Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Behavioral Sciences

How massive online experiments (MOEs) can illuminate critical and sensitive periods in development

Joshua K Hartshorne

The fact that children are far more likely to successfully acquire a variety of new skills and knowledge than are adults is so clearly evidenced in everyday life that it hardly needs scientific confirmation. However, despite four decades of intensive research, the reason *why* remains controversial. In fact, the terms of the debate about critical periods in language have hardly changed since the 1960s. I argue that this is because the standard in-lab research paradigms that have otherwise served psychology well are fundamentally ill-suited to the study of critical and sensitive periods. In particular, this research requires samples that are far more diverse and orders of magnitude larger than can be achieved in the lab. I show that massive online experiments provide an exciting and productive alternative.

Address
Department of Psychology, Boston College, United States

Current Opinion in Behavioral Sciences 2020, **36**:xx-yy

This review comes from a themed issue on **Sensitive and critical periods**

Edited by **Willem Frankenhuis** and **Cate Hartley**

https://doi.org/10.1016/j.cobeha.2020.09.005

The fact that children are far more likely to successfully acquire new skills and knowledge than are adults is so clearly evidenced in everyday life that it hardly needs formal scientific confirmation. Whether these 'critical periods' are due to a difference of ability or circumstance is less obvious. Note that for simplicity, I will not distinguish between 'critical', 'sensitive', or 'optimal' periods (etc.); instead, I use 'critical' throughout. These distinctions are controversial and will have no bearing on the present discussion.

While there has been phenomenal progress in understanding critical periods in perceptual processes in animals [1••] and, to a lesser extent, in humans [2••,3] progress on understanding higher-level cognitive functions has been frustratingly slow [4••,5]. In the case of second-language acquisition — by far the most emphatically studied of these phenomena — researchers remain as divided today as they were 40 years ago, and for roughly the same reasons [6–9]. This reflects not stubbornness on the part of researchers but rather a limitation of the empirical literature: many key questions have proven difficult to test, and those that have been tested have provided contradictory answers [5••].

Given five decades of intense focus on this question by some of our field's greatest minds, the slow progress is humbling, or even discouraging. Below, I argue that progress has been slow because the standard experimental paradigms of psychology are singularly ill-equipped to shed light on age-related changes in learning abilities. Next, I show that massive online experiments address many — not all — of these limitations, and thus hold the potential for unprecedented progress. I then present a case study and conclude with a discussion of the limitations of massive online experiments and future directions.

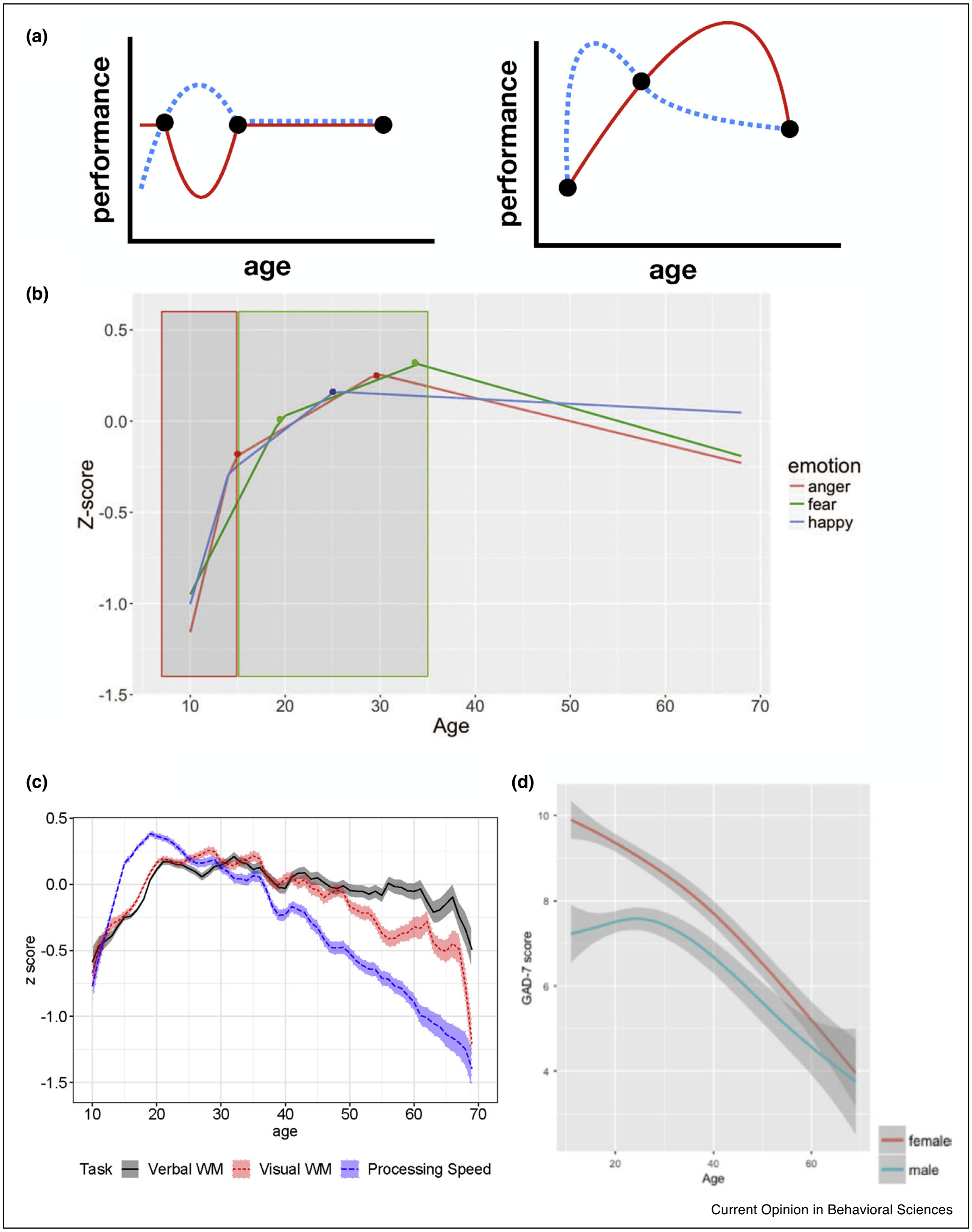
The problem(s) with studying critical periods in the lab

