




Commentary on the scientific rigor of Sen and Gredebäck's simulation: Why empirical parameters are necessary to build simulations

Kimberly Cuevas¹  | Scott A. Adler²  | Rachel Barr³ | John Colombo⁴ |
Peter Gerhardstein⁵  | Harlene Hayne⁶ | Pamela S. Hunt⁷ | Rick Richardson⁸

¹University of Connecticut, Storrs, Connecticut, USA

²York University, Toronto, Ontario, Canada

³Georgetown University, Washington, District of Columbia, USA

⁴University of Kansas, Lawrence, Kansas, USA

⁵Binghamton University, State University of New York, Binghamton, New York, USA

⁶Curtin University, Perth, Australia

⁷College of William & Mary, Williamsburg, Virginia, USA

⁸University of New South Wales, Sydney, New South Wales, Australia

Correspondence

Kimberly Cuevas, University of Connecticut, Storrs, CT, USA.

Email: kimberly.cuevas@uconn.edu

The operant mobile paradigm has been used in conjunction with rigorous and systematic experimental design to characterize infant memory and its development for over 50 years. This body of research has involved testing thousands of 3- to 6-month-old infants in their own homes on multiple occasions. The basic findings obtained using the paradigm (i.e., older infants learn faster than younger infants, older infants remember longer than younger infants, and changes to the experimental context or test stimuli disrupt memory retrieval, particularly if the infant is very young), have been repeatedly replicated both within the mobile paradigm and with two other memory paradigms with preverbal infants (i.e., deferred imitation, visual recognition; see Cuevas & Davinson, 2022; Hayne, 2004, for reviews).

In the accompanying paper, Sen and Gredebäck (2022) have raised questions about this literature. Their evaluation of 77 publications using the mobile paradigm confirmed the scientific rigor of the existing literature with no evidence of methodological biases in terms of p-hacking, reporting errors (“unintentional errors and fraud”), and variations in sample size (“opportunistic

use of the stopping rule”). However, they then conclude that “...the literature has been contaminated by methodological artifacts due to the opportunistic use of researcher degrees of freedom.” (abstract). Sen and Gredebäck base this assertion on simulation analyses of random kicking data. We argue here that this conclusion is misleading, especially for readers without expertise in infant memory. In this commentary, we seek to articulate the limitations of Sen and Gredebäck's conclusion while providing a broader context necessary to understand why their claims are misguided and unsupported by extant data.

In the following sections, we provide an overview of the measures used in the mobile paradigm to assess infant memory. We also address Sen and Gredebäck's (2022) critique regarding the use of a learning criterion and the operationalization of memory. We then discuss flaws in the simulation analyses and assumptions, and highlight converging evidence of the fundamental principles of infant memory development that have been documented using the mobile paradigm. We conclude by making recommendations for future approaches to open science

Abbreviation: VRM, visual recognition memory

This commentary was written in tribute to Carolyn Rovee-Collier (1942–2014) as a humble effort to highlight some of the issues she would have raised, although her retort would have been more poignant and eloquent! Carolyn never assumed “young infants could learn or remember,” rather it was an empirical question that carefully designed empirical research addressed.



collaborations that are designed to address the scientific rigor of existing literature.

UNFOUNDED PREMISE

Sen and Gredebäck (2022) begin by questioning the robustness of the mobile memory literature, stating "... statistically significant results appear consistently both within and across studies ... while null or contradictory findings are rare (figure 1b). In the present paper, for instance, 99% of the results were statistically significant in the screened articles..." (pp. 1–2). While evidence supports the first part of this statement, the second portion is misleading; approximately half of the analyses in figure 1b indicate significant forgetting (retention ratios). In fact, there are *many* non-significant findings in the mobile literature. For example, when infants are tested over long delays, or with stimuli that differ from those encountered at the time of original training, infants provide no evidence of memory. We find it remarkable that the authors have excluded all studies using the mobile paradigm in which infants were tested with a novel mobile based on the proviso that these studies involved generalization, discrimination, and categorization. In our view, this is a fatal flaw for two reasons. First, testing infants with a novel mobile is a *memory* problem. Second (and more importantly), a failure to consider these studies obscures the fact that changes in the independent variables (e.g., test stimulus, context, delay, and age) are what drives the outcome of these studies—not choices regarding the way in which the dependent variables have been operationalized. Putting this fundamental problem to one side, we go on to address their concerns.

