HHS Public Access

Author manuscript

Infant Child Dev. Author manuscript; available in PMC 2024 March 21.

Published in final edited form as:

Infant Child Dev. 2024 ; 33(1): . doi:10.1002/icd.2348.

Developmental psychologists should adopt citizen science to improve generalization and reproducibility

Wei Li[1], **Laura Thi Germine**[2,3], **Samuel A. Mehr**[4,5], **Mahesh Srinivasan**[6], **Joshua Hartshorne**[1]

[1]Department of Psychology and Neuroscience, Boston College, Chestnut Hill, MA, USA;

[2]McLean Hospital, Belmont, MA, USA;

[3]Department of Psychiatry, Harvard Medical School, Cambridge, MA;

[4]Data Science Initiative, Harvard University, Cambridge, MA;

[5]School of Psychology, Victoria University of Wellington, Wellington, New Zealand;

[6]Department of Psychology, University of California, Berkeley, CA

Abstract

Widespread failures of replication and generalization are, ironically, a scientific triumph, in that they confirm the fundamental metascientific theory that underlies our field. Generalizable and replicable findings require testing large numbers of subjects from a wide range of demographics with a large, randomly-sampled stimulus set, and using a variety of experimental parameters. Because few studies accomplish any of this, meta-scientists predict that findings will frequently fail to replicate or generalize. We argue that to be more robust and replicable, developmental psychology needs to find a mechanism for collecting data at greater scale and from more diverse populations. Luckily, this mechanism already exists: Citizen science, in which large numbers of uncompensated volunteers provide data. While best-known for its contributions to astronomy and ecology, citizen science has also produced major findings in neuroscience and psychology, and increasingly in developmental psychology. We provide examples, address practical challenges, discuss limitations, and compare to other methods of obtaining large datasets. Ultimately, we argue that the range of studies where it makes sense *not* to use citizen science is steadily dwindling.

Keywords

citizen science; online experiments; replicability; generalizability

Widespread failures of replication and generalization pose a existential challenge to psychological research (Clark, 1973; Hartshorne & Schachner, 2012; Henrich et al., 2010; Judd et al., 2012; Open Science Collaboration, 2015; Pashler & Harris, 2012; Stanley et al., 2018; Yarkoni, 2019). If we cannot trust the data we collect or the inferences we draw, what is the point of doing research?

Correspondence concerning this article should be addressed to Joshua Hartshorne, Department of Psychology and Neuroscience, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, jkhartshorne@gmail.com.

Issues of reproducibility and generalizability are often treated as procedural errors, resulting from researchers employing contingent stopping, circular analyses, or other "p-hacking" methods, and addressed by encouraging or enforcing best practices. In fact, there is a much more fundamental problem: even if researchers did everything "correctly", they are nonetheless extremely unlikely to obtain replicable, generalizable results without collecting many orders of magnitude more data than is typical. Because this is often not how these issues are framed – and because researchers often vastly underestimate the scale of the problem (Bakker et al., 2016; Pashler & Harris, 2012) – we spend the first part of the paper reviewing the evidence.

The solution is, of course, to collect more data, and there are a few methods for doing so, including "many lab" collaborations (Byers-Heinlein et al., 2020; Ebersole et al., 2016; Ebersole et al., 2020; Frank, Bergelson, et al., 2017; Jones et al., 2021; Klein et al., 2018; ManyBabies Consortium, 2020; Van Essen et al., 2012), aggregate data repositories (Frank, Braginsky, et al., 2017; Hall et al., 2012; MacWhinney, 2000), and targeted large-scale data-collection projects run by governments and other large agencies (Gilmore, 2016; Harris et al., 2019; Olsen et al., 2001; Trouton et al., 2002; West, 2000). However, one method has been substantially underutilized, particularly given its flexibility, speed, low cost, and track record of success: citizen science. Citizen science is a paradigm in which volunteer researchers collect and/or code data in order to contribute to a larger scientific or societal purpose (Bonney et al., 2014). Critically, it can and has been used to collect datasets on the necessary scale.