The ideal study would precisely measure learning over a wide range of ages from a representative group of subjects using an ecologically-valid task. None of these desiderata are easy to meet in human research labs.

Most studies focus on a handful of ages. This helps keep total subject numbers manageable. Moreover, different ages require different recruitment strategies, with most laboratories having the expertise and resources for only a few. Unfortunately, such studies provide very little constraint on theory (Figure 1a). This problem is compounded by the overreliance on college students as one of these groups. College students are already in decline on some cognitive abilities and still improving on others [10,11,12,5••], a fact which can muddy direct comparisons with older adults (Figure 1).

This is not merely a theoretical concern. For instance, early studies comparing college students with older adults found no evidence of age-related decline in social cognition [13]. This proved to be an artifact of the sampling method: the peak in social cognition lies in middle age [11•,10,14]. Similarly, critical periods researchers long assumed that college students had fully acquired grammar — an assumption that proved false, and which has demonstrably skewed results [5••]. More broadly, recent large-scale online studies have revealed a number of theoretically important results that could not be easily captured by sampling only a few ages (Figure 1b–d).

Figure 1



Current Opinion in Behavioral Sciences

Measuring at three time points can mask important differences. **(a)** illustrates how two very different developmental trajectories can be consistent with the same three measurements. Looking at only three time points could result in missing important developmental change (*left*) or misunderstanding when it happens (*right*). Panels b–d illustrate theoretically important effects that would be difficult to detect by sampling only a few ages. **(b)** shows clearly different developmental trajectories for sensitivity to anger, fear, and happiness in facial expressions ([10], $N = 9546$). **(c)**. Processing speed (as measured by digit symbol coding) peaks earlier and more sharply than visual and verbal working memory (WM) ([11•] = 10,394). **(d)** Generalized Anxiety Disorder symptoms decline with age for both men and women, but nonetheless on distinct trajectories ([12], $N = 7176$). Panels b and d are reprinted from the originals, with permission. Panel c was created from the raw data.

Note that while developmentalists may be interested in shorter time windows, this does not necessarily mitigate the issue: developmentalists are usually interested in characterizing changes that happen on the order of months, which requires densely sampling ages.

Testing more age groups would be helpful, but is complicated by the fact that most studies already test orders of magnitude too few subjects per age group. For instance, a common question in the language literature is, 'What is the oldest age at which one can start learning a language and still eventually achieve native-like proficiency?' It can be shown mathematically that obtaining statistically meaningful results using standard instruments (i.e. language quizzes) requires thousands of subjects, whereas the typical study comprises merely dozens [5••,15••]. This has been less obvious than it should have been because researchers have not typically included error bars for the critical analyses. Through simulation, Hartshorne *et al.* [5••] showed that for most studies, the error bars would include most or all of the available range. This was true not just for the aforementioned question, but for a range of analyses commonly reported in the literature.

As a result, even relatively large studies are undersized by at least an order of magnitude. For instance, Salthouse [16] cross-sectional results from 2350 individuals (ages 18–60), tested on a battery of 12 tasks. Despite being one of the largest studies to date, the results proved too coarse-grained to reveal effects captured by larger online studies, such as the earlier peak and decline for processing speed than for memory (compare Figures 1c and 2).

This lack of precision is not unique to the critical periods literature [17•], and has predictable effects on replicability. Mathematical analysis taking into account typical statistical power suggests that more than half of findings in the literature are false positives — a prediction with increasingly strong empirical support [18,19••,20]. Laboratory studies face other limitations beyond simply obtaining enough data. Bringing subjects to the lab inevitably skews subject demographics towards people who happen to live near the lab. Subject demographics are highly skewed towards the high-SES and well-educated, with psychology majors at North American research universities make up the bulk of all research subjects — all with predictable consequences for generalizability [21,22].

Another challenge is ecological validity. Of particular relevance to critical period research, learning and development typically unfolds over the course of months if not years, whereas the typical laboratory experiment lasts less than an hour. This seemingly bland fact can confound research in unexpected ways. Early laboratory studies into how language learning ability changes with time found counterintuitively that adults learned far more rapidly than young children [7]. This was shown not to be an artifact of laboratory testing: in the 'wild', adults show superior learning during the initial *months* of acquisition [7]. Thus, it is unclear how learning observed in the lab relates to the original phenomenon of interest: learning out in the world. Thus, researchers on "natural experiments", studying individuals who began learning languages at different ages. One can then estimate learning rates by comparing the linguistic knowledge of learners by repeatedly measuring the same learner over time or by comparing different learners who have been learning for different amounts of time. While such natural experiments have obvious advantages in ecological validity, they present an enormous logistical problem in that they require a great deal of data. Thus, researchers have mostly focused on a methodologically simpler problem of determing the greatest level of proficiency that can be achieved by learners who started at different ages [7,8,9,23]. While such measurements have considerable practical implications, they do not provide much insight into how learning ability changes with age, and thus say little about critical periods (Figure 3). Note that the same issues likely apply to the acquisition of any ability that takes a long time to learn, such as music or chess.