USE OF A LEARNING CRITERION

Sen and Gredebäck's (2022) critique of the mobile conjugate reinforcement procedure rests entirely on their objection to the use of a learning criterion as a basis for participant inclusion. Their criticism of this practice reflects a fundamental misunderstanding of the relation between learning and memory. In order to remember (or, for that matter, to forget), one has to learn in the first place. For this very reason, specification of a learning criterion in studies of memory is quite commonplace. If a research question pertains to memory or retrieval processes, then establishing a minimum behavioral criterion for learning is essential. By way of a simple illustration, suppose that a group of 6-month-old infants were trained on the mobile task and their memory was assessed after a 4-month retention interval. At the time of the test, none of the infants exhibited performance indicative of memory. What can we conclude? Did the infants forget over the 4-month interval, or did they fail

to learn in the first place? Employing a learning criterion establishes whether learning did indeed occur. In the mobile paradigm, the *learning criterion* establishes that the infant has learned the contingency between foot kicking and mobile movement. The learning criterion is operationally defined as kicking at 1.5 times the baseline kick rate (i.e., the kick rate prior to the introduction of the contingency) in 2 of any 3 consecutive minutes during the acquisition phase. Infants who do not meet this learning criterion are excluded from further analysis of memory and forgetting because there is no evidence that they learned in the first place. Furthermore, exclusion on this basis is infrequent. However, Sen and Gredebäck make several claims about why this practice yields biased samples and false positives. As we will show below, none of their claims are supported by the actual data.

There is no evidence of “researcher degrees of freedom” in the mobile literature

Although we (and others) find considerable merit in the use of a learning criterion when studying memory (note that training to a criterion has been used in the infant-control habituation procedure since the early 1970s; Colombo & Mitchell, 2009), we also recognize that it is still fair to ask whether the use of this criterion has biased the nature of the findings in the mobile task. In order to answer this question, we reviewed the 77 papers using the mobile task that was included in Sen and Gredebäck's (2022) critique (see <https://osf.io/P7RYX>). Opportunistic use of researcher degrees of freedom is typically defined as different criteria across different studies. Of the 77 papers, 65 (reflecting just under 200 experiments) explicitly stated that the criterion was a kick rate of 1.5 times the baseline kick rate. In other words, across these 65 papers, starting from about 1984 onward, researchers consistently applied the exact same learning criterion. Consistent use of the same criterion provides no evidence of an opportunistic use of researcher degrees of freedom. Further, and particularly relevant to Sen and Gredebäck's argument, the average number of participants excluded for failing to meet the learning criterion across these 65 papers was only 8.4%. Only experiments with infants 6 months of age and younger trained on the mobile task were included in this examination, as this is the bulk of the work critiqued. [Note that one paper reported the overall number of exclusions across 5 experiments (i.e., 4), rather for each individual experiment, so that paper was not included in obtaining these estimates.] In fact, there were 63 individual experiments (32% of the total) that had zero exclusions due to the learning criterion. Of the remaining 12 papers, nine did not state an explicit learning criterion nor did they report excluding any infants on the basis of a failure to learn. In the other three papers, a total of nine infants were excluded for a failure to learn even though

the explicit learning criterion was not stated. Taken together, our review of the 77 papers illustrates that when articulated, the learning criterion was identical across papers, and the number of infants who were excluded due to a failure to meet the learning criterion was remarkably small, making Sen and Gredebäck's primary point moot.

