternative link functions if we encounter any issues related to a restricted range of the dependent measure.



## S6. Exploratory Analyses

Below, we outline a series of analyses we plan to conduct to explore other possible sources of variation in infant-looking behavior and the impact of specific methodological choices (e.g., inclusion criteria, stimulus choices).

### S6.1. Infant-Controlled vs. Fixed-Length Familiarization

We will test for a moderating effect of familiarization method (infant-controlled vs. fixed-length familiarization) (a) in the overall dataset and (b) in the subsample of the data contributed by labs randomly assigned to collect both infant-controlled and fixed-length samples. Specifically, we will fit the 4-way interaction model predicting novelty preference from stimulus type, age, familiarization time, and familiarization method.

novelty\_preference ~

$$1 + \text{age} * \text{familiarization\_time} * \text{stimulus\_complexity} * \text{familiarization\_method} +$$

$$(1 + \text{familiarization\_time} * \text{stimulus\_complexity} \mid \text{participant}) +$$

$$(1 + \text{age} * \text{familiarization\_time} * \text{stimulus\_complexity} * \text{familiarization\_method} \mid \text{lab}) +$$

$$(1 + \text{age} * \text{familiarization\_time} * \text{familiarization\_method} \mid \text{item})$$

If familiarization method exerts a moderating influence on any of the main predictors of interest, we will interpret its implications in the General Discussion.

### S6.2. Stimulus Fixation Time vs. Familiarization Manipulation

In the main confirmatory model, we do not distinguish between infants' accumulated looking to the familiarized stimulus and the duration of the familiarization phase. Since labs can choose to either use an infant-controlled or a fixed familiarization design, this will allow us to assess each of these predictors independently. We will therefore fit a linear mixed-effects model in which we predict novelty preference from the duration of the familiarization phase

(familiarization\_duration) and infants' time fixating the stimulus during the familiarization phase (stimulus\_fixation) while controlling for age and stimulus complexity.

novelty\_preference ~ 1 +

age + familiarization\_duration + stimulus\_fixation + stimulus\_complexity +

(1 + familiarization\_duration + stimulus\_fixation + stimulus\_complexity | participant) +

(1 + age + familiarization\_duration + stimulus\_fixation + stimulus\_complexity | lab) +

(1 + age + familiarization\_duration + stimulus\_fixation | item)

We expect that familiarization duration and stimulus fixation will be correlated. We will therefore run diagnostic tests to check for issues related to multicollinearity (specifically, we will test for variance inflation). If the variance inflation factor indicates a concerning degree of multicollinearity (i.e., exceeds a value of at least 5), we will treat any results from the model with caution. If the model including both factors is high, we will instead consider a model comparisons approach in which we fit two models, one including only familiarization duration (in addition to the other main effects of age and stimulus complexity) and one with only stimulus fixation. This will allow us to investigate how well each of the two possible operationalizations of familiarization time predict preferential looking.

### **S6.3. Investigating Stimulus Types Separately**

One possible concern with the analysis outlined above is that there may be fundamental differences in how infants process the two types of stimuli, fribbles and fractals (and, e.g., their associated levels of relative complexity). In particular, fitting a model including both item kinds may lead to poor model fit if the estimates for the two item types systematically differ. To check for inconsistencies in the model predictions across the two stimulus types, we will therefore fit a linear mixed-effects model in which we predict novelty preference from the 4-way interaction

between the main effects of interest (age, familiarization time, and stimulus complexity) and stimulus type (centered; coded as fribbles = 0.5 vs. fractals = -0.5):

novelty\_preference  $\sim 1 +$

age \* familiarization\_time \* stimulus\_complexity \* stimulus\_type +  
 (1 + familiarization\_time \* stimulus\_complexity \* stimulus\_type | participant) +  
 (1 + age \* familiarization\_time \* stimulus\_complexity \* stimulus\_type | lab) +  
 (1 + age \* familiarization\_time | item)

We will follow the same random effects pruning approach as with the main model. Effects of the model – especially interactions between stimulus type and the effects of age, familiarization time, and stimulus complexity – will be qualified in light of any interactions observed with stimulus type.

#### S6.4. Inclusion Criteria

We will also explore the consequences of varying the inclusion criteria on the main effects of interest, in particular:

- **Minimum looking times:** We will also fit the main model at various cutoffs for minimum looking times. Specifically, we will fit the main model after excluding any trials that do not meet successively more restrictive inclusion criteria for a given trial:
  - Trials must include 1 second of looking in each of the two 5s test periods.
  - Trials must include 2 seconds of looking in each of the two 5s test periods.
- **Number of trials completed:** We will fit the main model with successively more strict inclusion criteria based on trial number. We will fit the model when only including infants who complete at least 2, 4, 6, 8, 10, and 12 trials and discuss any observed systematic changes in effect magnitude.

### S6.5. Variation across Demographic Predictors

We will explore the degree to which preferential-looking behavior varies across a range of demographic factors and participant characteristics that will be collected along with infants' looking behavior, including variation across testing location, cultural and linguistic experiences, and the backgrounds of infants and their families. We do not have specific hypotheses about cultural differences, given that a very restricted set of populations from outside the United States and Western Europe have been investigated previously, often with small sample sizes (Singh et al., 2023). However, these exploratory analyses will serve to identify possible sources of variation in infant looking behavior across a broad range of characteristics that could inform future confirmatory investigations.

### References

- Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, 23(3), 389–411. <https://doi.org/10.1037/met0000159>
- Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47(5), 1287. <https://doi.org/10.2307/1911963>
- Lüdtke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Rose, S. A., Melloy-Carminar, P., & Bridget, W. H. (1982). Familiarity and Novelty Preferences in Infant Recognition Memory: Implications for Information Processing. *Developmental Psychology*, 18(5), 704–713.

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. M. (2023). Diversity and representation in infant research: Barriers and bridges toward a globalized science of infant development. *Infancy*, 28(4), 708–737. <https://doi.org/10.1111/infa.12545>

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

# ManyBabies 5: The Hunter and Ames Model of Infant Looking Preference

## Supplementary Materials: Data Simulation and Power Analysis

ManyBabies Analysis Team

## Contents

<b>1</b>	<b>Data Simulation</b>	<b>2</b>
<b>2</b>	<b>Visualisation of Simulated Data</b>	<b>4</b>
2.1	Familiarization: . . . . .	4
2.2	Complexity: . . . . .	5
2.3	Age: . . . . .	6
2.4	Age*Familiarization: . . . . .	7
2.5	Age*Complexity: . . . . .	8
2.6	Familiarization*Complexity: . . . . .	9
2.7	Variability by Lab . . . . .	11
2.8	Variability by Item . . . . .	12
<b>3</b>	<b>Overview of Power Simulation Results</b>	<b>13</b>
3.1	Summary Statistics for Power Calculation with Full Data and Varying Intercepts and Varying Slopes: . . . . .	13
3.2	Summary Statistics for Power Calculation with Full Data and Varying Intercepts: . . . . .	13
3.3	Summary Statistics for Power Calculation with 20 pct. Missing Data and Varying Intercepts: . . . . .	13
3.4	Summary Statistics for Power Calculation with 50 pct. Missing Data and Varying Intercepts: . . . . .	13
<b>4</b>	<b>Overview of Bias Results</b>	<b>14</b>
<b>5</b>	<b>Grid Search Code</b>	<b>14</b>
<b>6</b>	<b>Power Calculation with Full Data and Varying Intercepts and Varying Slopes</b>	<b>20</b>
6.1	Effect Size = 0.5 . . . . .	20
6.2	Effect Size = 0.4 . . . . .	21
6.3	Effect Size = 0.3 . . . . .	21
6.3.1	Visualise Estimates for Fixed Effects: . . . . .	21
6.3.2	Visualise Estimates for Random Effects: . . . . .	23
6.4	Effect Size = 0.2 . . . . .	24
6.5	Effect Size = 0.1 . . . . .	24

<b>7</b>	<b>Power Calculation with Full Data and Varying Intercepts</b>	<b>24</b>
7.1	Effect Size = 0.5 . . . . .	24
7.2	Effect Size = 0.4 . . . . .	25
7.3	Effect Size = 0.3 . . . . .	25
7.3.1	Visualise Estimates for Fixed Effects: . . . . .	25
7.3.2	Visualise Estimates for Random Effects: . . . . .	26
7.4	Effect Size = 0.2 . . . . .	27
7.5	Effect Size = 0.1 . . . . .	27
<b>8</b>	<b>Power Calculation with 20 pct. Missing Data and Varying Intercepts</b>	<b>28</b>
8.1	Effect Size = 0.5 . . . . .	28
8.2	Effect Size = 0.4 . . . . .	28
8.3	Effect Size = 0.3 . . . . .	29
8.3.1	Visualise Estimates for Fixed Effects: . . . . .	29
8.3.2	Visualise Estimates for Random Effects: . . . . .	29
8.4	Effect Size = 0.2 . . . . .	29
8.5	Effect Size = 0.1 . . . . .	30
<b>9</b>	<b>Power Calculation with 50 pct. Missing Data and Varying Intercepts</b>	<b>30</b>
9.1	Effect Size = 0.5 . . . . .	30
9.2	Effect Size = 0.4 . . . . .	31
9.3	Effect Size = 0.3 . . . . .	31
9.3.1	Visualise Estimates for Fixed Effects: . . . . .	31
9.3.2	Visualise Estimates for Random Effects: . . . . .	31
9.4	Effect Size = 0.2 . . . . .	32
9.5	Effect Size = 0.1 . . . . .	32

# 1 Data Simulation

```
my_sim_data <- function(  
  n_subj      = 1280,  # number of subjects  
  n_simple   = 6,     # number of complex stimuli  
  n_complex  = 6,     # number of complex stimuli  
  n_small_fam = 4,     #small familiarization time  
  n_medium_fam = 4,   #medium familiarization time  
  n_high_fam  = 4,    #high familiarization time  
  n_lab      = 40,  
  
  beta_0 = 0, # intercept; i.e., the grand mean  
  beta_c = 0.3, # main effect for complexity  
  beta_f = 0.3, # main effect for familiarization time  
  beta_a = 0.3, # main effect for age  
  
  beta_ca = 0.3,  
  beta_af = 0.3,  
  beta_cf = 0.3,  
  
  beta_cfa = 0.3, #main effect for interaction between complexity and familiarization.  
  
  subject_0 = 0.2, # by-subject random intercept sd  
  
  subject_c = 0.2, # by-subject slope complexity sd  
  subject_f = 0.2, # by-subject slope familiarization sd  
  subject_a = 0.2, # by-subject slope age sd  
  
  subject_ca = 0.2, # by-subject slope for interaction between age and complexity sd  
  subject_af = 0.2, # by-subject slope for interaction between age and familiarization sd  
  subject_cf = 0.2, # by-subject slope complexity*familiarization sd  
  
  subject_cfa = 0.2, # by-subject slope for interaction between age, complexity and familiarization sd  
  
  subj_rho = .2, # correlations between by-subject random effects  
  
  lab_0 = 0.2, # by-lab random intercept sd  
  
  lab_c = 0.2, # by-lab slope complexity sd  
  lab_f = 0.2, # by-lab slope familiarization sd  
  lab_a = 0.2, # by-lab slope age sd  
  
  lab_ca = 0.2, # by-lab slope for interaction between age and complexity sd  
  lab_af = 0.2, # by-lab slope for interaction between age and familiarization sd  
  lab_cf = 0.2, # by-lab random slope complexity*familiarization sd  
  
  lab_cfa = 0.2, # by-lab slope for interaction between age, complexity and familiarization sd  
  
  lab_rho = 0.2, # correlations between by-lab random effects  
  
  item_0 = 0.05, # by-item random intercept sd  
  
  item_c = 0.05, # by-item slope complexity sd  
  item_f = 0.05, # by-item slope familiarization sd  
  item_a = 0.05, # by-item slope age sd  
  
  item_ca = 0.05, # by-item slope for interaction between age and complexity sd  
  item_af = 0.05, # by-item slope for interaction between age and familiarization sd  
  item_cf = 0.05, # by-item random slope complexity*familiarization sd  
  
  item_cfa = 0.05, # by-item slope for interaction between age, complexity and familiarization sd
```



```

item_rho = 0.2, # correlations between by-item random effects

sigma = 0.3 # residual (error) sd
) { # residual (standard deviation)

# simulate a sample of items
items <- data.frame(
  category = rep(c("simple", "complex"), c(n_simple, n_complex)),
  X_c = rep(c(-0.5, 0.5), c(n_simple, n_complex)),
  familiarization = rep(c("short", "medium", "long"), (n_simple + n_complex)/3),
  X_f = rep(c(-0.5, 0, 0.5), (n_simple + n_complex)/3),
  faux::rnorm_multi(
    n = n_simple + n_complex, mu = 0, sd = c(item_0,
                                              item_c,
                                              item_f,
                                              item_a,
                                              item_ca,
                                              item_af,
                                              item_cf,
                                              item_cfa), r = item_rho,
    varnames = c("I_0", "I_c", "I_f", "I_a",
                  "I_ca", "I_af", "I_cf",
                  "I_cfa"))
) %>%
mutate(item_id = faux::make_id(nrow(.), "I"))

# simulate a sample of subjects
subjects <-
  faux::rnorm_multi(
    n = n_subj, mu = 0, sd = c(subject_0,
                                subject_c,
                                subject_f,
                                subject_a,
                                subject_ca,
                                subject_af,
                                subject_cf,
                                subject_cfa), r = subj_rho,
    varnames = c("S_0", "S_c", "S_f", "S_a",
                  "S_ca", "S_af", "S_cf",
                  "S_cfa")
) %>%
  mutate(subj_id = faux::make_id(nrow(.), "S")) %>%
  mutate(X_a = runif(n_subj, min = -0.5, max = 0.5))
#add subject age measure, sample from distribution from -0.5 to 0.5. #subjects$subj_id <- 1:n_subj

labs <- faux::rnorm_multi(
  n = n_lab, mu = 0, sd = c(lab_0, lab_c, lab_f, lab_a,
                            lab_ca, lab_af, lab_cf,
                            lab_cfa), r = lab_rho,
  varnames = c("L_0", "L_c", "L_f", "L_a",
                "L_ca", "L_af", "L_cf",
                "L_cfa")
) %>%
  mutate(lab_id = faux::make_id(nrow(.), "L"))

#create lab and subj nesting structure
#Number of subjects must be a multiple of number of labs
lab_multiplier = n_subj/n_lab
lab_subj_dict <- data.frame(
  subj_id = subjects$subj_id,