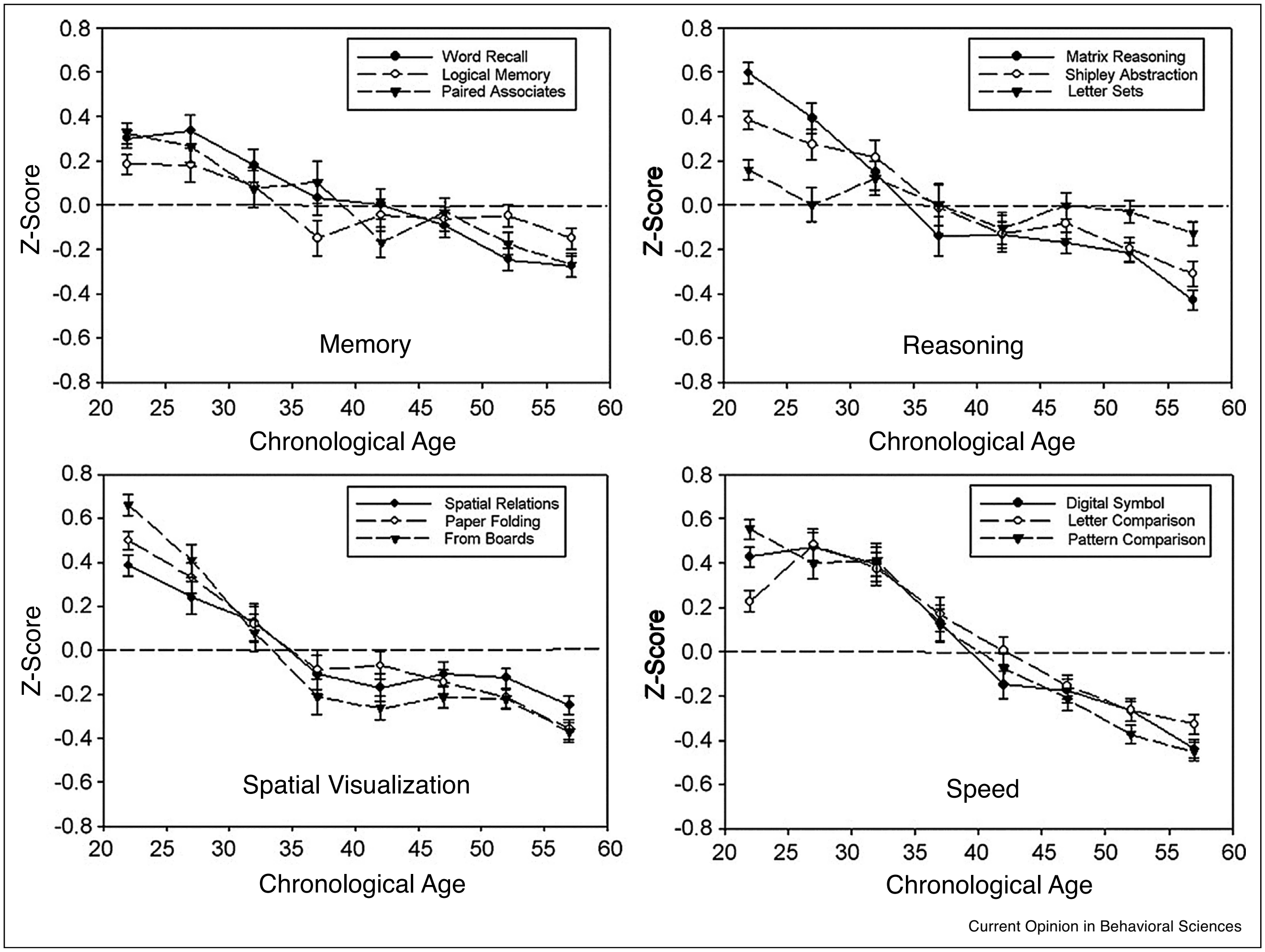
Note that all these issues — age range, sample size, representativeness, and ecological validity — apply whether measurement is cross-sectional, longitudinal, or both. Note also that while our review focuses on studies of humans, some of these same issues apply in animal studies [24,25].

The solution: massive online experiments?

Researchers now routinely collect datasets with tens or hundreds of thousands of subjects — and occasionally millions — by testing them over the Internet [26•• ,27,28]. Some studies adopt a 'citizen science' model, where subjects are volunteers donating their time to science through special-purpose apps (eBird, KidTak) or websites (gameswithwords.org, testmybrain.org), while others make use of 'naturally occurring' data, such as performance on games or interactions on social media. For a variety of reasons, these studies have **not** used Amazon Mechanical Turk or other online labor markets. Of particular relevance to the study of development, these platforms ban users under the age of 18.