OPERATIONALIZATION OF MEMORY

Sen and Gredebäck (2022) also defined researcher degrees of freedom in terms of choosing methodological practices and analyses that produce biased results for an entire literature. For the mobile paradigm, they point to the use of a learning criterion as well as the retention measures and analyses as influencing findings, and use this line of reasoning as the impetus for their simulation analyses. For example, they state that "...mobile paradigm studies that focus on memory development use a very particular way of operationalizing memory measures (e.g., baseline ratio) and what is considered evidence for an infant's ability to remember past events." (p. 5). We agree that the way in which the dependent variables are operationalized will influence our conclusions; indeed, this fact applies to all experimental studies of infant memory. The way in which we define what constitutes evidence of memory will determine when we conclude whether an infant has remembered or forgotten. However, such critiques do not account for systematic patterns of null findings (control groups and experimental manipulations) in the mobile literature (i.e., limits of infant memory using the same approach) or the correspondence of the findings from mobile procedure to those from other infant memory paradigms.

Sen and Gredebäck (2022) also state that "...only infants who increase kicking from baseline to acquisition are included in any analysis, creating by default a condition in which all statistical tests are run on infants with a relatively high kicking rate when connected to the mobile" (p. 4). Although the first part of the sentence is correct, the second is not. What they fail to acknowledge is that kick rates vary dramatically across infants (e.g., Hartshorn et al., 1998; Merz et al., 2017). While some infants have relatively high kick rates during baseline and acquisition, others show relatively low rates.

They also express concerns that the learning criterion results in low-variance samples in the mobile studies (see Sen & Gredebäck, 2022, p. 12). In fact, studies with the mobile paradigm have relatively high variance, particularly in comparison to studies of imitation where the range is typically 0–3 or 0–4 actions. Indeed, the standard error bars of actual raw kick rates indicate that scores vary dramatically across infants. All of this begs the point relative to deferred imitation. In most imitation studies, infants are shown a small set of actions (typically 3–4). Memory is inferred relative to either a baseline control group (deferred imitation) or relative to the

infant's own behavior prior to the demonstration (elicited imitation). Memory is typically inferred on the basis of a difference of 0.5 to 1.5 actions.

SIMULATION RESULTS

We now consider the simulation analyses that Sen and Gredebäck (2022) use as evidence for opportunistic use of researcher degrees of freedom in the mobile literature. The simulation proceeds by generating a distribution for infant baseline kicking, and then generates the same distribution once for each of the 9 min of acquisition (explicitly with no provision for learning; it is exactly the same distribution). When drawing a baseline and comparing it to acquisition as specified by the learning criterion, they find that the simulation produces a significant fraction of the set of simulated infants in each iteration that meets the learning criterion. They then draw multiple random samples from the subset of baseline ratios that meet the learning criterion and test them, with the outcome that a substantial fraction of these samples are significant, supporting their contention that the criterion causes researchers using the approach to conclude that infants are learning and remembering when no learning is actually taking place.

Faulty assumptions

The simulation is based on a set of unjustifiable assumptions. First, the distributions generated are either normal or uniform (level). Neither of these is likely to be a reasonable approximation of baseline kicking activity. As kicking is a 'count' data type that is bounded at zero with an expectation of a rightward (positive) skew, a Poisson distribution appears the most reasonable distribution for a model, and applying a Poisson distribution appears to have a significant impact on the model's output (see below). Additionally, generating a normal distribution with a negative component, which the authors then reflect back into positive territory by taking the absolute value, will result in a potentially non-normal distribution. The authors' choice of a mean of 10, standard deviation (SD) of 3 likely reduces this concern, but those choices can be questioned. The simulation learning criterion (2 consecutive minutes) also differs from the criterion used in the mobile literature (2 out of any 3 consecutive minutes).

Questions regarding choice of parameters

The authors chose a distribution with a mean of 10 and a SD of 3 to model the baseline data, and then generated the same distribution for each of the 9 min of acquisition (table 2 lists output for alternative choices of the SD).