```

```

  lab_id = rep(labs$lab_id, lab_multiplier)
)

# cross subject and item IDs
temp <- crossing(subjects, items) %>%
  left_join(lab_subj_dict, by = "subj_id") %>%
  left_join(labs, by = "lab_id") %>%
  group_by(subj_id, item_id) %>%
  mutate(item_id = sample(item_id)) %>%
  ungroup() %>%
  mutate(trial_num = rep(seq(n_simple + n_complex), n_subj))

temp <- temp %>%
  mutate(
    B_0 = beta_0 + S_0 + L_0 + I_0,

    B_c = beta_c + S_c + L_c + I_c,
    B_f = beta_f + S_f + L_f + I_f,
    B_a = beta_a + S_a + L_a + I_a,

    B_ca = beta_ca + S_ca + L_ca + I_ca,
    B_af = beta_af + S_af + L_af + I_af,
    B_cf = beta_cf + S_cf + L_cf + I_cf,

    B_cfa = beta_cfa + S_cfa + L_cfa + I_cfa,

    e_si = rnorm(nrow(temp), mean = 0, sd = sigma),

    DV = B_0 +
      (B_a * X_a) + (B_c * X_c) + (B_f * X_f) +
      (B_cf * X_c * X_f) + (B_af * X_a * X_f) + (B_ca * X_c * X_a) +
      (B_cfa * X_c * X_f * X_a) + e_si
  )
}

dat_sim <- my_sim_data()

```

## 2 Visualisation of Simulated Data

### 2.1 Familiarization:

```

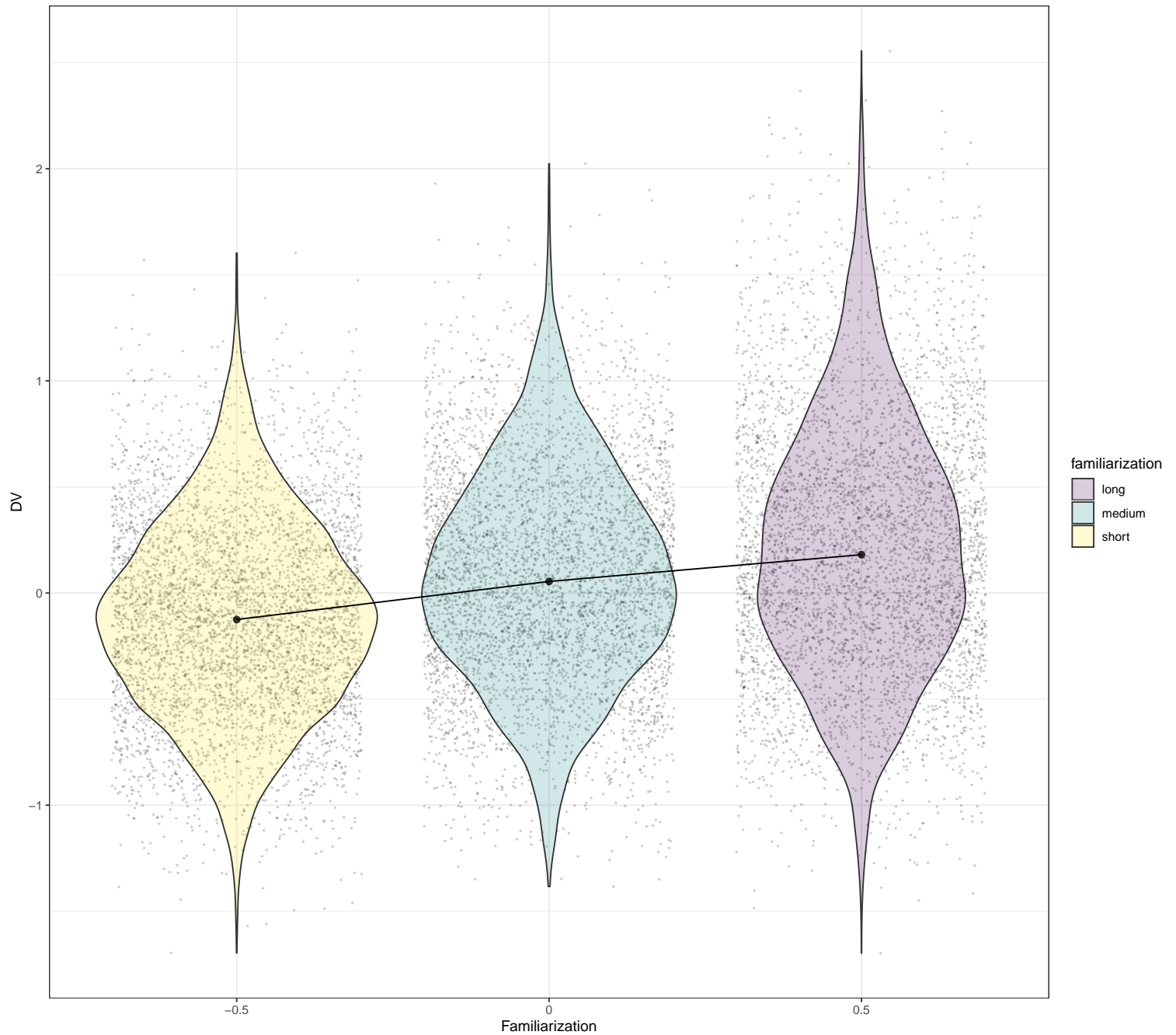
dat_sim_plot_familiarization <- dat_sim %>%
  group_by(X_f) %>%
  dplyr::summarise(med_DV = median(DV))

plot_familiarization <- dat_sim %>%
  mutate(X_f = as.factor(X_f)) %>%
  ggplot() + geom_point(aes(y = DV, x = X_f), position = "jitter",
    alpha = 0.2, size = 0.2) + geom_violin(aes(y = DV, x = X_f,
    fill = familiarization), alpha = 0.2) + geom_line(aes(y = med_DV,
    x = as.factor(X_f), group = 1), data = dat_sim_plot_familiarization) +
  geom_point(aes(y = med_DV, x = as.factor(X_f)), alpha = 0.8,
    size = 2, data = dat_sim_plot_familiarization) + scale_fill_manual(values = viridis(n = 3)) +
  ggtitle("Familiarization") + xlab("Familiarization") + theme_bw()

plot_familiarization <- plot_familiarization + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
plot_familiarization

```

## Familiarization



## 2.2 Complexity:

```
dat_sim_plot_complexity <- dat_sim %>%
  group_by(X_c) %>%
  dplyr::summarise(med_DV = median(DV))

plot_complexity <- dat_sim %>%
  mutate(X_c = as.factor(X_c)) %>%
  ggplot() + geom_point(aes(y = DV, x = X_c), position = "jitter",
    alpha = 0.2, size = 0.2) + geom_violin(aes(y = DV, x = X_c,
    fill = category), alpha = 0.2) + geom_line(aes(y = med_DV,
    x = as.factor(X_c), group = 1), data = dat_sim_plot_complexity) +
  geom_point(aes(y = med_DV, x = as.factor(X_c)), alpha = 0.8,
    size = 2, data = dat_sim_plot_complexity) + scale_fill_manual(values = viridis(n = 2)) +
  ggtitle("Stimulus Complexity") + xlab("Stimulus Complexity") +
  theme_bw()
```

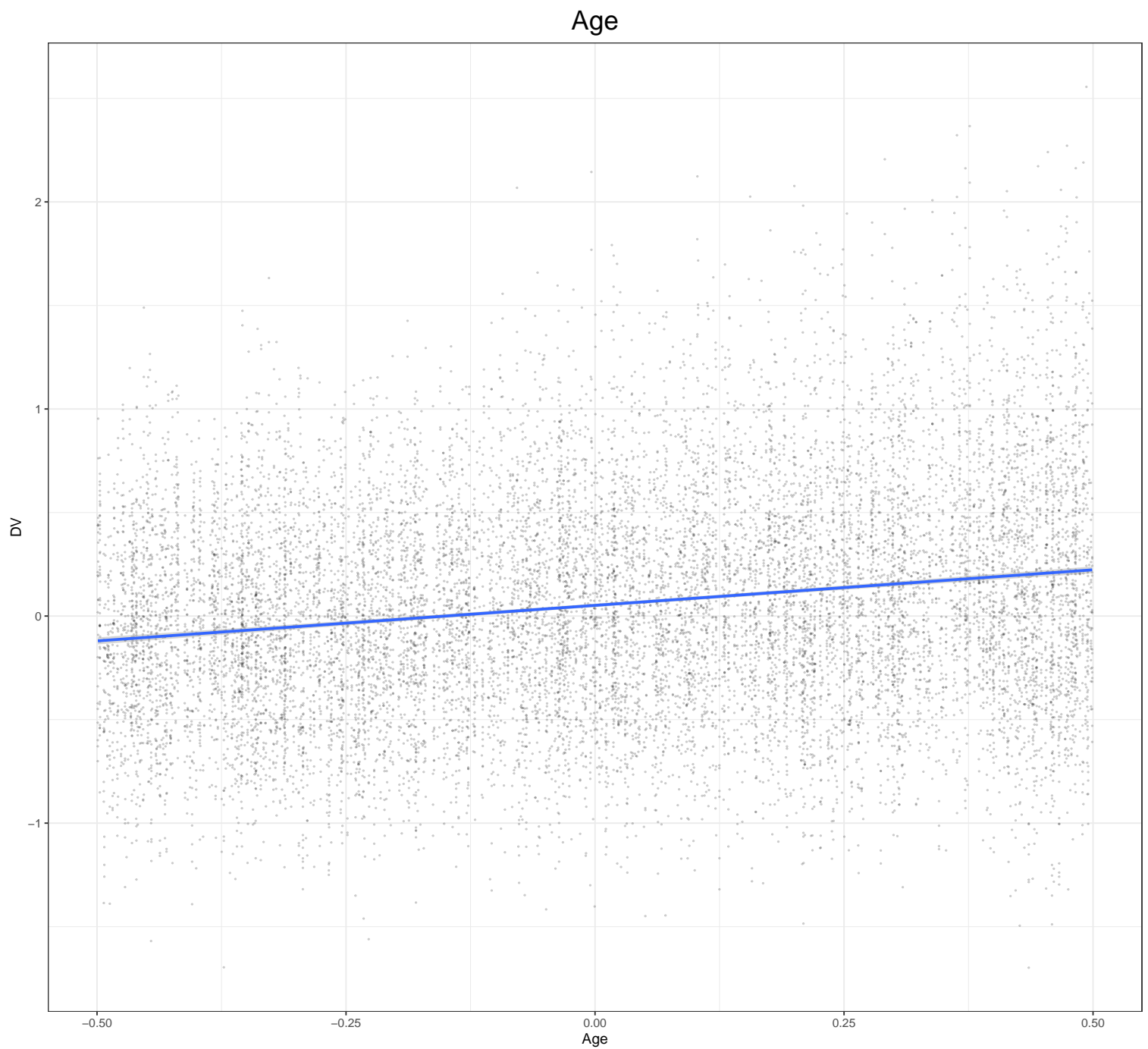
```
plot_complexity <- plot_complexity + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
plot_complexity
```