Given that half the world's population has internet access [29], any study that can be run on a computer or mobile device can be run with nearly any demographic anywhere in the world, and in large numbers. This includes not just surveys, but studies involving grammatical judgments, reaction times, decision-making, economics games, eye-tracking, priming, sentence completion, skill acquisition, and even virtual reality — which is to say, most human behavioral experiments [26••,30•,31]. The broad demographic reach facilitates investigation of demographic

Figure 2



Current Opinion in Behavioral Sciences

An unusually large in-lab lifespan dataset, compiled from several studies ($N = 2350$). Figure reprinted from Salthouse [16], used with permission.

variation; testing outside the laboratory facilitates examination of ecologically-valid behavior; and the fact that subjects can participate on their own time without traveling to and from the lab facilitates studying behaviors that unfold over extended periods of time, such as learning [32–34,26••,27,5••].

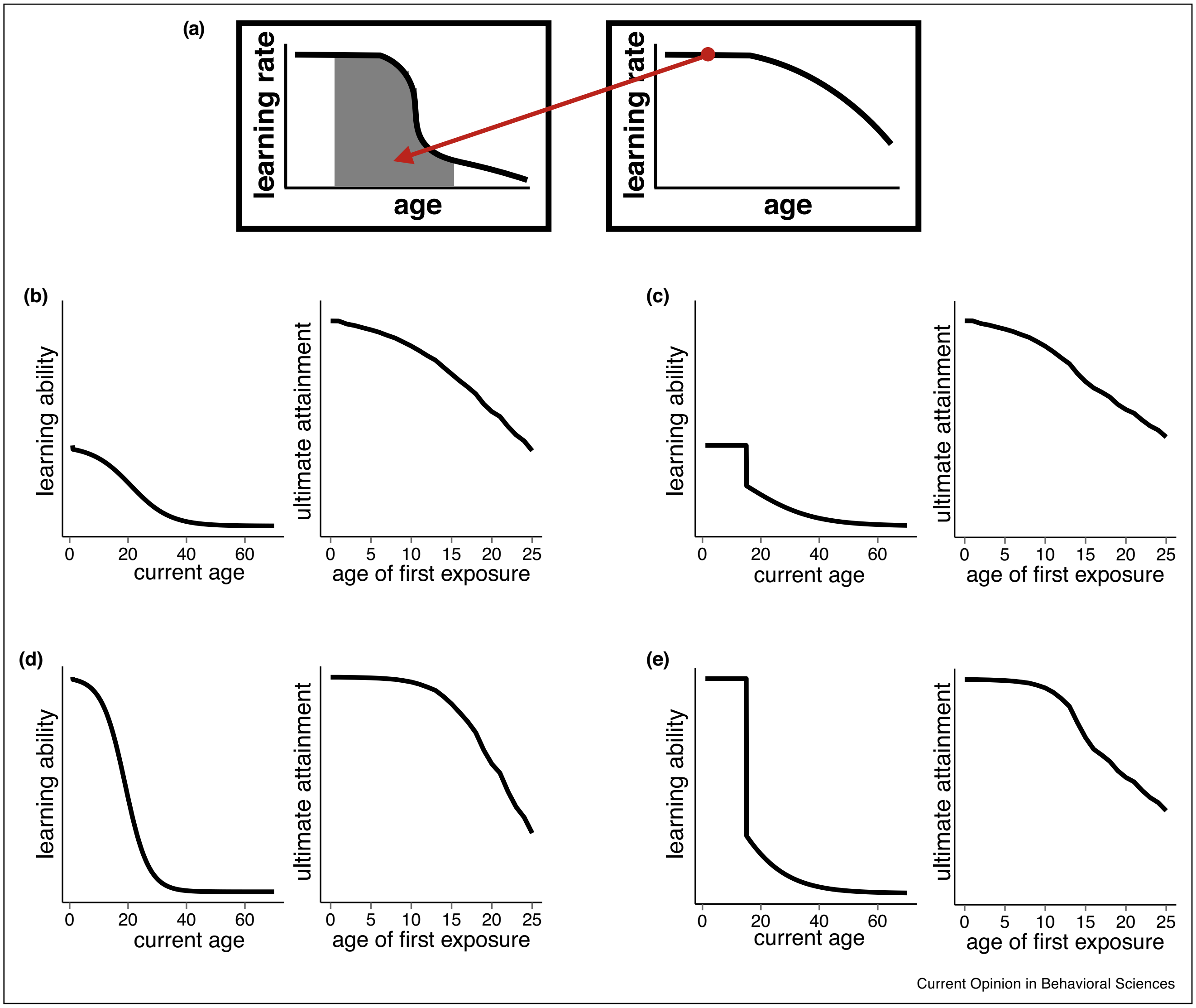
Critically, extensive research has shown that data from online studies is, if anything, higher-quality than what is typically achieved in the lab [26••]. In retrospect, this is not surprising. Unlike traditional subjects who must be enticed with extrinsic rewards, online volunteers participate purely for intrinsic motivation, and indeed MOEs are designed to be intrinsically motivating [30] or make use of naturally-occurring data [27]. Subjects who are not interested simply do not participate [35]. While differential dropout is certainly worth keeping in mind, it is easier to monitor in online studies than in traditional lab-based studies, where dropout usually happens before the subject comes in, as they decide whether to sign up or choose whether to no-show (cf. [35]).

MOEs are particularly well-suited for critical periods research because, by default, they recruit subjects from a wide range of ages. Not surprisingly, studying age-related change has been one of the most common uses (recently: [10,12,5••,11•,34,33,36]). As reviewed above, such studies have frequently revealed theoretically important findings that were missed in prior, smaller-scale studies (see also Figure 1b–d).