Few data are offered to justify these parameters, and re-running their simulation with lower means (likely a better choice, given our collective experience and familiarity with published papers in this literature) is not discussed. Lower means can lower the outcome in terms of (spuriously) significant random samples across the comparison of the baseline to the 9 min of acquisition, in particular if a distribution without a significant tail on the lower end is used (altering the mean and SD together also may have significant effects on the data).

Number of simulation teaching criterion varies

Additionally, the authors indicate that “For the simulations in which the learning criteria were implemented, the infants who did not meet the criteria were excluded after the simulation of the acquisition phase was completed” (Sen & Gredebäck, 2022, p. 7). They make no further comment on this issue, which equates to an assumption that the “drop” rate equals that of actual experiments (approximately 8% by Sen and Gredebäck's estimate). Their “keep” rate (meaning that the ‘infant’ met the criterion for learning), however, for uniform distributions, appears to be approximately 40%–45%. In other words, Sen and Gredebäck drop 55%–60% of their simulated samples, a loss rate that is far higher than is the case in actual experiments with infants. For normal distributions ($M=10$, $SD=3$), it appears that the “keep” rate is even lower, approximately 24% or less, meaning that even more simulated “infants” are dropped for failing to learn. The authors might assert that *all* infants should fail to learn in their simulation, but some level of spurious activity exceeding criterion is likely. The fact that their *t*-tests are based on data following a substantial removal (and thus, a highly selective sample of simulated cases) will skew the outcome. Tests of the model with a Poisson distribution applied (mean of seven), with the caveat that a zero baseline is not permitted, result in the number of “kept” learning ratios declining to 4%–5%, absent any learning (that is, using only a baseline distribution). If a Poisson distribution is the most accurate means of modeling these data, then in the simulation (with no learning), the vast majority of simulated infant responses fail to meet the criterion, as compared to a real-life data set, in which the vast majority (with learning applied) will meet the learning criterion. Thus, the structure of the simulation appears to (inappropriately) produce quite high drop rates because many of the simulated ‘infants’ did not meet the learning criterion, which does not match actual infant data. More generally, this means that the simulation is highly sensitive to both the choice of distribution and to the choice of parameters. Without a substantial exploration of the actual infant

data that are being modeled, the assertions made are at least highly questionable and may be simply incorrect.

Simulation design issue

There appears to be an even more central issue with the design of the simulation. Sen and Gredebäck (2022) are taking a normal (or uniform) distribution as baseline, and then another such distribution as learning, and comparing the two by randomly choosing a baseline to compare to a (randomly) chosen pair of values for each 2 min of acquisition. This is equivalent to treating these two distributions as independent samples. This is a problem; it is clear that the infant baseline and infant acquisition data are correlated, even disregarding any learning, because *their source is the same infant participant*. For example, infants who kick at high frequencies are likely to kick at high frequencies across both baseline and acquisition phases. Thus, when applying the learning criterion, the data essentially represent a within-subjects design. There appears to be no attempt to include in the model the expected correlation between baseline and acquisition scores. This issue alone might cause the extent of random matches meeting the 1.5 criterion to be much higher than is the case in the actual infant data.

The simulation focuses solely on the likelihood of obtaining false positive results for the memory retention measures. Most mobile studies include control conditions that, as expected, showed no memory effects (i.e., long-term retention tests were not significantly above baseline); as we understand the computational model here, these should have been included in the simulations, and would have affected the predicted probabilities if in fact no learning was actually taking place. Along with the lack of consideration of these control conditions, the simulation fails to account for replication of findings across multiple studies, and systematic changes in infant performance (i.e., shift from significance to non-significance) across various manipulations (e.g., increasing retention interval). It is unclear to us how conclusions about false positives can be derived in the absence of consideration of these points. If there were a high degree of false positives in the paradigm because of the use of a learning criterion and baseline ratio, then there should also be a roughly equivalent number of false positives in control conditions that were included across numerous studies.

Overall, the determination of which parameters to include in these models should be based on actual data, and such models can lead to affirmation of hypotheses and new testable predictions. However, using parameters that are not based upon the actual behavior of live participants to simulate behavioral outcomes, and then using such outcomes to criticize decades of data from carefully implemented experiments is problematic.