As should become clear below, we believe the main reason citizen science is not more widely used in developmental psychology is that many researchers are not familiar with it and have not (yet) acquired necessary skills. This of course has been true at the advent of nearly every major new experimental paradigm, including the notion of doing experiments at all. (At the dawn of experimental developmental psychology, vanishingly few researchers knew how to run experiments with children.) We hope to show below that the value of citizen science is so profound that it will very often be worth the cost of adopting.

After first reviewing the relationship between access to subjects and reproducibility/generalizability, we introduce citizen science, give examples of successful projects, and address common concerns about data quality and feasibility (particularly for developmental samples). We outline a number of areas where we feel there is untapped potential. We then discuss some of the outstanding challenges and how they might be addressed.

The Crises of Replication and Generalization

Systematic investigations in psychology and neuroscience typically report replication rates between 30% and 70% (Camerer et al., 2018; Cova et al., 2021; Ebersole et al., 2020; Hartshorne, Skorb, et al., 2019; Klein et al., 2018; LeBel et al., 2018; Open Science Collaboration, 2015). Even where findings replicate, they may not generalize. Unfortunately, we have less in the way of empirical estimates of generalizability – in part because it is difficult to estimate generalizability when replicability is so low – but long experience gives plenty of evidence of findings that do not generalize across slight variations in experimental

parameters (Byers-Heinlein et al., 2020; Klein et al., 2018; ManyBabies Consortium, 2020; Yarkoni, 2019), to moderately different stimuli (Hartshorne & Snedeker, 2013; Peterson et al., 2021), or to different subject populations (Evans & Levinson, 2009; Henrich, 2020; Nisbett, 2004).

Replicability

While empirical reports of low replicability have surprised many, they should not have. The typical study in psychology has less than a 50/50 chance of detecting a average-sized effect — an issue first identified in the 1960s and re-confirmed regularly since then (Bezeau & Graves, 2001; Button et al., 2013; Cafri et al., 2010; Chase & Chase, 1976; Clark-Carter, 1997; Cohen, 1962, 1994; Fraley & Vazire, 2014; Hartshorne & Schachner, 2012; Maddock, 1999; Maxwell, 2004; Mone et al., 1996; Osborne, 2008; Rossi, 1997; Schäfer & Schwarz, 2019; Sedlmeier & Gigerenzer, 1992; Stanley et al., 2018; Szucs & Ioannidis, 2017; Vankov et al., 2014; Ward, 2002).[1] That is, average statistical power is less than 50%. To reach this conclusion, statisticians have calculated the average effect sizes across studies as well as average sample sizes. It turns out that even when the null hypothesis is false, an experiment will more likely than not fail to reject the null hypothesis unless it involves an unusually large number of subjects or unless the effect under investigation is unusually large.

It is obvious that this chronically low statistical power renders null results uninformative in most cases. What is perhaps less intuitive is that it also renders *significant* results less informative. The reason is that if the null hypothesis being false rarely results in a significant p-value, then significant p-values are probably due to something else: random chance, mathematical errors, or p-hacking (Pashler & Harris, 2012). For example, suppose you thought there was only a 10% chance that a particular intervention (say, playing Mozart to a baby) would have some particular effect (raising the baby's IQ). Then you run a study with typical levels of statistical power (35%) and get a significant result. Some simple math shows there is a 56% chance this was a false positive.[2] In practice, the probability of a false alarm is actually probably higher, since many common research practices tend to increase the false alarm rate, such as contingent stopping, failure to correct for multiple comparisons, and treating items as fixed effects (Clark-Carter, 1997; John et al., 2012). Given the emphasis our field places on unexpected results, the math suggests that high-profile findings are particularly unlikely to be true.

How large a sample size is required to make a null result meaningful? Most studies are framed as testing for the existence of an effect, so they should have the statistical power to detect any reasonably sized effect. Maxwell et al. (2015) suggest ensuring enough statistical power to detect any effect larger than 1/20 of a standard deviation, resulting in 20,794 subjects for a two-sampled t-test.[3] (Smaller effects are not only of limited practical

[1] We note that by centering statistical power we are adopting a paradigm in which science is about detecting the presence or absence of an effect. There are good reasons to dispute this as the main desideratum (Newell, 1973; Wilson et al., 2020). However, the alternatives are no less data-hungry, and so land in roughly the same place with regards to our central argument here about the need for more data.