Studying critical periods with MOEs: An example

Hartshorne *et al.* [5••] henceforth, 'HTP' — present a study of age-related change in syntax learning that utilizes MOEs to address the challenges raised above. Specifically, we recruited 680 333 English-speakers from around the globe to take an English grammar quiz. Of these, around a quarter million were monolingual native

Figure 3



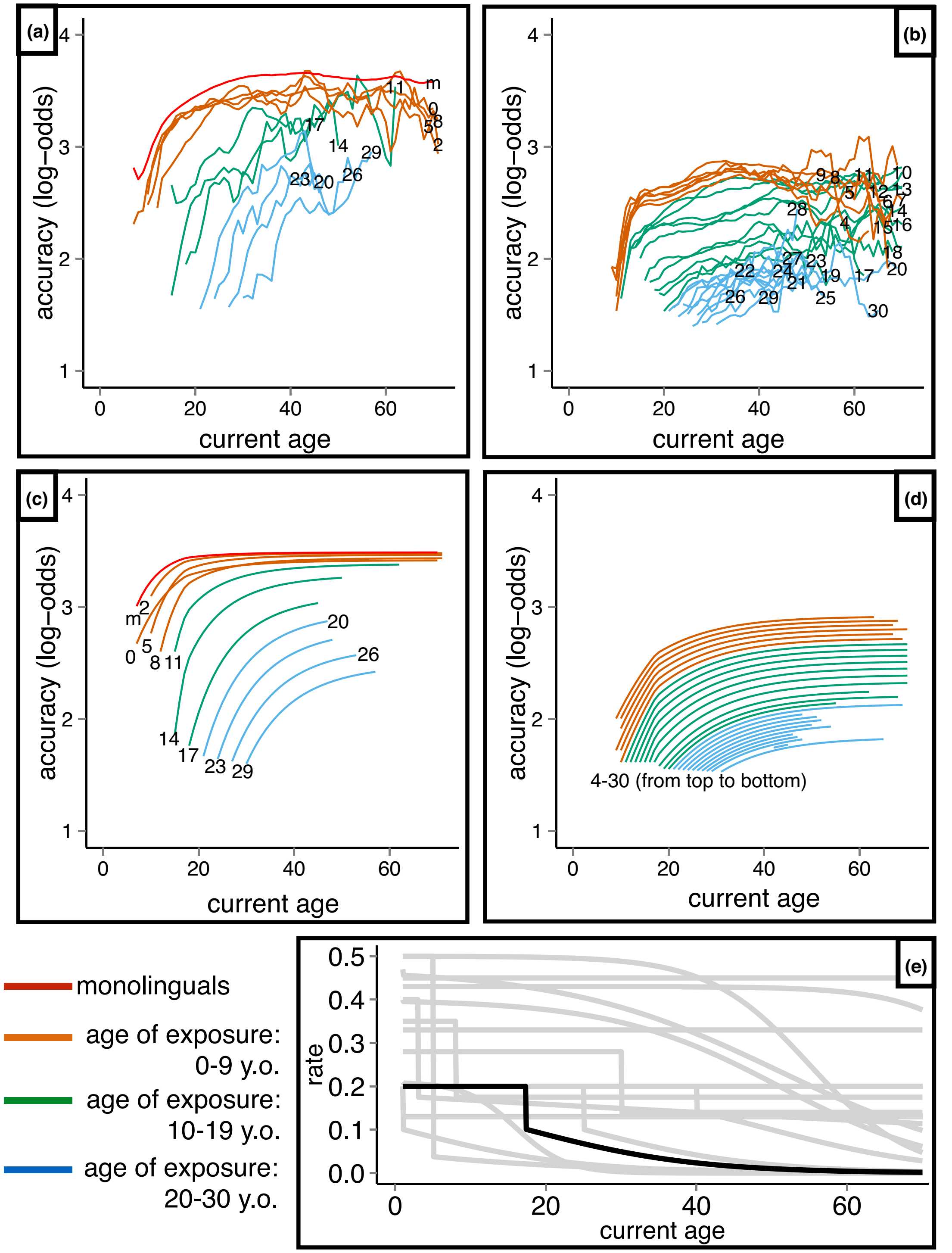
There is a complex relationship between how learning ability changes over time (left side of each panel) and knowledge measured after years of learning — or 'ultimate attainment' — as a function of the age at which learning started (right side of each panel). **(a)** Ultimate attainment is a function of how long someone has been learning and how rapidly they learned at each time point. That is, ultimate attainment is related to the integral under the learning ability curve over the period of learning. **(b)–(e)** Each panel shows a simulated ability curve and its corresponding ultimate attainment curve [5••], used with permission.

English speakers; a quarter million were 'non-immersion learners,' who had learned English outside of the home and primarily in a non-English-speaking country; and around fifty thousand were 'immersion learners,' who immigrated to an English-speaking country around the same time they began learning English.

Critically, subjects ranged widely in terms of current age (7–89) and years of exposure to English (1–89). This allowed HTP to directly measure how syntax knowledge increases with years of experience, conditioned on age of first exposure (Figure 4a,b). Statistical analysis showed that these 'learning curves' begin to become more shallow for learners who began after around the age of 10.