## 2.3 Age:

```
plot_age <- dat_sim %>%
  ggplot() + geom_point(aes(y = DV, x = X_a), position = "jitter",
    alpha = 0.2, size = 0.2) + geom_smooth(method = "lm", se = TRUE,
    formula = y ~ x, aes(y = DV, x = X_a)) + ggtitle("Age") +
    xlab("Age") + theme_bw()

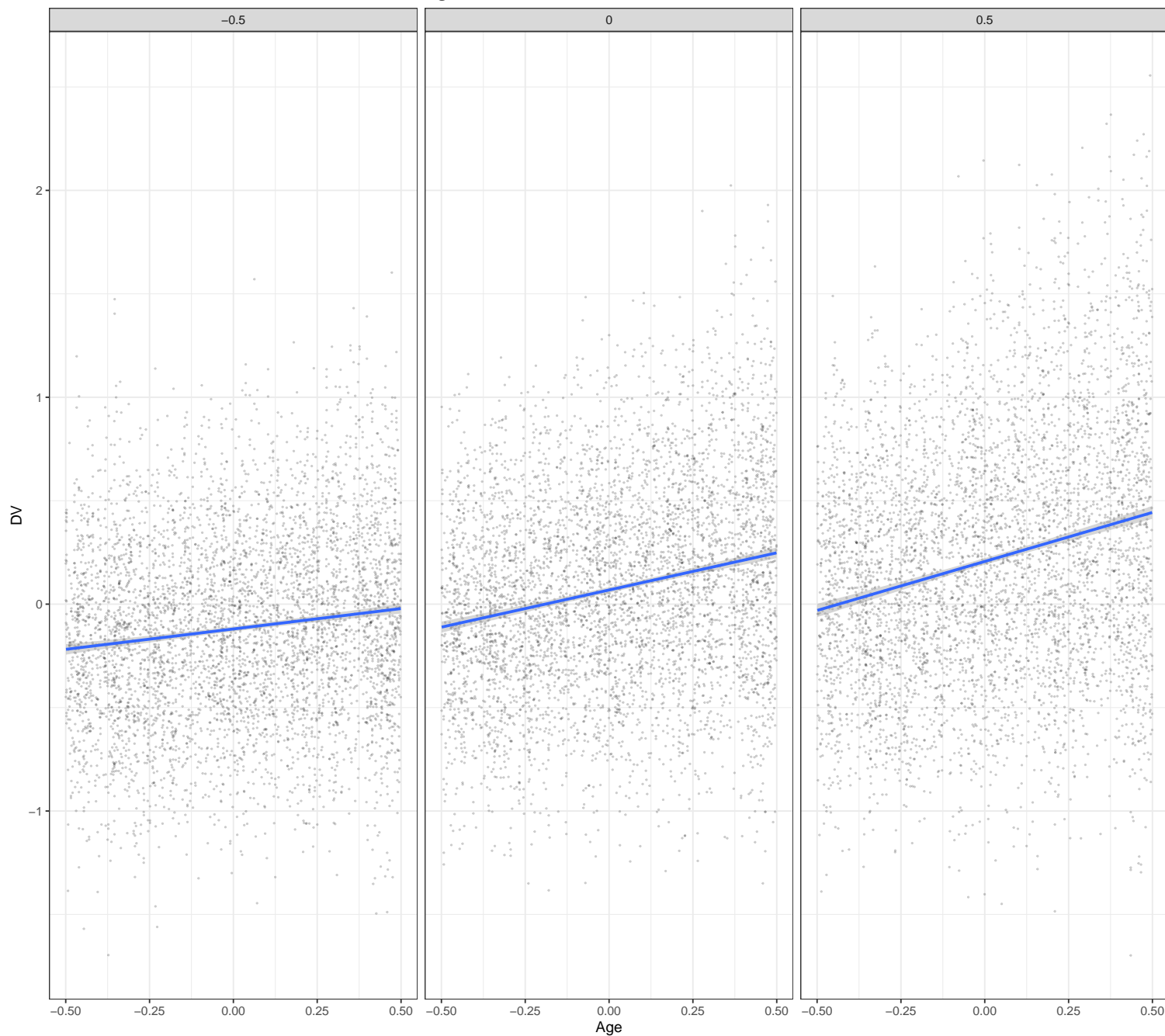
plot_age <- plot_age + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
plot_age
```



## 2.4 Age\*Familiarization:

```
plot_age_familiarization <- dat_sim %>%
  ggplot() + geom_point(aes(y = DV, x = X_a), position = "jitter",
    alpha = 0.2, size = 0.2) + geom_smooth(method = "lm", formula = y ~
    x, se = TRUE, aes(y = DV, x = X_a)) + facet_wrap(~X_f) +
  ggtitle("Age x Familiarization Interaction") + xlab("Age") +
  theme_bw()
plot_age_familiarization <- plot_age_familiarization + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
plot_age_familiarization
```

## Age x Familiarization Interaction



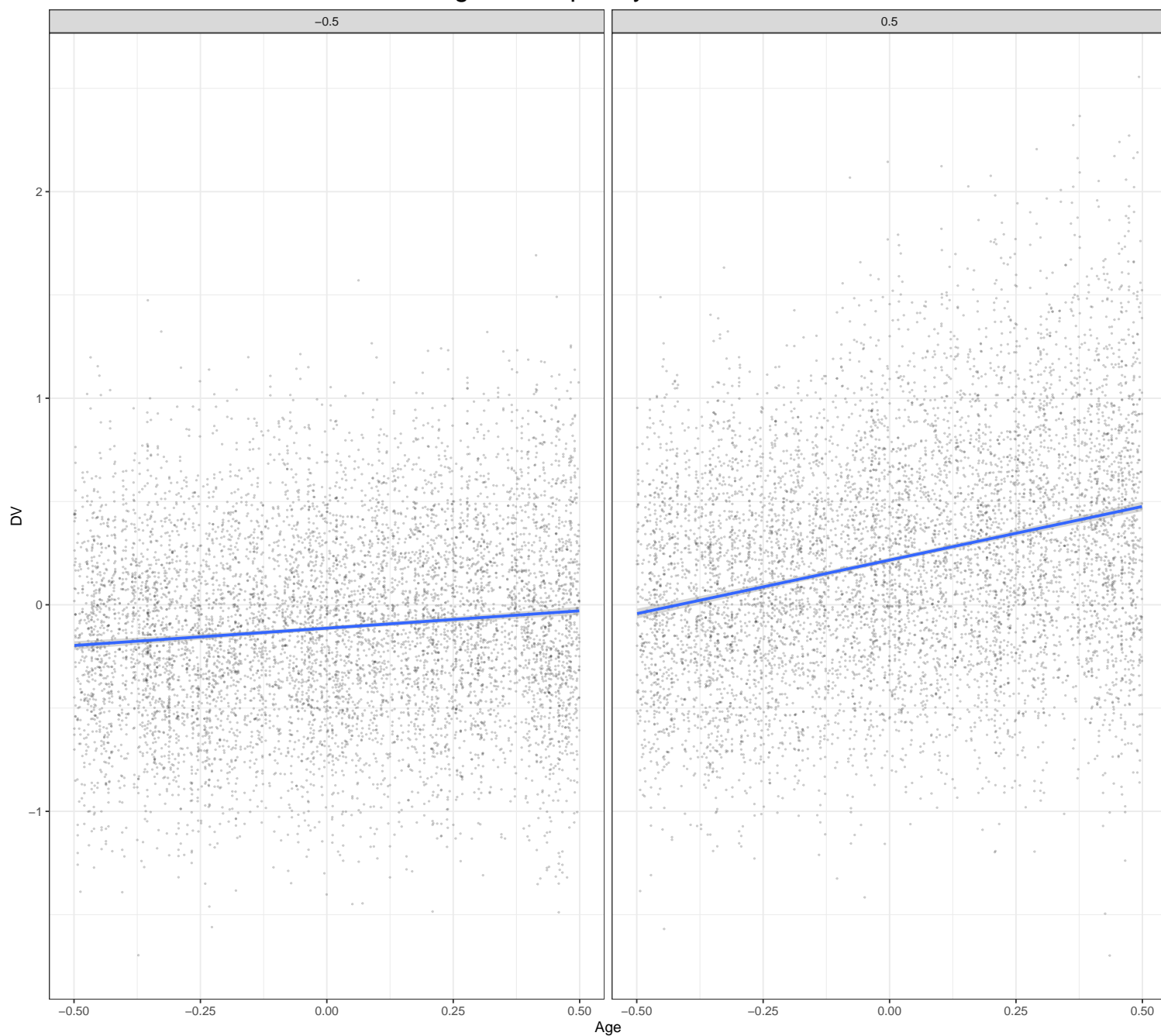
### 2.5 Age\*Complexity:

```
plot_age_complexity <- dat_sim %>%
  ggplot() + geom_point(aes(y = DV, x = X_a), position = "jitter",
    alpha = 0.2, size = 0.2) + geom_smooth(method = "lm", formula = y ~
    x, se = TRUE, aes(y = DV, x = X_a)) + facet_wrap(~X_c) +
  ggtitle("Age x Complexity Interaction") + xlab("Age") + theme_bw()

plot_age_complexity <- plot_age_complexity + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
plot_age_complexity
```



## Age x Complexity Interaction



## 2.6 Familiarization\*Complexity:

```
dat_f_c_interaction <- dat_sim %>%
  mutate(X_c = as.factor(X_c)) %>%
  mutate(X_f = as.factor(X_f)) %>%
  group_by(X_f, X_c) %>%
  dplyr::summarise(med_DV = median(DV))
```