[2] $P(H_0|sig) = \frac{P(H_0) * P(sig|H_0)}{P(sig)} = \frac{P(H_0) * P(sig|H_0)}{P(sig|H_0) * P(H_0) + P(sig|H_1) * P(H_1)}$. $P(H_0|sig)$ is the probability of false alarm, the probability of that null hypothesis is true given the significant result. $P(H_0)$ and $P(H_1)$ are the prior of the effect. $P(sig/H_0)$ is the Type I error. $P(sig/H_1)$ is the statistical power.

importance but are reasonably likely to be due to tiny uncontrolled confounds or minor imperfections in the stimuli.)

Researchers sometimes mistakenly suggest that large samples are only necessary if one wishes to detect effects too small to be theoretically meaningful (Combs, 2010). This is not the case. Ironically, large effects are so rare in psychology that — all else equal — reports of large effects are more likely to be spurious or at least overestimated (Funder & Ozer, 2019). Second, detecting typically-sized effects in psychology requires far more subjects than researchers usually intuit (Bakker et al., 2016). Indeed, if one wishes to successfully detect a randomly-selected effect from the psychology literature in a two-sample t-test 95% of the time, one would need 7,000 subjects.[4] The problem is actually worse if one uses Bayesian statistics, where 7,000 subjects only gives one an 86% chance of obtaining "very strong" evidence for the alternative hypothesis (a rough Bayes-factor equivalent of statistical significance). Even with 20,794 subjects — the number suggested by Maxwell and colleagues — this probability rises only to 93%. Note, moreover, that these numbers were estimated assuming a two-sample t-test: investigation of a 2×2 interaction would generally require at minimum 4 times as many subjects (Blake & Gangestad, 2020). In short, many of the effects that our field focuses on are simply too small to be reliably detected by the typical study, which only has sufficient power to detect unusually large effects.

Unfortunately, interventions to control the false discovery rate, such as preregistration and registered reports (Chambers, 2019; Nosek et al., 2012), do not change this math. As long as most studies are unlikely to detect a significant result even when the alternative hypothesis is true, then widespread use of registered reports should result in journals full of mostly null results, which indeed appears to be happening (Scheel et al., 2021).[5]

Before continuing, we note that statistical power can also be increased by decreasing the amount of noise in the data. This follows because, for purposes of statistical analysis, effects are measured relative to variability in the data, so decreasing variability increases the effect size. Part of this variability is often measurement error. Intuitively, one would have a better chance of detecting the effect of some educational intervention on children's linguistic

[3]This is based on ensuring 95% power. Some commentators focus on achieving 80% statistical power, but this accepts a fairly high error rate, namely accepting the null hypothesis one out of every 5 times that it is false. Since our focus here is on obtaining robust and reliable results, we adopt a more conservative 5% Type II error rate, similar to the widely-adopted 5% Type I error rate.

[4]We calculated power by sampling from a skewed half-normal distribution built to roughly match the histogram of effect sizes reported in Richard et al. (2003). While this sample includes only social psychology studies, its mean is similar to what has been reported by psychology as a whole (Stanley et al., 2018) (unfortunately, studies of the entire field do not report the shape of the distribution). We used each sampled effect size to create a synthetic two-sample dataset, which we then analyzed with a t-test. For each sample size being considered, we conducted 5,000 simulated studies and calculated statistical power. Note that several factors may result in this being an overestimate or an underestimate. On the one hand, the strong bias against publishing null results in psychology means that reported effect sizes – and, consequently, power estimates – are substantially inflated (e.g., Open Science Collaboration, 2015). On the other hand, the distribution of published effect sizes is necessarily a mixture of samples from both the null and alternative hypotheses; inclusion of the former will tend to left-skew the distribution, decreasing observed power. Likewise, the distribution includes not just main effects but also interactions, which tend to be smaller, also decreasing power. Thus, our estimate of 7,000 subjects may be too large, though it may also be too small. At the moment, it is the best available estimate.