This finding suggests that the rate of learning begins to decline sometime well after 10 years of age. To see why, note that if learning began to decline at age 11, then this should be noticeable in people who began learning at age 9, who would then have only two years to learn at the fast, initial rate. To estimate how learning changes with age, HTP developed a novel analytic model in which learning is governed by an exponential decay function with a rate r that depends on age:

Figure 4



Current Opinion in Behavioral Sciences

Learning curves for English syntax, conditioned on age of first exposure, are shown for monolingual and immersion learners **(a)** and non-immersion learners **(b)**. Age of first exposure is indicated by color (see legend) and labeled on the lines themselves. Best model fits for these curves are shown in **(c)** and **(d)**. Best-fitting estimate for how learning rate/ability changes with age is shown in **(e)** (gray lines show examples of other hypotheses considered but rejected by the model). Figure reprinted from Hartshorne *et al.* [5••], used with permission.

$$r = \begin{cases} r_0 t \leq t_c \\ r_0\left(1 - \dfrac{1}{1 + e^{-\alpha*(t - t_e - \delta)}}\right) t > t_c \end{cases}$$

where t is current age, t_e is the age at which learning began, and t_c is a critical inflection point. Before t_c, learning rate is constant (r_0). Afterwards, it declines sigmoidally with shape parameters α and δ, which stretch and shift the sigmoid left or right.

The resulting model was able to fit the raw data quite closely (compare Figure 4a–d), and revealed a sharp decline in learning, beginning at age 17–18 (Figure 4e). Note that this was the first-ever estimate of how syntax learning ability changes with age — a feat made possible by the ability to collect large, diverse samples using MOEs.

HTP's results raise numerous questions, including just how robust and generalizable the findings are [37•]. Conveniently, HTP also illustrate a method for answering these questions: MOEs.

Limitations of MOEs

The most obvious limitation of MOEs is that (sufficiently many) subjects must have access to the necessary equipment. For instance, high-quality virtual reality systems (e.g. Oculus) are uncommon, consumer EEG systems (e.g. Muse) even more so, and wearables that measure skin conductance are only just coming onto the market (Fitbit Sense) [38–40]. Certain neuroscience methods like fMRI are unlikely to ever be widely available.

Similarly, the subjects must be available. MOEs may not work for some specialized populations such as pre-technological societies, individuals with congenital cataracts or delayed exposure to language, either because these populations are too small or have limited access to the Internet. It probably excludes many types of animal studies — particularly those involving controlled rearing — though the existence of some large-scale, Internet-enabled studies involving animals suggests this is an under-tapped resource [41,42].

More broadly, many familiar lab paradigms are optimized for the affordances of the laboratory and thus may not work well online. While we have over 150 years of experience with laboratory experiments to draw upon, we are only just beginning to develop paradigms optimized for MOEs. Of particular relevance for critical period research, while there is no obvious reason one cannot conduct longitudinal MOEs — and indeed on might expect them to be easier (e.g. no problems with subjects moving away) — we have much more experience navigating the potential pitfalls of longitudinal in-person studies than longitudinal MOEs. For these reasons, and given the many paradigms that can *only* be run online and not in the laboratory, it is unclear exactly how much these constraints on *paradigms* constrain research *questions*. This illustrates perhaps the most significant limitation of all: the novelty of MOEs means they may require more ingenuity, up-front investment, and risk. For instance, while there is no shortage of young children online, the field has not yet worked out reliable methods for recruiting large numbers of children under the age of 8 (but see [43]).

Relatedly, the meta-science of MOEs is still quite new, and we are only beginning to best practices [30•]. For instance, although differential dropout and self-selection may not be any worse than in laboratory studies [11•,44], they are nonetheless an issue [45,33]. Recently, researchers have begun capitalizing on the affordances of MOEs to better understand these issues and address them, both in research design and in analysis [35,33].

Conclusion

Massive online experiments (MOEs) provide an opportunity to collect the large, diverse samples that are necessary to understand age-related changes in high-level cognitive abilities. MOEs are not the only option: the same can be achieved by one very extraordinarily-resourced research group, by consortiums of many labs, or by pooling data from many studies [46–54]. However, MOEs are fast, cost-effective, and allow measurement of behavior difficult to observe in the lab. Thus, they provide significant opportunities for the study of cognition and behavior in general, and for critical periods in particular.

Conflict of interest statement

Nothing declared.

Acknowledgement

The author thanks Katharina Reinecke, Tianhu Chen, and Anna Petti for comments and suggestions.

Declaration of Competing Interest

The authors report no declarations of interest.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hensch TK: **Critical periods in cortical development**. *The Neurobiology of Brain and Behavioral Development* 2018:133-151. This paper provides an excellent recent review of the biology of critical and sensitive periods, primarily in animal perceptual systems.
••

2. Vogelsang L, Gilad-Gutnick S, Ehrenberg E, Yonas A, Diamond S, Held R, Sinha P: **Potential downside of high initial visual acuity**. *Proc Natl Acad Sci U S A* 2018, **115**:11333-11338
••
Evidence for a critical period in face perception in humans comes from children who do not receive visual input during infancy due to congenital cataracts that are later corrected. Such children do not develop normal

face perception. This paper argues that this is not due to a loss of plasticity, but rather due to differences in experience: whereas infants have poor visual accuity, the restored-sight children miss this phase of development. The authors present a computational model suggesting that the initial period of poor visual accuity supports learning to recognize faces. This account is reminiscent of Newport's 'Less is More' hypothesis for language acquisition, and helps expand the discussion of why learning abilities may change with age

3. Werker JF, Hensch TK: **Critical periods in speech perception: new directions**. *Annu Rev Psychol* 2015, **66**:173-196.

4. •• Fuhrmann D, Knoll LJ, Blakemore S-J: **Adolescence as a sensitive period of brain development**. *Trends Cogn Sci* 2015, **19**:558-566.