SIMILARITIES ACROSS INFANT MEMORY PARADIGMS

The ultimate question for the field of infant memory development is whether the mobile procedure represents a “unique way of measuring memory” that yields conclusions that are different from those obtained using other paradigms. A careful analysis of the experimental literature demonstrates that it does not.

Use of learning criteria across memory paradigms

Establishing a learning criterion, and using that criterion as the basis of the exclusion of participants, is by no means unique to the mobile paradigm. Such exclusion criteria can be found in a variety of other areas of research. In infant-controlled habituation tasks, infants' memory is evaluated relative to their terminal level of looking (Colombo & Mitchell, 1990) after they have attained a habituation criterion derived from some aspect of their initial looking. For instance, “learning” criteria are often used in the visual recognition memory (VRM) task. In these studies, infants are shown an object, and then later shown that object along with a novel object. Performance in the task is based on a young infant's natural tendency to preferentially look at a novel object. In some of these studies, the initial object is shown to the participant until they have accumulated a predetermined amount of looking time directed at that object (e.g., Robinson & Pascalis, 2004), while in others, the initial object is shown for a preselected amount of time (e.g., 20 s) and if an infant does not accumulate a pre-set amount of time looking at it (e.g., 5 s); then, they are excluded from subsequent analysis (e.g., Barr et al., 2014; Courage & Howe, 1998; Jones et al., 2011). This most definitely qualifies as a learning criterion for inclusion. After a learning criterion for inclusion is established, a separate memory test is conducted that is relative to meeting the learning criterion. In the VRM paradigm, the memory measure is indexed as looking time to a novel stimulus that is significantly above 50%. The assumption is that, during the initial learning phase, the infant encoded the image and at the time of test remembered the training item as familiar and therefore looked significantly longer at the novel item demonstrating a significant novelty preference. Also like the mobile paradigm, there is evidence of forgetting across time when after longer delays infants move to a null or even a familiarity preference.

The VRM task is identical in principle to the Novel Object or Object Recognition tasks that are used extensively in nonhuman subjects (Ennaceur & Delacour, 1988). Many studies of object recognition in nonhuman animals

employ a criterion regarding the amount of time that subjects must interact with, or explore, the object (i.e., a learning criterion) in order for subjects to be included in the analysis of memory (e.g., Cohen et al., 2022; Gaskin et al., 2010; Jablonski et al., 2013; Shimoda et al., 2021). It is also not uncommon in studies of operant conditioning in nonhuman animals (e.g., learning the contingency between some response and an appetitive outcome; akin to the mobile paradigm) for a learning criterion to be applied (e.g., must correctly respond on 80% of choices or responses; Boulougouris et al., 2007; Schoenbaum et al., 1999; Šlipogor et al., 2022). Furthermore, in tasks that assess more complex cognitive processes that require sequential phases of training, criteria for moving from one to the next phase are often set; failure to meet that criterion results in either extended training until the criterion is met, or else the subject is dropped from the experiment (e.g., Alamy et al., 2005; Howe & Courage, 1997; Overman et al., 1992). Similarly, in the deferred imitation paradigm, “memory” is inferred relative to either the infant's own baseline or the average baseline of a control group. The underlying rationale for all of these experimental choices are akin to those used in the mobile task; that is, it is impossible to study memory or forgetting until you have first established that learning has taken place.

Employing a learning criterion is especially important when examining how memory, and memory-related processes, change across development. For example, as noted by Courage and Howe (2004) “The question of developmental changes in long-term retention of information has been more difficult to address. Indeed, the methods that have been used to study infant memory contain a potentially serious threat to the validity of conclusions about development—namely, the failure to control for the effect of age differences in initial learning” (p. 11). Courage and Howe note that the work with the mobile procedure is an exception to this serious threat. Given this, we are surprised that Sen and Gredebäck (2022) consider this fundamental strength of the mobile procedure to be a methodological flaw.