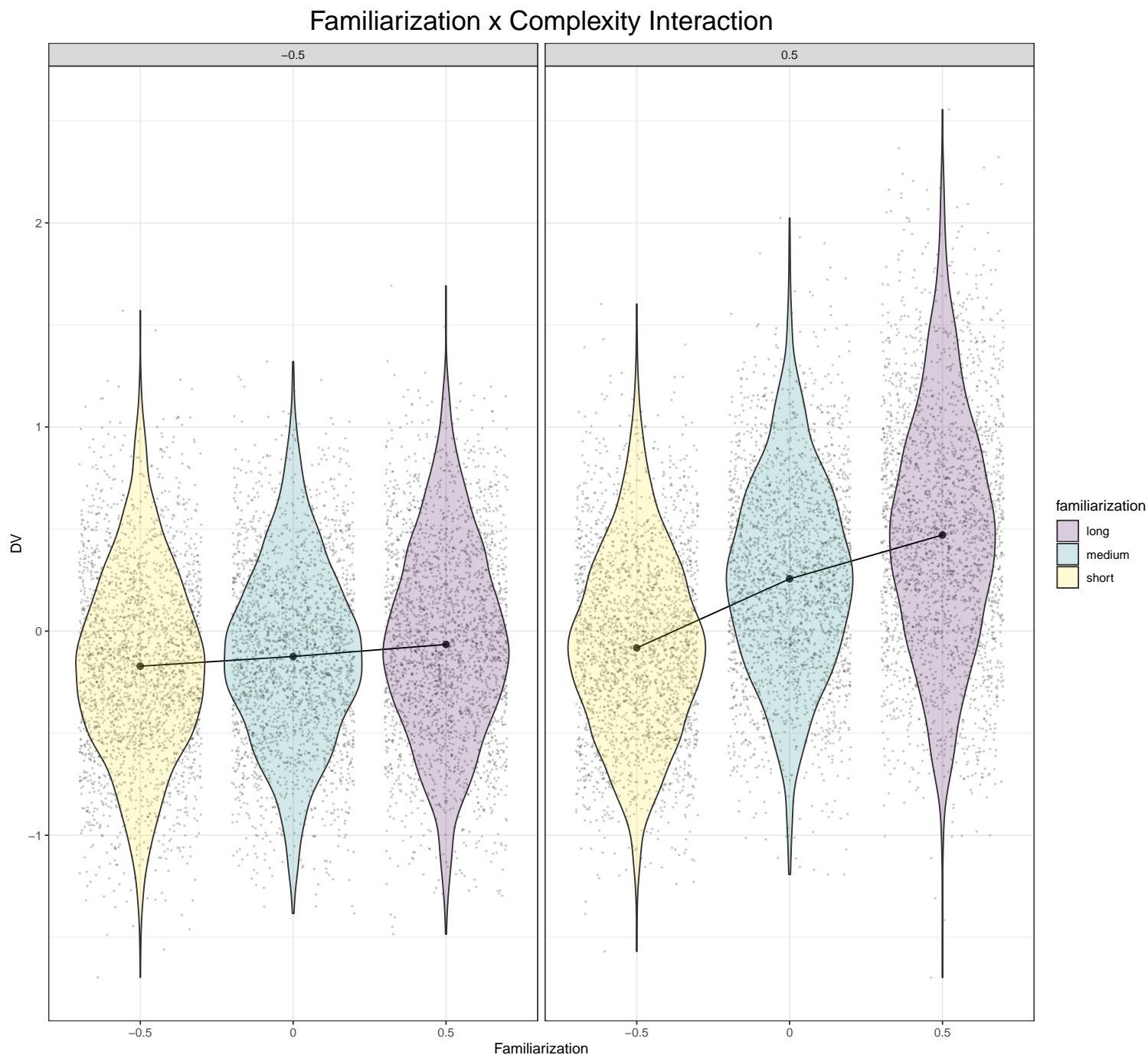
## `summarise()` has grouped output by 'X\_f'. You can override using the `.groups`  
## argument.

```
plot_familiarization_complexity <- dat_sim %>%
  mutate(X_c = as.factor(X_c)) %>%
  mutate(X_f = as.factor(X_f)) %>%
  ggplot() + geom_point(aes(y = DV, x = X_f), position = "jitter",
```

```

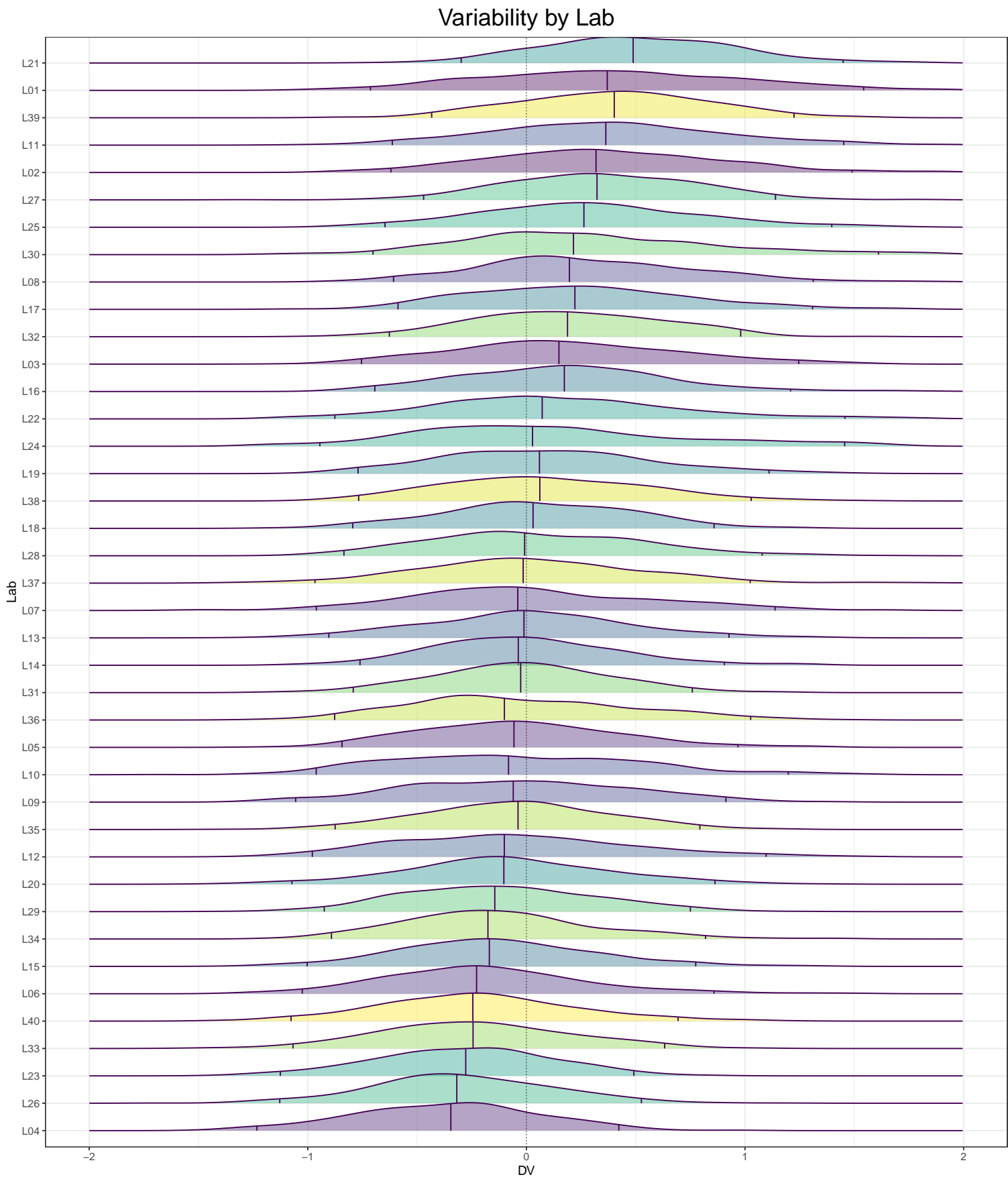
alpha = 0.2, size = 0.2) + geom_point(aes(y = med_DV, x = as.factor(X_f)),
alpha = 0.8, size = 2, data = dat_f_c_interaction) + geom_line(aes(y = med_DV,
x = as.factor(X_f), group = 1), data = dat_f_c_interaction) +
geom_violin(aes(y = DV, x = X_f, fill = familiarization),
alpha = 0.2) + scale_fill_manual(values = viridis(n = 3)) +
facet_wrap(~X_c) + ggtitle("Familiarization x Complexity Interaction") +
xlab("Familiarization") + theme_bw()
plot_familiarization_complexity <- plot_familiarization_complexity +
theme(plot.title = element_text(hjust = 0.5, size = 20))
plot_familiarization_complexity

```

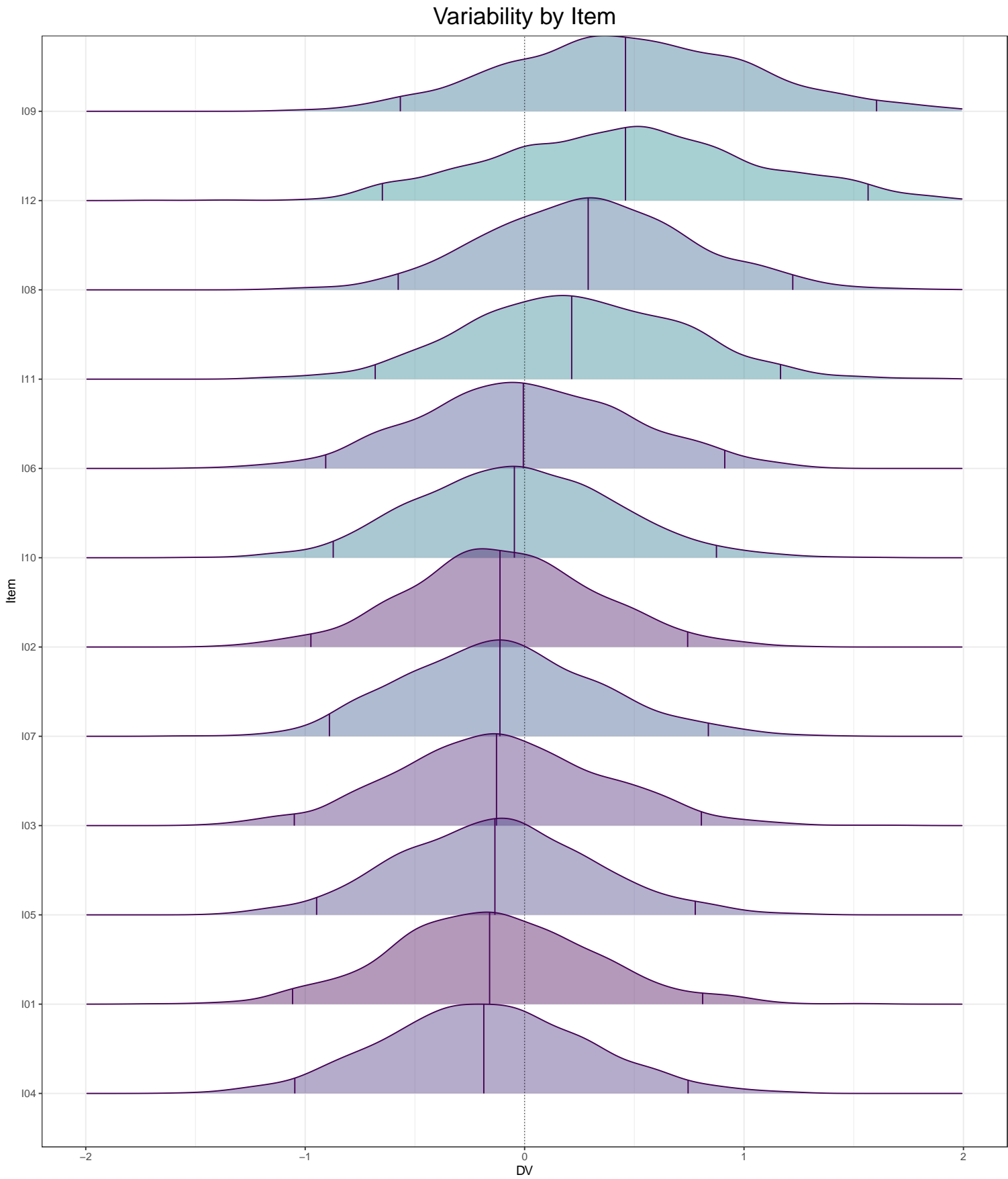




2.7 Variability by Lab



2.8 Variability by Item



### 3 Overview of Power Simulation Results

#### 3.1 Summary Statistics for Power Calculation with Full Data and Varying Intercepts and Varying Slopes:

Table 1: Power for Simulations with Full Data and Varying Intercepts and Varying Slopes

term	power, ef = 0.1	power, ef = 0.2	power, ef = 0.3	power, ef = 0.4	power, ef = 0.5
(Intercept)	0.056	0.044	0.056	0.038	0.044
X_a	0.892	1.000	1.000	1.000	1.000
X_c	0.648	0.988	1.000	1.000	1.000
X_f	0.514	0.966	1.000	1.000	1.000
X_a:X_c	0.744	0.994	1.000	1.000	1.000
X_a:X_f	0.702	0.976	0.998	1.000	1.000
X_c:X_f	0.190	0.600	0.824	0.968	1.000
X_a:X_c:X_f	0.468	0.878	0.972	1.000	1.000

#### 3.2 Summary Statistics for Power Calculation with Full Data and Varying Intercepts:

Table 2: Power for Simulations with Full Data and Varying Intercepts

term	power, ef = 0.1	power, ef = 0.2	power, ef = 0.3	power, ef = 0.4	power, ef = 0.5
(Intercept)	0.060	0.070	0.052	0.038	0.048
X_a	0.896	1.000	1.000	1.000	1.000
X_c	0.638	0.988	1.000	1.000	1.000
X_f	0.476	0.944	1.000	1.000	1.000
X_a:X_c	0.854	0.994	1.000	1.000	1.000
X_a:X_f	0.778	0.992	1.000	1.000	1.000
X_c:X_f	0.192	0.572	0.840	1.000	0.968
X_a:X_c:X_f	0.536	0.830	0.974	1.000	0.996

#### 3.3 Summary Statistics for Power Calculation with 20 pct. Missing Data and Varying Intercepts:

Table 3: Power for Simulations with 20 pct. Missing Data and Varying Intercepts and Slopes

term	power, ef = 0.1	power, ef = 0.2	power, ef = 0.3	power, ef = 0.4	power, ef = 0.5
(Intercept)	0.056	0.050	0.062	0.038	0.046
X_a	0.890	1.000	1.000	1.000	1.000
X_c	0.660	0.988	1.000	1.000	1.000
X_f	0.550	0.940	1.000	1.000	1.000
X_a:X_c	0.818	0.994	1.000	1.000	1.000
X_a:X_f	0.768	0.984	1.000	1.000	1.000
X_c:X_f	0.266	0.534	0.852	0.974	1.000
X_a:X_c:X_f	0.486	0.816	0.954	0.998	1.000