5. •• Hartshorne JK, Tenenbaum J, Pinker S: **A critical period for second language acquisition: evidence from 2/3 million English speakers**. *Cognition* 2018, **177**:263-277

This study of syntax-learning utilized the large, diverse subject sample available through a massive online experiment to disentangle subject's current age, the age at which they started learning, and how long they have been learning. This allowed for the first-ever estimate of how syntax-learning ability changes with age

6. McLaughlin B: **Second-language learning in children**. *Psychol Bull* 1977, **84**:438.

7. Krashen SD, Long MA, Scarcella RC: **Age, rate, and eventual attainment in second language acquisition**. *RESOL Q* 1979:1-168.

8. Flege J: **A non-critical period for second-language learning**. *A Sound Approach to Language Matters. In Honor of Ocke-Schwen Bohn* 2018.

9. DeKeyser R: **Age effects in second language learning, so obvious and so misunderstood**. *Est Lingüíst Ing Aplicada* 2019, **19**:235-242.

10. Rutter LA, Dodell-Feder D, Vahia IV, Forester BP, Ressler KJ, Wilmer JB, Germine L: **Emotion sensitivity across the lifespan: mapping clinical risk periods to sensitivity to facial emotion intensity**. *J Exp Psychol: Gen* 2019.

11. • Hartshorne JK, Germine LT: **When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span**. *Psychol Sci* 2015, **26**:433-443

While there were several earlier MOEs investigating age-related change in cognition, to my knowledge this was the first to draw on such data to make a sweeping theoretical contribution: specifically, challenging prevailing theories of age-related decline. It also provides the most comprehensive comparison of MOEs and traditional methods with respect to the measurement of age-related changes in cognition.

12. Rutter LA, Scheuer L, Vahia IV, Forester BP, Smoller JW, Germine L: **Emotion sensitivity and self-reported symptoms of generalized anxiety disorder across the lifespan: a population-based sample approach**. *Brain Behav* 2019, **9**:e01282.

13. Moran JM: **Lifespan development: The effects of typical aging on theory of mind**. *Behavioural Brain Res* 2013, **237**:32-40.

14. Hartshorne JK, Vidal J, Pinker S: *The Search for S: Individual Difference and Developmental Evidence for A Common Component in Linguistic and Nonlinguistic Social Cognition*. 2016. (under revision).

15. •• Vanhove J: **The critical period hypothesis in second language acquisition: a statistical critique and a reanalysis**. *PLOS ONE* 2013, **8**:e69172

This paper provides a valuable critique of the statistical methods used in earlier critical periods research. It also provides clear evidence that such studies are underpowered by orders of magnitude.

16. Salthouse TA: **When does age-related cognitive decline begin?** *Neurobiol Aging* 2009, **30**:507-514.

17. • Stanley T, Carter EC, Doucouliagos H: **What meta-analyses reveal about the replicability of psychological research**. *Psychol Bull* 2018, **144**:1325

The most recent large-scale study of statistical power in psychology, showing that psychology studies continue to be significantly underpowered.

18. Pashler H, Harris CR: **Is the replicability crisis overblown? Three arguments examined**. *Perspect Psychol Sci* 2012, **7**:531-536.

19. •• Open Science Collaboration: **The reproducibility project: a model of large-scale collaboration for empirical research on reproducibility**. In *Implementing Reproducible Computational Research (A Volume in the R Series).* Edited by Stodden F, Leisch F, Peng R. New York, NY: Taylor & Francis; 2014

A helpful description of the behind-the-scenes work on one of the largest multi-lab collaborations in the hisotry of psychology.

20. LeBel EP, McCarthy RJ, Earp BD, Elson M, Vanpaemel W: **A unified framework to quantify the credibility of scientific findings**. *Adv Methods Pract Psychol Sci* 2018, **1**:389-402.

21. Henrich J, Heine SJ, Norenzayan A: **The weirdest people in the world?** *Behav Brain Sci* 2010, **33**:61-83.

22. Nielsen M, Haun D, Kärtner J, Legare CH: **The persistent sampling bias in developmental psychology: a call to action**. *J Exp Child Psychol* 2017, **162**:31-38.

23. Birdsong D: **Plasticity, variability and age in second language acquisition and bilingualism**. *Front Psychol* 2018, **9**:81.

24. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR: **Power failure: why small sample size undermines the reliability of neuroscience**. *Nat Rev Neurosc* 2013, **14**:365-376.

25. Lazic SE, Essioux L: **Improving basic and translational science by accounting for litter-to-litter variation in animal models**. *BMC Neurosci* 2013, **14**:1-11.

26. •• Hartshorne JK, de Leeuw JR, Goodman ND, Jennings M, O'Donnell TJ: **A thousand studies for the price of one: accelerating psychological science with Pushkin**. *Behav Res Methods* 2019, **51**:1782-1803

This work describes the technical and methodological challenges involved in conducting massive online experiments. It also introduces Pushkin, a software ecosystem that my colleagues and I have developed to help meet those challenges

27. Adjerid I, Kelley K: **Big data in psychology: a framework for research advancement**. *Am Psychol* 2018, **73**:899.

28. Heigl F, Kieslinger B, Paul KT, Uhlik J, Dörler D: **Opinion: toward an international definition of citizen science**. *Proc Natl Acad Sci U S A* 2019, **116**:8089-8092.