Converging evidence across infant memory paradigms

Sen and Gredebäck (2022) argue that “Based on our results, we suggest that memory findings in the mobile paradigm literature present a case of scientific endemism, in which the phenomenon under investigation exists only within the ecosystem created by a specific methodological protocol” (p. 13) and that “Scientists should look for independent sources of evidence (e.g., different research groups, different paradigms, replications), in order to protect the theory from researcher bias.” (p. 14). We find these conclusions to be totally unfounded. What the authors fail to acknowledge is that all of the data with the



mobile paradigm with 6-month-olds has been replicated in studies of deferred imitation and preferential looking with infants of the same age and map to adult memory principles as well (Rovee-Collier, 1997). For example, Gross et al. (2002) tested 6-month-olds on a VRM task, a deferred imitation task, and a mobile task and found evidence of learning and memory across all three paradigms in the same infants. Furthermore, researchers have demonstrated that infants rapidly forget across time in deferred imitation (e.g., Barr & Hayne, 2000) and in the VRM paradigm (e.g., Fagan, 1970; Morgan & Hayne, 2006), just like what has been reported in the mobile task.

In a *Developmental Review* paper, Hayne (2004) directly compared the results from the most common memory paradigms (VRM, deferred imitation) with the mobile conjugate reinforcement paradigm highlighting key parallels across the paradigms. The following four principles of memory development replicate across all three paradigms:

1. Older infants encode information faster than younger infants.
2. Older infants retain information for longer durations than younger infants.
3. Memory retrieval is specific to the cues present at the time of original encoding.
4. Older infants are increasingly able to exploit retrieval cues and apply their knowledge to more situations.

STRONG ASSERTIONS

In our view, Sen and Gredebäck's (2022) failure to acknowledge the aforementioned evidence from the broader infant memory literature is problematic. In providing a historical context of infant memory and the mobile paradigm, they refer to "...critical voices from the outside (Bauer, 1996; Bauer et al., 2007; Millar & Weir, 2015; Pomerleau et al., 1992; Schacter & Moscovitch, 1984) had little impact on the workings within the ecosystem" (p. 13). Although the infant memory literature has been full of lively debates regarding the "types of memory" measured by various memory paradigms (e.g., implicit vs. explicit memory) and manipulations, the scientific rigor of the mobile paradigm, its methodology, and analyses, has never constituted one of those debates.

Sen and Gredebäck (2022) also express concern that only a small group of researchers have used this traditional methodology over decades which, according to them, has created a unique way of measuring memory in the mobile paradigm that is different from the other paradigms measuring memory, such as deferred imitation and preferential looking. If the definition of a "small group of researchers" includes a senior academic whose students subsequently go on to use the

same paradigm, then this concern would also apply to other groups who fit this definition. In the case of deferred imitation, for example, there are probably three senior (i.e., approaching retirement) researchers who are prominent in the field. Of these, the bulk of the most recent research has been conducted by these scholars and their students and their students' students. This is how science often works, no matter what the field. In our view, the number of researchers who use a particular paradigm is much less important than the rigor with which the paradigm is used, no matter whose hands it passes through.

CONCLUSIONS AND FUTURE DIRECTIONS


We fully recognize the importance of the open science framework and meta-analytic approaches to science. Scientific dialogue between groups of experts can, when orchestrated appropriately, most definitely improve the quality of our science. Respectful and objective feedback and critique regarding our underlying assumptions, the broader literature, and the specifics of the paradigm of interest are all essential steps for strengthening the products of our scientific inquiry, including experimental design, behavioral data analysis, meta-analysis, or simulation models. However, in order for this to occur, the dialogue needs to be open from the beginning and informed by the extant literature. In this specific case, the Sen and Gredebäck (2022) manuscript would have certainly benefited from discussions with researchers with expertise in using the mobile paradigm to assess memory as well as with those with expertise in other learning and memory paradigms. Future studies should adopt open science collaborative practices.