#### 3.4 Summary Statistics for Power Calculation with 50 pct. Missing Data and Varying Intercepts:

Table 4: Power for Simulations with 50 pct. Missing Data and Varying Intercepts and Slopes

term	power, ef = 0.1	power, ef = 0.2	power, ef = 0.3	power, ef = 0.4	power, ef = 0.5
(Intercept)	0.044	0.036	0.056	0.042	0.046
X_a	0.858	0.998	1.000	1.000	1.000
X_c	0.630	0.990	1.000	1.000	1.000
X_f	0.528	0.944	1.000	1.000	1.000
X_a:X_c	0.756	0.990	1.000	1.000	1.000
X_a:X_f	0.666	0.976	1.000	1.000	1.000
X_c:X_f	0.170	0.480	0.840	0.972	0.998
X_a:X_c:X_f	0.370	0.696	0.914	0.994	0.998

## 4 Overview of Bias Results

Table 5: Bias for Simulations with Full Data and Varying Intercepts and Varying Slopes

term	bias, ef = 0.1	bias, ef = 0.2	bias, ef = 0.3	bias, ef = 0.4	bias, ef = 0.5
(Intercept)	0.001	-0.004	0.001	-0.003	-0.003
X_a	-0.001	-0.005	0.001	-0.001	0.001
X_c	0.000	-0.003	0.001	-0.003	0.000
X_f	0.003	-0.003	0.000	-0.003	-0.001
X_a:X_c	-0.009	0.002	0.002	-0.001	0.000
X_a:X_f	-0.001	-0.003	-0.005	0.002	0.000
X_c:X_f	-0.004	-0.011	0.009	0.001	-0.001
X_a:X_c:X_f	0.012	-0.009	-0.005	-0.005	-0.008

Table 6: Bias for Simulations with Full Data and Varying Intercepts

term	bias, ef = 0.1	bias, ef = 0.2	bias, ef = 0.3	bias, ef = 0.4	bias, ef = 0.5
(Intercept)	-0.001	-0.002	0.002	0.000	-0.003
X_a	0.000	0.002	0.003	-0.100	0.100
X_c	0.000	0.001	-0.002	-0.104	0.098
X_f	0.007	0.004	-0.002	-0.100	0.096
X_a:X_c	0.000	-0.002	-0.002	-0.094	0.092
X_a:X_f	-0.002	0.009	0.002	-0.100	0.098
X_c:X_f	0.003	-0.010	0.005	-0.096	0.103
X_a:X_c:X_f	0.001	-0.001	0.000	-0.103	0.092

Table 7: Bias for Simulations with 20 pct. Missing Data and Varying Intercepts and Slopes

term	bias, ef = 0.1	bias, ef = 0.2	bias, ef = 0.3	bias, ef = 0.4	bias, ef = 0.5
(Intercept)	-0.001	0.000	0.001	0.002	0.001
X_a	-0.001	-0.002	-0.002	0.001	-0.004
X_c	-0.001	0.000	0.000	-0.002	0.001
X_f	-0.001	0.006	-0.003	0.001	-0.001
X_a:X_c	0.000	0.005	-0.003	-0.008	-0.001
X_a:X_f	0.005	-0.003	0.000	0.003	0.003
X_c:X_f	-0.007	-0.005	-0.001	0.011	-0.004
X_a:X_c:X_f	-0.002	0.001	-0.003	-0.002	0.003

Table 8: Bias for Simulations with 50 pct. Missing Data and Varying Intercepts and Slopes

term	bias, ef = 0.1	bias, ef = 0.2	bias, ef = 0.3	bias, ef = 0.4	bias, ef = 0.5
(Intercept)	-0.003	-0.002	-0.002	-0.004	0.002
X_a	0.006	0.001	0.000	-0.004	0.002
X_c	0.003	-0.008	0.005	-0.002	0.004
X_f	0.000	0.003	-0.004	-0.003	0.003
X_a:X_c	0.000	0.002	-0.002	0.001	0.001
X_a:X_f	0.003	-0.002	-0.001	0.001	-0.001
X_c:X_f	0.007	0.004	-0.002	-0.009	-0.001
X_a:X_c:X_f	0.020	0.001	0.002	-0.003	-0.001

## 5 Grid Search Code

```
n_subjects <- c(1280, 1920, 2560)
n_trials <- c(12, 18, 24)
b_parameter <- c(0.3, 0.5, 0.7)
sd_parameter <- c(0.1, 0.2, 0.3)
```

```
sigma_parameter <- c(0.1)
```

```
run_sims_grid_point <- function(filename_full, subj_n, trial_n, ef, residual_sd, random_var, random_var_item, rho_val)
```

```
  participants_per_lab<-32
```

```
  n_simple<- trial_n/2
```

```
  n_complex<- trial_n/2
```

```
  n_small_fam<- trial_n/3
```

```
  n_medium_fam<- trial_n/3
```

```
  n_high_fam<- trial_n/3
```

```
  n_lab<-floor(subj_n/participants_per_lab)
```

```
  dat_sim <- my_sim_data(n_subj = subj_n,  
                        n_simple = n_simple,  
                        n_complex = n_complex,  
                        n_small_fam = n_small_fam,  
                        n_medium_fam = n_medium_fam,  
                        n_high_fam = n_high_fam,  
                        n_lab = n_lab,  
                        beta_c = ef,  
                        beta_f = ef,  
                        beta_a = ef,  
  
                        beta_ca = ef,  
                        beta_af = ef,  
                        beta_cf = ef,  
                        beta_cfa = ef,  
  
                        subject_0 = random_var,  
                        subject_c = random_var,  
                        subject_f = random_var,  
                        subject_a = random_var,  
                        subject_ca = random_var,  
                        subject_af = random_var,  
                        subject_cf = random_var,  
                        subject_cfa = random_var,  
                        subj_rho = rho_val,  
  
                        lab_0 = random_var,  
                        lab_c = random_var,  
                        lab_f = random_var,  
                        lab_a = random_var,  
                        lab_ca = random_var,  
                        lab_af = random_var,  
                        lab_cf = random_var,  
                        lab_cfa = random_var,  
                        lab_rho = rho_val,  
  
                        item_0 = random_var_item,  
                        item_c = random_var_item,  
                        item_f = random_var_item,  
                        item_a = random_var_item,  
                        item_ca = random_var_item,  
                        item_af = random_var_item,  
                        item_cf = random_var_item,  
                        item_cfa = random_var_item,  
                        item_rho = rho_val,  
  
                        sigma = residual_sd,  
                        )
```

```

mod_sim <- lmer(DV ~ 1 + X_a * X_c * X_f +
               (1 | subj_id) +
               (1 | lab_id) +
               (1 | item_id),
               data=dat_sim)

sim_results <- broom.mixed::tidy(mod_sim)

# append the results to a file
append <- file.exists(filename_full)
write_csv(sim_results, filename_full, append = append)

# return the tidy table
sim_results
}

reps <- 50

n_subj_values <- c(1280, 1920, 2560)
trial_n_values <- c(12, 18, 24)
random_var <- c(0.1, 0.2)
random_var_item <- c(0.05, 0.1, 0.2)
b_parameter <- c(0.3, 0.5, 0.7)
residual_sd <- c(0.1, 0.3)

param_combinations <- expand.grid(n_subj = n_subj_values,
                                  trial_n = trial_n_values,
                                  random_var = random_var,
                                  random_var_item = random_var_item,
                                  b_parameter = b_parameter,
                                  residual_sd = residual_sd)

for (i in seq_len(nrow(param_combinations))) {
  start_time <- Sys.time()

  sim_params <- param_combinations[i, ]
  filename_full <- paste0('sims_grid_search/test_grid_search_',
                          sim_params$n_subj, '_',
                          sim_params$trial_n, '_',
                          sim_params$random_var, '_',
                          sim_params$random_var_item, '_',
                          sim_params$b_parameter, '_',
                          sim_params$residual_sd, '.csv')

  sims <- purrr::map_df(1:reps, ~run_sims_grid_point(filename_full = filename_full,
                                                    subj_n = sim_params$n_subj,
                                                    trial_n = sim_params$trial_n,
                                                    ef = sim_params$b_parameter,
                                                    random_var = sim_params$random_var,
                                                    random_var_item = sim_params$random_var_item,
                                                    residual_sd = sim_params$residual_sd,
                                                    rho_val = 0.2))

  end_time <- Sys.time()
  cat("Simulation", i, "Time elapsed:", end_time - start_time, "\n")
}

# set the directory where the CSV files are located
setwd("/work/sims_grid_search")

# get the file names for CSV files

```

```
file_names <- list.files(pattern = "*.csv")

# read in all CSV files into a list of dataframes
df_list <- purrr::map(file_names, ~{
  df <- read.csv(.x)
  df$filename <- .x
  df
})

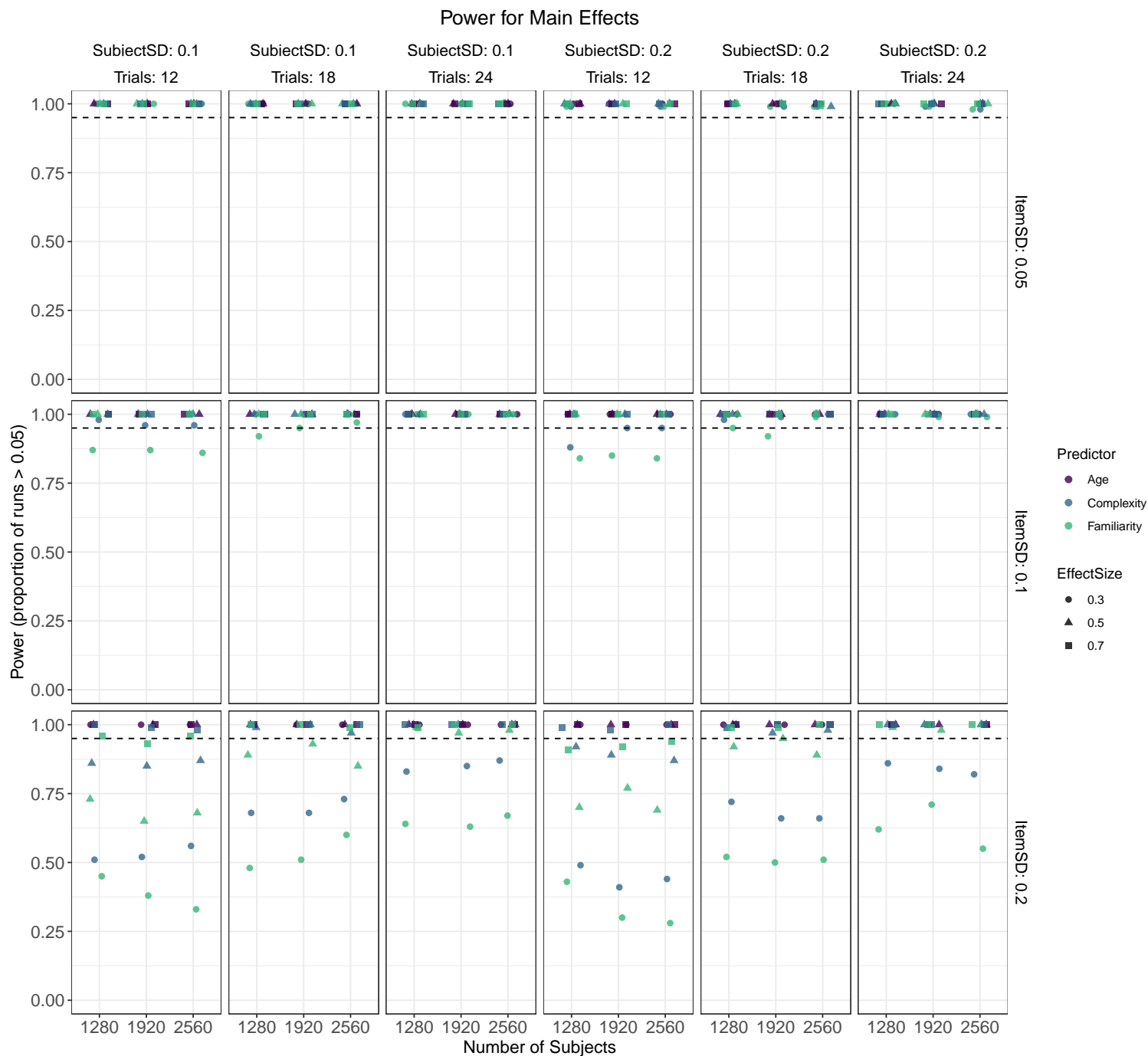
df <- purrr::reduce(df_list, dplyr::bind_rows)

df_per_sim <- df %>%
  filter(effect == "fixed") %>%
  group_by(filename, term) %>%
  summarise(median_estimate = median(estimate), median_se = median(std.error),
    power = mean(p.value < 0.05)) %>%
  mutate(n_subj = sapply(strsplit(filename, "_"), `[`, 4),
    Trials = sapply(strsplit(filename, "_"), `[`, 5), SubjectSD = sapply(strsplit(filename,
      "_"), `[`, 6), ItemSD = sapply(strsplit(filename,
        "_"), `[`, 7), EffectSize = sapply(strsplit(filename,
          "_"), `[`, 8), residual_sd = as.numeric(str_replace(sapply(strsplit(filename,
            "_"), `[`, 9), pattern = ".csv", "")))

error_bars <- df_per_sim %>%
  filter(term != "(Intercept)") %>%
  group_by(EffectSize, SubjectSD, Trials, ItemSD, n_subj, term) %>%
  dplyr::summarise(power_mean = median(power), power_lci = quantile(power,
    prob = 0.025), power_hci = quantile(power, prob = 0.975)) %>%
  filter(term %in% c("X_a", "X_c", "X_f")) %>%
  dplyr::rename(Predictor = term) %>%
  mutate(Predictor = ifelse(Predictor == "X_a", "Age", Predictor)) %>%
  mutate(Predictor = ifelse(Predictor == "X_c", "Complexity",
    Predictor)) %>%
  mutate(Predictor = ifelse(Predictor == "X_f", "Familiarity",
    Predictor))

MainEffectGridSearchPlot <- ggplot() + geom_point(aes(x = n_subj,
  y = power_mean, color = Predictor, shape = EffectSize), data = error_bars,
  size = 2, alpha = 0.8, position = position_jitter(width = 0.2,
    height = 0)) + geom_hline(yintercept = 0.95, linetype = 2) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_color_manual(values = viridis(n = 4)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +
  scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1), limits = c(0,
    1)) + facet_grid(ItemSD ~ SubjectSD + Trials, labeller = label_both) +
  ggtitle("Power for Main Effects") + xlab("Number of Subjects") +
  ylab("Power (proportion of runs > 0.05)") + theme_bw() +
  theme(strip.background = element_rect(color = "white", fill = "white",
    size = 5, linetype = "solid"), plot.title = element_text(hjust = 0.5,
    size = 15), strip.text.x = element_text(size = 12, color = "black"),
    strip.text.y = element_text(size = 12, color = "black"),
    axis.text.x = element_text(size = 13), axis.title.x = element_text(size = 13),
```

```
axis.text.y = element_text(size = 13), axis.title.y = element_text(size = 13))
MainEffectGridSearchPlot
```



```
error_bars <- df_per_sim %>%
  filter(term != "(Intercept)") %>%
  group_by(EffectSize, SubjectSD, Trials, ItemSD, n_subj, term) %>%
  dplyr::summarise(power_mean = median(power), power_lci = quantile(power,
    prob = 0.025), power_hci = quantile(power, prob = 0.975)) %>%
  filter(term %in% c("X_a:X_c", "X_a:X_f", "X_c:X_f", "X_a:X_c:X_f")) %>%
  dplyr::rename(Predictor = term) %>%
  mutate(Predictor = ifelse(Predictor == "X_a:X_c", "Age:Complexity",
    Predictor)) %>%
  mutate(Predictor = ifelse(Predictor == "X_a:X_f", "Age:Familiarity",
    Predictor)) %>%
  mutate(Predictor = ifelse(Predictor == "X_c:X_f", "Complexity:Familiarity",
    Predictor)) %>%
  mutate(Predictor = ifelse(Predictor == "X_a:X_c:X_f", "Familiarity:Age:Complexity",
```