29. ITU Telecommunication Development Sector: *ICT Facts and Figures*. 2017.

30. • Oliveira N, Jun E, Croxson T, Gajos KZ, Reinecke K: **Labinthewild: how to design uncompensated, feedback-driven online experiments**. *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 2017:25-28

This short paper describes many of the design choices that go into designing a successive massive online experiment. In addition to running a successful online laboratory (Labinthewild), this research team has published a number of significant quantitative analyses of and experiments into how to best design massive online experiments

31. Huber B, Gajos KZ: **Conducting online virtual environment experiments with uncompensated, unsupervised samples**. *PLOS ONE* 2020, **15**:e0227629.

32. Kumar A, Killingsworth MA, Gilovich T: **Waiting for merlot: anticipatory consumption of experiential and material purchases**. *Psychol Sci* 2014, **25**:1924-1931.

33. Steyvers M, Benjamin AS: **The joint contribution of participation and performance to learning functions: exploring the effects of age in large-scale data sets**. *Behav Res Methods* 2019, **51**:1531-1543.

34. Steyvers M, Hawkins GE, Karayanidis F, Brown SD: **A large-scale analysis of task switching practice effects across the lifespan**. *Proc Natl Acad Sci U S A* 2019, **116**:17735-17740.

35. Jun E, Hsieh G, Reinecke K: **Types of motivation affect study selection, attention, and dropouts in online experiments**. *Proc ACM Hum–Comput Interact* 2017, **1**.

36. Dodell-Feder D, Germine L: **Epidemiological dimensions of social anhedonia**. *Clin Psychol Sci* 2018, **6**:735-743.

37. Frank MC: **With great data comes great (theoretical)**
• **opportunity**. *Trends Cogn Sci* 2018, **22**:669-671

Frank delves into the analyses of [5], challenging certain statistical assumptions and raising important questions about analysis. This paper is useful for interpreting [5], but also has generalizable implications for the analysis of MOEs.

38. Charara S: *A New Fitbit Claims to Track Your Stress Levels. Can It Really Do It? Wired*. 2020. Retrieved from https://www.wired.co.uk/article/fitbit-stress-tracking-eda.

39. Martín B: *Video Games Controlled By Thoughts*. 2020. Retrieved from https://www.bbvaopenmind.com/en/technology/innovation/video-games-controlled-by-thoughts/.

40. Tankovska H: **Unit shipments of virtual reality (VR) devices worldwide from 2017 to 2019 (in millions), by vendor**. *Statistica* 2020. Retrieved from https://www.statista.com/statistics/671403/global-virtual-reality-device-shipments-by-vendor/.

41. Howard E, Aschen H, Davis AK: **Citizen science observations of monarch butterfly overwintering in the southern United States**. *Psyche* 2010:2010.

42. Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, Damoulas T, Dhondt AA, Dietterich T, Farnsworth A, Fink D, Fitzpatrick JW, Fredericks T, Gerbracht J, Gomes C, Hochachka WM, Iliff MJ, Lagoze C, La Sorte FA, Merrifield M, Morris W, Phillips TB, Reynolds M, Rodewald AD, Rosenberg KV, Trautmann NM, Wiggins A, Winkler DW, Wong W-K, Wood CL, Yu J, Kelling S: **The eBird enterprise: an integrated approach to development and application of citizen science**. *Biol Conserv* 2014, **169**:31-40.

43. Scott K, Schulz L: **Lookit (Part 1): a new online platform for developmental research**. *Open Mind* 2017, **1**:4-14.

44. Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB: **Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments**. *Psychon Bull Rev* 2012, **19**:847-857.

45. Coutrot A, Silva R, Manley E, De Cothi W, Sami S, Bohbot VD, Wiener JM, Holscher C, Dalton RC, Hornberger M, Spiers HJ: **Global determinants of navigation ability**. *Curr Biol* 2018, **28**:2861-2866.

46. MacWhinney B: *The Childes Project: The Database, vol 2*. Psychology Press; 2000.

47. Bergmann C, Tsuji S, Piccinini P, Lewis M, Braginsky M, Frank MC, Cristia A: **Promoting replicability in developmental research through meta-analyses: insights from language acquisition research**. *Child Dev* 2018, **89**:1996-2009.

48. Frank MC, Braginsky M, Yurovsky D, Marchman VA: **Wordbank: an open repository for developmental vocabulary data**. *J Child Lang* 2016.

49. Crosas M, King G, Honaker J, Sweeney L: **Automating open science for big data**. *Ann Am Acad Polit Soc Sci* 2015, **659**:260-273.

50. Whitley E, Deary IJ, Ritchie SJ, Batty GD, Kumari M, Benzeval M: **Variations in cognitive abilities across the life course: Cross-sectional evidence from understanding society: the UK household longitudinal study**. *Intelligence* 2016, **59**:39-50.

51. Smith TW, Davern M, Freese J, Hout M: *General Social Surveys, 1972–2018 [Machine-Readable Data File]*. NORC at the University of Chicago; 2018. [Producer and Distributor].

52. Ebersole CR, Mathur M, Baranski E, Bart-Plange D-J, Buttrick N, Chartier CR *et al.*: *Many Labs 5: Testing Pre-Data Collection Peer Review As An Intervention to Increase Replicability*. 2019. (results-blind manuscript).

53. Byers-Heinlein K, Bergmann C, Davies C, Frank MC, Hamlin JK, Kline M *et al.*: **Building a collaborative psychological science: lessons learned from ManyBabies 1**. *Can Psychol* 2020.

54. Harms MP, Somerville LH, Ances BM, Andersson J, Barch DM, Bastiani M *et al.*: **Extending the human connectome project across ages: Imaging protocols for the lifespan development and aging projects**. *NeuroImage* 2018, **183**:972-984.