[5]It might seem surprising that registered reports are underpowered, given that many journals require explicit power analyses for such studies. However, it appears that these power analyses are usually based on previously-reported effect sizes, which tend be be substantially inflated (Open Science Collaboration, 2015). This has been recognized by researchers working on replications, who discovered that replications that are powered based on the effect size in the original paper will rarely successfully replicate even when the null hypothesis is false. There is now a trend to power to detect effects much smaller than what was previously reported (Ebersole et al., 2020). As noted above, a more realistic power analysis would generally indicate needing more subjects that the typical researcher can obtain. It follows that studies with realistic power analyses are unlikely to be accepted as registered reports.

knowledge using a comprehensive standardized measure of linguistic knowledge rather than a 20-word vocabulary test. It is currently unknown just how much of the variability in psychological data is due to measurement error as opposed to inherent variability between subjects or items, but it stands to reason that improved measurement would at least somewhat decrease the number of subjects needed. There are a number of methods for improving measurement precision, though some of them — such as utilizing Item Response Theory (Embretson & Reise, 2013) — can require substantial numbers of subjects in their own right. In any case, addressing measurement error does not help with generalization, which we turn to in the next subsection. Thus, while we strongly endorse greater attention to reliable measurement (Chen & Hartshorne, 2021; Germine et al., 2019; Passell et al., 2019), we do not discuss it further here.

Generalization

The discussion so far has assumed that we only wish to know whether we can reliably obtain the same results with the same stimuli, experimental parameters, and subject pool. The problem of insufficient data gets worse if we care about generalization, which we usually do (Byers-Heinlein et al., 2020; Clark, 1973; Henrich, 2020; Judd et al., 2012; Klein et al., 2018; Moriguchi, 2021; Nisbett, 2004; Yarkoni, 2019).

In principle, there are statistical methods for assessing whether a finding is likely to generalize to the population of subjects, items, & procedures under consideration (Baayen et al., 2008; Clark, 1973). However, these methods depend on the subjects, items, and procedures being randomly sampled from the population, which is almost never the case. For instance, researchers may wish to generalize to the population “humans” but in fact only sample from introductory psychology students at the local university. Similarly, researchers may wish to generalize to “aversive stimuli” but in fact sample only from photos of open wounds. In any case, many studies use too few items – often as few as a single stimulus per condition – to statistically estimate generalization (Judd et al., 2012). This is certainly true in developmental psychology, where single-trial studies are common, particularly in infant research. Similarly, most studies in psychology consider at most a handful of procedures, and typically only one. It is difficult to quantify exactly how problematic all this is – measuring likelihood of generalization when the sample is non-representative remains a difficult, unsolved, and perhaps unsolvable problem (D’Amour et al., 2020) – but there is little reason to be optimistic.

As a practical matter, researchers rarely test whether an effect varies between populations of subjects, likely because they only have easy access to one or two populations (Henrich et al., 2010; Hilton & Mehr, *in press*). As a result, most of what we know about human psychology is restricted to a relatively narrow segment of the species (Henrich et al., 2010; Kidd & Garcia, 2021; Nielsen et al., 2017). However, even if researchers had access to multiple populations, and even if they could obtain representative samples of those populations, they would need unusually large subject samples. As noted above, detecting whether an effect varies across two populations requires at least 4x as many subjects as detecting that effect in one population. The same goes for testing for differences across populations of items or experimental parameters.

In short, given the way psychology is currently practiced, we should expect relatively low rates of replication and generalization. Indeed, high rates of replication and generalization would call into question the statistical regime that undergirds our field, thereby undermining our belief in those same results.

But what about developmental psychology?

While much of the discussion on replication and generalization has focused on social psychology and neuroscience, there is little reason to suspect the status of developmental psychology is better and good reason to suppose it is worse. In particular, the exigencies of testing young children mean that samples tend to small, the number of stimuli few, and subjects highly skewed towards the affluent North Americans (Bergmann et al., 2018; Kidd & Garcia, 2021; Nielsen et al., 2017; Oakes, 2017; Scott & Schulz, 2017).