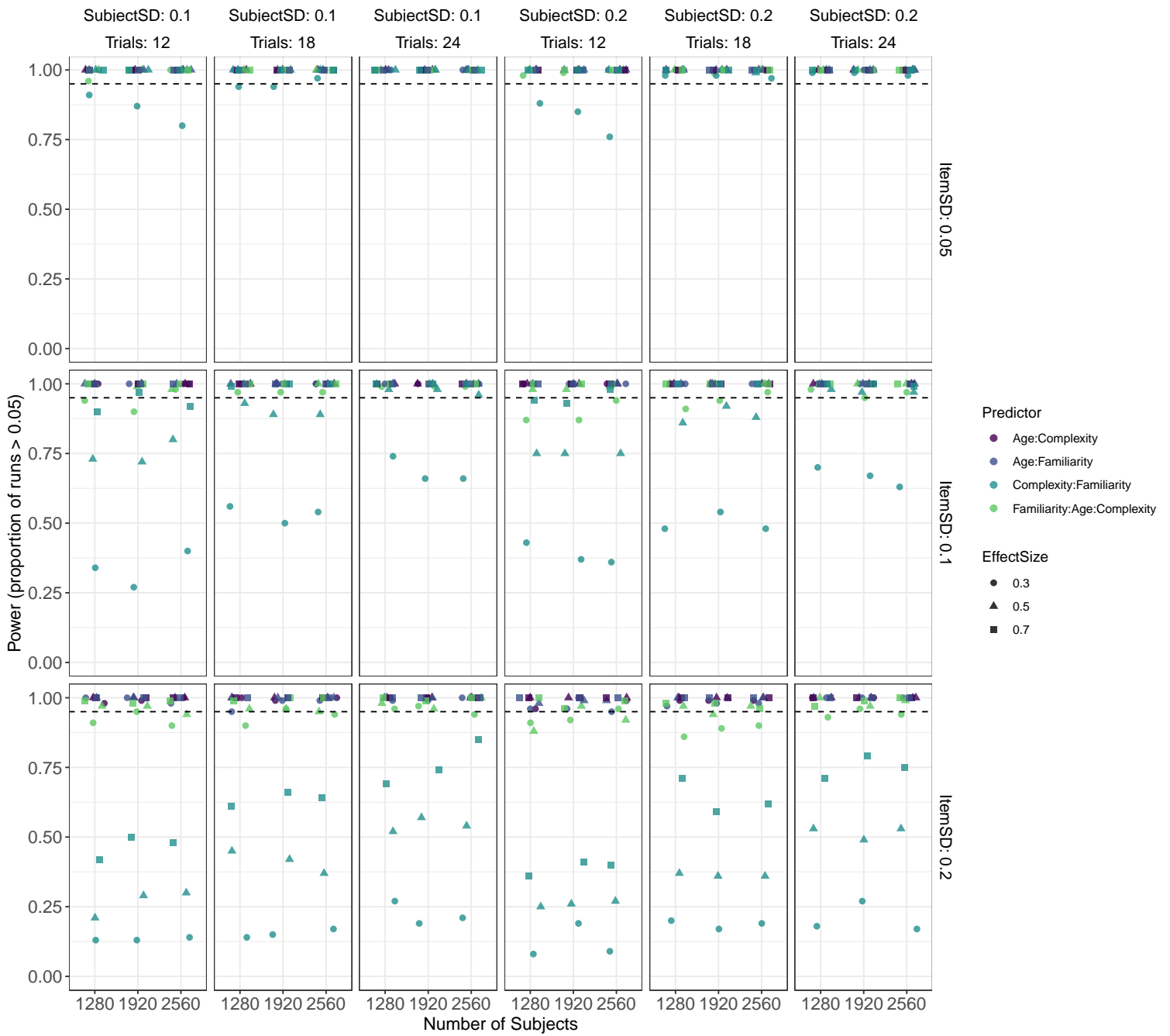


```
Predictor))
```

```
InteractionEffectGridSearchPlot <- ggplot() + geom_point(aes(x = n_subj,  
y = power_mean, color = Predictor, shape = EffectSize), data = error_bars,  
size = 2, alpha = 0.8, position = position_jitter(width = 0.25,  
height = 0)) + geom_hline(yintercept = 0.95, linetype = 2) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_color_manual(values = viridis(n = 5)) + #scale_fill_manual(values = brewer.pal(n=8, name = 'Dark2')) +  
scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1), limits = c(0,  
1)) + facet_grid(ItemSD ~ SubjectSD + Trials, labeller = label_both) +  
ggtitle("Power for Interaction Effects") + xlab("Number of Subjects") +  
ylab("Power (proportion of runs > 0.05)") + theme_bw() +  
theme(strip.background = element_rect(color = "white", fill = "white",  
size = 5, linetype = "solid"), plot.title = element_text(hjust = 0.5,  
size = 15), strip.text.x = element_text(size = 12, color = "black"),  
strip.text.y = element_text(size = 12, color = "black"),  
axis.text.x = element_text(size = 13), axis.title.x = element_text(size = 13),  
axis.text.y = element_text(size = 13), axis.title.y = element_text(size = 13))
```

InteractionEffectGridSearchPlot

## Power for Interaction Effects



```
ggsave(plot = MainEffectGridSearchPlot, file = "MainEffectGridSearchPlot.png",
        height = 10, width = 15)
ggsave(plot = InteractionEffectGridSearchPlot, file = "InteractionEffectGridSearchPlot.png",
        height = 10, width = 15)
```

## 6 Power Calculation with Full Data and Varying Intercepts and Varying Slopes

### 6.1 Effect Size = 0.5

```
# Number of simulations:
reps <- 500

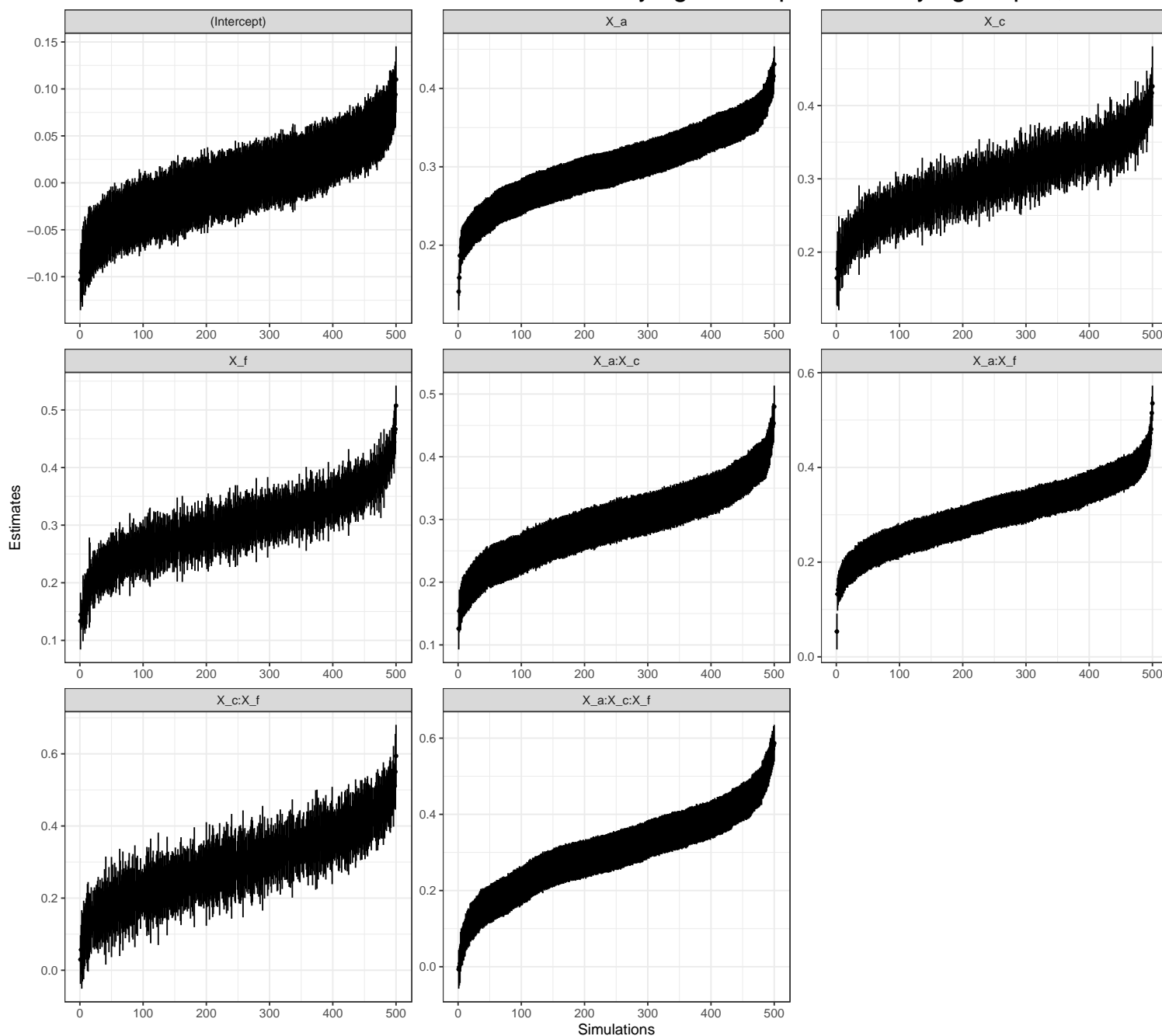
# Simulation function:
run_sims <- function(filename_full, ef) {
```



```
fixed_full_plot <- sims_full_0.3 %>%
  filter(effect == "fixed") %>%
  ungroup() %>%
  arrange(term, estimate) %>%
  mutate(row = rep(seq(1:reps), 8)) %>%
  ggplot(aes(x = row, y = estimate, ymin = estimate - std.error,
             ymax = estimate + std.error)) + facet_wrap(~term, scales = "free") +
  geom_pointrange(fatten = 1/2) + ylab("Estimates") + xlab("Simulations") +
  ggtitle("Estimates of Fixed Effects for Full Data and Varying Intercepts and Varying Slopes, ef = 0.3") +
  theme_bw()

fixed_full_plot <- fixed_full_plot + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
fixed_full_plot
```