In conclusion, as Sen and Gredebäck (2022) note, one key characteristic of the mobile paradigm literature is that across decades, researchers have faithfully followed a particular methodological protocol. Rather than seeing this as a weakness, we see it as one of the fundamental strengths of the paradigm. In our view, one key feature of all rigorous research is that it faithfully follows a particular methodological protocol. The consistency with which the mobile paradigm has been used has made it possible to isolate the impact of a particular independent variable (e.g., age, delay, context, stimulus, and number of learning trials) Furthermore, in a field where replication is remarkably rare, the basic findings obtained with the mobile paradigm have been replicated over and over again *under conditions that allow direct comparison across laboratories and experiments*. These basic principles of memory development have been cemented in the literature due to the careful way in which researchers have followed this particular methodological paradigm.

ORCID

Kimberly Cuevas  <https://orcid.org/0000-0003-1811-1131>

Scott Adler  <https://orcid.org/0000-0001-6795-1239>

Peter Gerhardstein  <https://orcid.org/0000-0002-4500-5150>

REFERENCES

- Alamy, M., Errami, M., Taghzouti, K., Saddiki-Traki, F., & Bengelloun, W. A. (2005). Effects of postweaning undernutrition on exploratory behavior, memory and sensory reactivity in rats: Implication of the dopaminergic system. *Physiology & Behavior*, *86*, 195–202.
- Barr, R., & Hayne, H. (2000). Age-related changes in imitation: Implications for memory development. In C. Rovee-Collier, L. P. Lipsitt, & H. Hayne (Eds.), *Progress in infancy research* (Vol. 1, pp. 21–67). Erlbaum.
- Barr, R., Walker, J., Gross, J., & Hayne, H. (2014). Age-related changes in spreading activation during infancy. *Child Development*, *85*, 549–563. <https://doi.org/10.1111/cdev.12163>
- Boulougouris, V., Dalley, J. W., & Robbins, T. W. (2007). Effects of orbitofrontal, infralimbic and prelimbic cortical lesions on serial spatial reversal learning in the rat. *Behavioural Brain Research*, *179*, 219–228. <https://doi.org/10.1016/j.bbr.2007.02.005>
- Cohen, S. J., Cinalli, D. A., Ásgeirsdóttir, H. N., Hindman, B., Barenholtz, E., & Stackman, R. W. (2022). Mice recognize 3D objects from recalled 2D pictures, support for picture-object equivalence. *Scientific Reports*, *12*, 4184. <https://doi.org/10.1038/s41598-022-07782-4>
- Colombo, J., & Mitchell, D. W. (1990). Individual differences in early visual attention: Fixation time and information processing. In J. Colombo & J. W. Fagen (Eds.), *Individual differences in infancy: Reliability, stability, prediction* (pp. 193–227). Lawrence Erlbaum Associates, Inc.
- Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*, 225–234. <https://doi.org/10.1016/j.nlm.2008.06.002>
- Courage, M. L., & Howe, M. L. (1998). The ebb and flow of infant attentional preferences: Evidence for long-term recognition memory in 3-month-olds. *Journal of Experimental Child Psychology*, *70*, 26–53. <https://doi.org/10.1006/jecp.1998.2444>
- Courage, M. L., & Howe, M. L. (2004). Advances in early memory development research: Insights about the dark side of the moon. *Developmental Review*, *24*, 6–32.
- Cuevas, K., & Davinson, K. (2022). The development of infant memory. In M. L. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (3rd ed., pp. 31–59). Routledge. <https://doi.org/10.4324/9781003016533>
- Ennaceur, A., & Delacour, J. (1988). A new one-trial test for neurobiological studies of memory in rats. I: Behavioral data. *Behavioural Brain Research*, *31*, 47–59. [https://doi.org/10.1016/0166-4328\(88\)90157-x](https://doi.org/10.1016/0166-4328(88)90157-x)
- Fagan, J. (1970). Memory in the infant. *Journal of Experimental Child Psychology*, *9*, 217–226.
- Gaskin, S., Tardif, M., Cole, E., Piterkin, P., Kayello, L., & Mumby, D. G. (2010). Object familiarization and novel-object preference in rats. *Behavioural Processes*, *83*, 61–71.
- Gross, J., Hayne, H., Herbert, J., & Sowerby, P. (2002). Measuring infant memory: Does the ruler matter? *Developmental Psychobiology*, *40*, 183–192. <https://doi.org/10.1002/dev.10020>
- Hartshorn, K., Wilk, A. E., Muller, K. L., & Rovee-Collier, C. (1998). An expanding training series protracts retention for 3-month-old infants. *Developmental Psychobiology*, *33*, 271–282.
- Hayne, H. (2004). Infant memory development: Implications for childhood amnesia. *Developmental Review*, *24*, 33–73. <https://doi.org/10.1016/j.dr.2003.09.007>
- Howe, M. L., & Courage, M. L. (1997). Independent paths in the development of infant learning and forgetting. *Journal of Experimental Child Psychology*, *67*, 131–163.
- Jablonski, S. A., Schreiber, W. B., Westbrook, S. R., Brennan, L. E., & Stanton, M. E. (2013). Determinants of novel object and location recognition during development. *Behavioural Brain Research*, *256*, 140–150.
- Jones, E. J. H., Pascalis, O., Eacott, M. J., & Herbert, J. S. (2011). Visual recognition memory across contexts. *Developmental Science*, *14*, 136–147. <https://doi.org/10.1111/j.1467-7687.2010.00964.x>
- Merz, E. C., McDonough, L., Huang, Y. L., Foss, S., Werner, E., & Monk, C. (2017). The mobile conjugate reinforcement paradigm in a lab setting. *Developmental Psychobiology*, *59*, 668–672. <https://doi.org/10.1002/dev.21520>
- Morgan, K., & Hayne, H. (2006). The effect of encoding time on retention by infants and young children. *Infant Behavior & Development*, *29*, 599–602. <https://doi.org/10.1016/j.infbeh.2006.07.009>
- Overman, W., Bachevalier, J., Turner, M., & Peuster, A. (1992). Object recognition versus object discrimination: Comparison between human infants and infant monkeys. *Behavioral Neuroscience*, *106*, 15–29.
- Robinson, A. J., & Pascalis, O. (2004). Development of flexible visual recognition memory in human infants. *Developmental Science*, *7*, 527–533.
- Rovee-Collier, C. (1997). Dissociations in infant memory: Rethinking the development of implicit and explicit memory. *Psychological Review*, *104*, 467–498. <https://doi.org/10.1037/0033-295X.104.3.467>
- Schoenbaum, G., Chiba, A. A., & Gallagher, M. (1999). Neural encoding of orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *Journal of Neuroscience*, *19*, 1876–1884. <https://doi.org/10.1523/JNEUROSCI.19-05-01876.1999>
- Sen, U., & Gredebäck, G. (2022). Methodological integrity assessment in the mobile paradigm literature: A lesson for understanding opportunistic use of researcher degrees of freedom in psychology. *Child Development*. <https://doi.org/10.1111/cdev.13850>
- Shimoda, S., Ozawa, T., Ichitani, Y., & Yamada, K. (2021). Long-term associative memory in rats: Effects of familiarization period in object-place-context recognition test. *PLoS One*, *16*, e0254570. <https://doi.org/10.1371/journal.pone.0254570>
- Šlipogor, V., Graf, C., Massen, J. J. M., & Bugnyar, T. (2022). Personality and social environment predict cognitive performance in common marmosets (*Callithrix jacchus*). *Scientific Reports*, *12*, 6702. <https://doi.org/10.1038/s41598-022-10296-8>

How to cite this article: Cuevas, K., Adler, S. A., Barr, R., Colombo, J., Gerhardstein, P., Hayne, H., Hunt, P. S., & Richardson, R. (2023). Commentary on the scientific rigor of Sen and Gredebäck's simulation: Why empirical parameters are necessary to build simulations. *Child Development*, *00*, 1–7. <https://doi.org/10.1111/cdev.14062>