Estimates of Fixed Effects for Full Data and Varying Intercepts and Varying Slopes, ef = 0.3

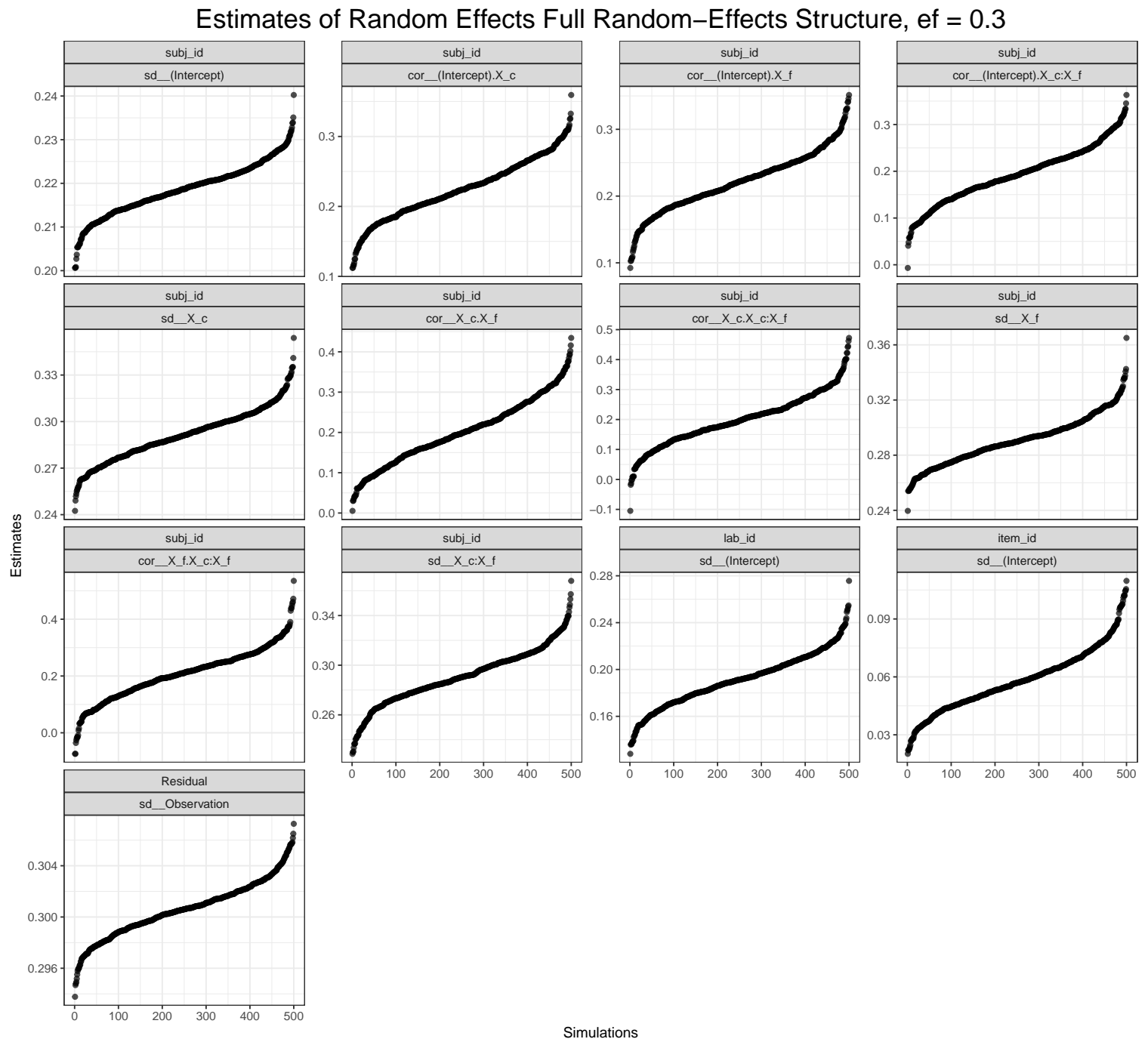


### 6.3.2 Visualise Estimates for Random Effects:

```

ran_full_plot <- sims_full_0.3 %>%
  filter(effect == "ran_pars") %>%
  ungroup() %>%
  arrange(group, term, estimate) %>%
  mutate(row = rep(seq(1:reps), 13)) %>%
  ggplot(aes(x = row, y = estimate)) + geom_point(alpha = 0.7) +
  facet_wrap(~group + term, scales = "free_y") + theme_bw() +
  ylab("Estimates") + xlab("Simulations") + ggtitle("Estimates of Random Effects Full Random-Effects Structure,
  theme_bw()
ran_full_plot <- ran_full_plot + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
ran_full_plot

```



## 6.4 Effect Size = 0.2

```
filename_full_0.2 = "run_sims_full_0.2.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_0.2,
  ef = 0.2))
end_time <- Sys.time()
end_time - start_time
```

## 6.5 Effect Size = 0.1

```
filename_full_0.1 = "run_sims_full_0.1.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_0.1,
  ef = 0.1))
end_time <- Sys.time()
end_time - start_time
```

# 7 Power Calculation with Full Data and Varying Intercepts

## 7.1 Effect Size = 0.5

```
# Simulation function:
run_sims <- function(filename_full, ef) {

  dat_sim <- my_sim_data(beta_c = ef,
    beta_f = ef,
    beta_a = ef,

    beta_ca = ef,
    beta_af = ef,
    beta_cf = ef,

    beta_cfa = ef)

  mod_sim <- lmer(DV ~ 1 + X_a * X_c * X_f +
    (1 | subj_id) +
    (1 | lab_id) +
    (1 | item_id),
    data=dat_sim)

  sim_results <- broom.mixed::tidy(mod_sim)

  # append the results to a file
  append <- file.exists(filename_full)
  write_csv(sim_results, filename_full, append = append)

  # return the tidy table
  sim_results
}

filename_full_int_0.5 = 'run_sims_full_int_0.5.csv'
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_int_0.5, ef = 0.5))
end_time <- Sys.time()
end_time - start_time
```

## 7.2 Effect Size = 0.4

```
filename_full_int_0.4 = "run_sims_full_int_0.4.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_int_0.4,
  ef = 0.4))
end_time <- Sys.time()
end_time - start_time
```

## 7.3 Effect Size = 0.3

```
filename_full_int_0.3 = "run_sims_full_int_0.3.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_int_0.3,
  ef = 0.3))
end_time <- Sys.time()
end_time - start_time
```

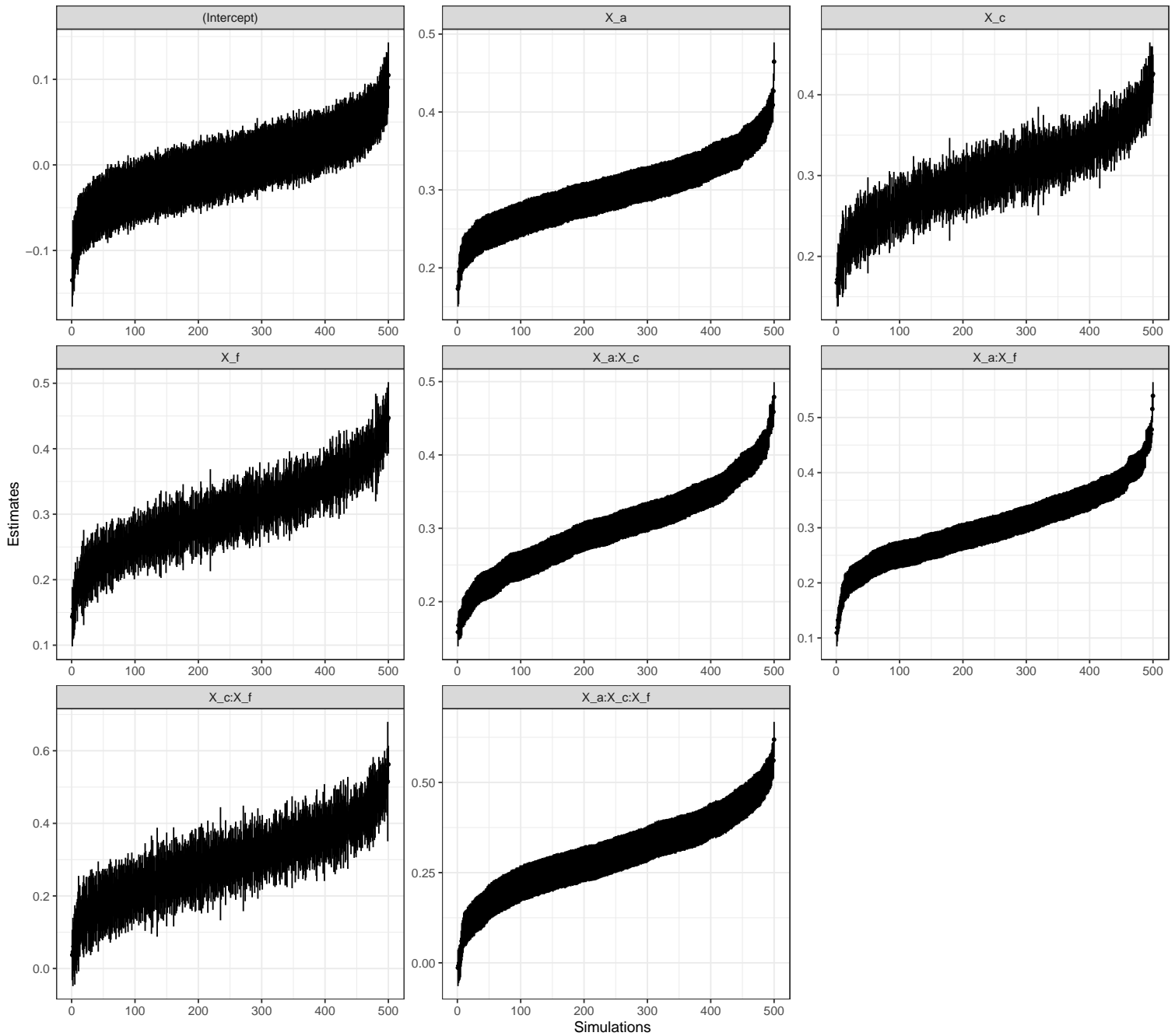
### 7.3.1 Visualise Estimates for Fixed Effects:

```
sims_full_int_0.3 <- read_csv(filename_full_int_0.3, col_types = cols(group = col_factor(ordered = TRUE),
  term = col_factor(ordered = TRUE)))

fixed_full_int_plot <- sims_full_int_0.3 %>%
  filter(effect == "fixed") %>%
  ungroup() %>%
  arrange(term, estimate) %>%
  mutate(row = rep(seq(1:reps), 8)) %>%
  ggplot(aes(x = row, y = estimate, ymin = estimate - std.error,
    ymax = estimate + std.error)) + facet_wrap(~term, scales = "free") +
  geom_pointrange(fatten = 1/2) + ylab("Estimates") + xlab("Simulations") +
  ggtitle("Estimates of Fixed Effects for Full Data and Random Intercepts, ef = 0.3") +
  theme_bw()

fixed_full_int_plot <- fixed_full_int_plot + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
fixed_full_int_plot
```

## Estimates of Fixed Effects for Full Data and Random Intercepts, ef = 0.3

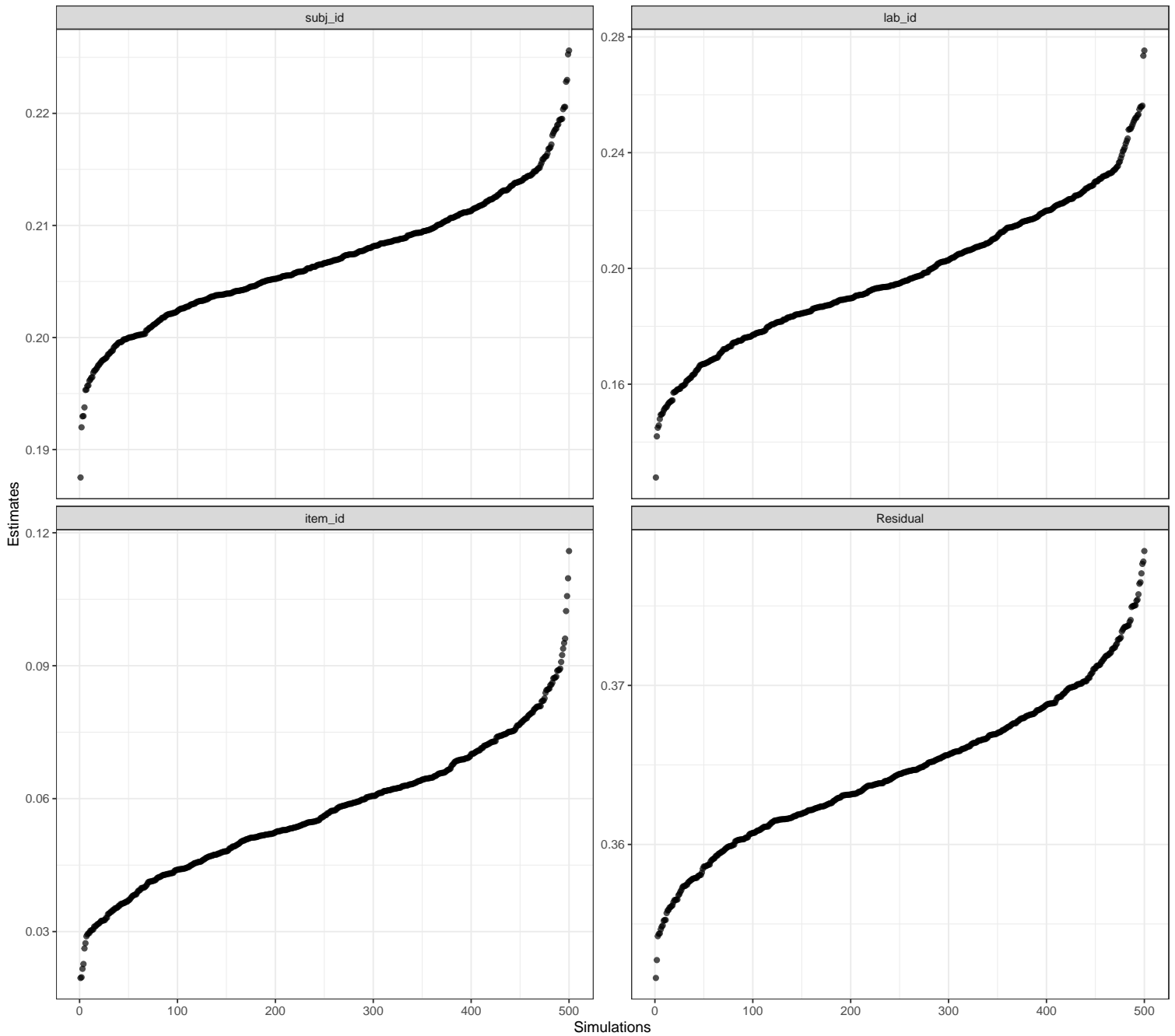


### 7.3.2 Visualise Estimates for Random Effects:

```
ran_full_int_plot <- sims_full_int_0.3 %>%
  filter(effect == "ran_pars") %>%
  ungroup() %>%
  arrange(group, estimate) %>%
  mutate(row = rep(seq(1:reps), 4)) %>%
  ggplot(aes(x = row, y = estimate)) + geom_point(alpha = 0.7) +
  facet_wrap(~group, scales = "free_y") + theme_bw() + ylab("Estimates") +
  xlab("Simulations") + ggtitle("Estimates of Random Effects for Full Data, ef = 0.3") +
  theme_bw()
ran_full_int_plot <- ran_full_int_plot + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
ran_full_int_plot
```



## Estimates of Random Effects for Full Data, ef = 0.3



### 7.4 Effect Size = 0.2

```
filename_full_int_0.2 = "run_sims_full_int_0.2.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_int_0.2,
  ef = 0.2))
end_time <- Sys.time()
end_time - start_time
```

### 7.5 Effect Size = 0.1

```
filename_full_int_0.1 = "run_sims_full_int_0.1.csv"
start_time <- Sys.time()
sims <- purrr::map_df(1:reps, ~run_sims(filename_full = filename_full_int_0.1,
```

```

    ef = 0.1))
end_time <- Sys.time()
end_time - start_time

```

## 8 Power Calculation with 20 pct. Missing Data and Varying Intercepts

### 8.1 Effect Size = 0.5

```

run_sims_missing <- function(filename_missing, ef) {

  dat_sim <- my_sim_data(beta_c = ef,
                        beta_f = ef,
                        beta_a = ef,

                        beta_ca = ef,
                        beta_af = ef,
                        beta_cf = ef,

                        beta_cfa = ef)

  missing_samples <- dat_sim %>%
    mutate(nas = rbinom(nrow(dat_sim), 1, 1 - .20)) %>%
    mutate(DV = ifelse(nas == 1, DV, NA)) %>%
    drop_na()

  mod_sim <- lmer(DV ~ 1 + X_a * X_c * X_f +
                (1 | subj_id) +
                (1 | lab_id) +
                (1 | item_id),
                data=missing_samples)

  sim_results <- broom.mixed::tidy(mod_sim)

  # append the results to a file
  append <- file.exists(filename_missing)
  write_csv(sim_results, filename_missing, append = append)

  # return the tidy table
  sim_results
}

filename_20_missing_0.5 = 'run_sims_20_missing_0.5.csv'
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_20_missing_0.5, ef = 0.5))
end_time <- Sys.time()
end_time - start_time

```

### 8.2 Effect Size = 0.4

```

filename_20_missing_0.4 = "run_sims_20_missing_0.4.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_20_missing_0.4,
    ef = 0.4))
end_time <- Sys.time()
end_time - start_time

```

## 8.3 Effect Size = 0.3

```
filename_20_missing_0.3 = "run_sims_20_missing_0.3.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_20_missing_0.3,
  ef = 0.3))
end_time <- Sys.time()
end_time - start_time
```

### 8.3.1 Visualise Estimates for Fixed Effects:

```
# read saved simulation data
sims_20_missing_0.3 <- read_csv(filename_20_missing_0.3, col_types = cols(
  # makes sure plots display in this order
  group = col_factor(ordered = TRUE),
  term = col_factor(ordered = TRUE)
))

fixed_missing_plot <- sims_20_missing_0.3 %>%
  filter(effect == "fixed") %>%
  ungroup() %>%
  arrange(term, estimate) %>%
  mutate(row = rep(seq(1:reps), 8)) %>%
  ggplot(aes(x = row, y = estimate, ymin = estimate-std.error, ymax = estimate+std.error)) +
  facet_wrap(~term, scales = "free") +
  geom_pointrange(fatten = 1/2) +
  ylab("Estimates") +
  xlab("Simulations") +
  ggtitle('Estimates of Fixed Effects for 20 pct. Missing Data, ef = 0.3') +
  theme_bw()

fixed_missing_plot <- fixed_missing_plot + theme(plot.title = element_text(hjust = 0.5, size=20))
fixed_missing_plot
```

### 8.3.2 Visualise Estimates for Random Effects:

```
ran_missing_plot <- sims_20_missing_0.3 %>%
  filter(effect == "ran_pars") %>%
  ungroup() %>%
  arrange(group, term, estimate) %>%
  mutate(row = rep(seq(1:reps), 13)) %>%
  ggplot(aes(x = row, y = estimate)) + geom_point(alpha = 0.7) +
  facet_wrap(~group + term, scales = "free_y") + theme_bw() +
  ylab("Estimates") + xlab("Simulations") + ggtitle("Estimates of Random Effects for 20 pct. Missing Data, ef = 0.3") +
  theme_bw()

ran_missing_plot <- ran_missing_plot + theme(plot.title = element_text(hjust = 0.5,
  size = 20))
ran_missing_plot
```

## 8.4 Effect Size = 0.2

```
filename_20_missing_0.2 = "run_sims_20_missing_0.2.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_20_missing_0.2,
  ef = 0.2))
```

```
end_time <- Sys.time()
end_time - start_time
```

## 8.5 Effect Size = 0.1

```
filename_20_missing_0.1 = "run_sims_20_missing_0.1.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_20_missing_0.1,
  ef = 0.1))
end_time <- Sys.time()
end_time - start_time
```

# 9 Power Calculation with 50 pct. Missing Data and Varying Intercepts

## 9.1 Effect Size = 0.5

```
run_sims_missing <- function(filename_missing, ef) {

  dat_sim <- my_sim_data(beta_c = ef,
                        beta_f = ef,
                        beta_a = ef,

                        beta_ca = ef,
                        beta_af = ef,
                        beta_cf = ef,

                        beta_cfa = ef)

  missing_samples <- dat_sim %>%
    mutate(nas = rbinom(nrow(dat_sim), 1, 1 - .50)) %>%
    mutate(DV = ifelse(nas == 1, DV, NA)) %>%
    drop_na()

  mod_sim <- lmer(DV ~ 1 + X_a * X_c * X_f +
                (1 | subj_id) +
                (1 | lab_id) +
                (1 | item_id),
                data=missing_samples)

  sim_results <- broom.mixed::tidy(mod_sim)

  # append the results to a file
  append <- file.exists(filename_missing)
  write_csv(sim_results, filename_missing, append = append)

  # return the tidy table
  sim_results
}

filename_50_missing_0.5 = 'run_sims_50_missing_0.5.csv'
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_50_missing_0.5, ef = 0.5))
end_time <- Sys.time()
end_time - start_time
```

## 9.2 Effect Size = 0.4

```
filename_50_missing_0.4 = "run_sims_50_missing_0.4.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_50_missing_0.4,
  ef = 0.4))
end_time <- Sys.time()
end_time - start_time
```

## 9.3 Effect Size = 0.3

```
filename_50_missing_0.3 = "run_sims_50_missing_0.3.csv"
start_time <- Sys.time()
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_50_missing_0.3,
  ef = 0.3))
end_time <- Sys.time()
end_time - start_time
```

### 9.3.1 Visualise Estimates for Fixed Effects:

```
# read saved simulation data
sims_50_missing_0.3 <- read_csv(filename_50_missing_0.3, col_types = cols(
  # makes sure plots display in this order
  group = col_factor(ordered = TRUE),
  term = col_factor(ordered = TRUE)
))

fixed_missing_plot <- sims_50_missing_0.3 %>%
  filter(effect == "fixed") %>%
  ungroup() %>%
  arrange(term, estimate) %>%
  mutate(row = rep(seq(1:reps), 8)) %>%
  ggplot(aes(x = row, y = estimate, ymin = estimate-std.error, ymax = estimate+std.error)) +
  facet_wrap(~term, scales = "free") +
  geom_pointrange(fatten = 1/2) +
  ylab("Estimates") +
  xlab("Simulations") +
  ggtitle('Estimates of Fixed Effects for 50 pct. Missing Data, ef = 0.3') +
  theme_bw()

fixed_missing_plot <- fixed_missing_plot + theme(plot.title = element_text(hjust = 0.5, size=20))
fixed_missing_plot
```

### 9.3.2 Visualise Estimates for Random Effects:

```
ran_missing_plot <- sims_20_missing_0.3 %>%
  filter(effect == "ran_pars") %>%
  ungroup() %>%
  arrange(group, term, estimate) %>%
  mutate(row = rep(seq(1:reps), 13)) %>%
  ggplot(aes(x = row, y = estimate)) + geom_point(alpha = 0.7) +
  facet_wrap(~group + term, scales = "free_y") + theme_bw() +
  ylab("Estimates") + xlab("Simulations") + ggtitle("Estimates of Random Effects for 50 pct. Missing Data, ef = 0.3") +
  theme_bw()

ran_missing_plot <- ran_missing_plot + theme(plot.title = element_text(hjust = 0.5,
```

```
size = 20))  
ran_missing_plot
```

## 9.4 Effect Size = 0.2

```
filename_50_missing_0.2 = "run_sims_50_missing_0.2.csv"  
start_time <- Sys.time()  
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_50_missing_0.2,  
  ef = 0.2))  
end_time <- Sys.time()  
end_time - start_time
```

## 9.5 Effect Size = 0.1

```
filename_50_missing_0.1 = "run_sims_50_missing_0.1.csv"  
start_time <- Sys.time()  
sims_missing <- purrr::map_df(1:reps, ~run_sims_missing(filename_missing = filename_50_missing_0.1,  
  ef = 0.1))  
end_time <- Sys.time()  
end_time - start_time
```