Although there have not yet been systematic studies of replicability in developmental psychology (but see Black & Bergmann, 2017; Byers-Heinlein et al., 2020), there are numerous examples of classic findings that have remained controversial over decades, or which have explicitly failed to replicate. These include: when and whether there is a critical period for language acquisition (Bialystok & Kroll, 2018; Birdsong & Molis, 2001; Hartshorne et al., 2018; Singleton & Leśniewska, 2021), the relative importance of pretend play in children's development (Lillard et al., 2013; Lillard et al., 2011; Weisberg, 2015), whether bilingualism affects executive function (Dick et al., 2019; Paap, 2019), and whether toddlers can succeed at "implicit" theory of mind tasks (Baillargeon et al., 2018; Burnside et al., 2018; Dörrenberg et al., 2018; Kulke, Reiß, et al., 2018; Kulke, von Duhn, et al., 2018; Poulin-Dubois & Yott, 2018; Powell et al., 2018; Wiesmann et al., 2018).

Reconciling abysmal math with the actuality of progress

The gloomy assessment above might seem to preclude psychology having made any progress or discovered (m)any clear facts. Indeed, some observers have concluded that most of what we believe we know must be false (Ioannidis, 2012). However, we take it as *prima facie* obvious that psychology — including developmental psychology — has made progress over the last 150 years, that we have indeed discovered some things. In this context, it is relevant that the concerns raised above apply to the state of our experimental evidence, not the state of our knowledge. These are separable.

First, much of the dismal math above followed from the fact that most effects in psychology are fairly small relative to noise. As a result, one needs a lot of data to establish that some difference between conditions is real and not just due to noise. However, some effects are quite large relative to noise, such as classic Gestalt perception effect or the fact that babies don't know language. Because these effects are robust and vary little from trial to trial or stimulus to stimulus, they can be statistically established for an individual, often within a few minutes. Because they vary so little from individual to individual, only a few subjects may be statistically sufficient to show that the effect is present for a specific subject population. Alternatively, if some subject population is not subject to the effect (their babies all have military posture), the difference will be readily apparent.

Second, data is not interpreted blindly. Rather, scientists think hard about its implications, in the context of other studies and in dialog with other scientists. This can lead to powerful insight. A compelling example comes from a recent study of risky choice: while a neural network trained on the results of large numbers of risky choice experiments vastly outperformed the best human-built theories — including the Nobel-winning Prospect Theory — human-built theories far outperformed the model when considering only the small amount of data previously available to theorists (Peterson et al., 2021).[6]

Third, while experimental evidence is important, it is not our only form of empirical evidence. Thus, while we are unaware of any systematic *experimental* data to this effect, it is clear both that every human culture uses language, and in no culture are babies born talking. Moreover, some reasonable conclusions can follow from these facts even without any direct experimental evidence (in no culture do babies prefer knock-knock jokes to puns). As a result, it is quite possible to reach the indisputably correct conclusion even if it is not statistically justified by the data. For instance, the experimenter who runs an experiment consisting of a single trial with a single subject and concludes that no people have extra-sensory perception is probably correct, even though their evidence is underwhelming.

Finally, even if each experiment has only a fraction of the data needed to test the broad claim of interest, each one still has some data, and many experiments eventually add up, giving us a progressively more clear understanding of the phenomenon. By the same token, however, this process requires a lot of false steps along the way. The results of the first few studies are unlikely to both replicate and generalize, and indeed *consistent* results across early studies would be statistically shocking and indicate researcher error (Francis, 2012). Mathematically, the optimal strategy would be to make *no inferences at all* until at least a few dozen standard-sized studies have been published, but even if that is technically correct, we assume it is obvious that this advice is very difficult to follow. (It would, however, result in much shorter papers, since Discussion and Conclusion sections would no longer be necessary, except for in the occasional meta-analysis paper.)

Thus, the abysmal math does not necessarily mean we have made no progress, but that the fact we have made progress is as much *despite of* as *because of* our experiments. If we believe in the scientific method, we should also believe that better data would result in substantially faster progress.

Methods of Obtaining Larger Datasets

To summarize the discussion above, even simple experiments require thousands of subjects just to reliably detect an effect, and orders of magnitude more if the goal is to test generalization across subjects, items, and procedures. It is generally not feasible for one laboratory to test that many subjects in a face-to-face setting. Moreover, laboratories are usually restricted by the diversity of the surrounding population, (though this challenge has relaxed somewhat in the videoconferencing era; Janghorban et al., 2014; Reñosa et al., 2021; Sheskin & Keil, 2018; Su & Ceci, 2021).

[6]This method, we should note, is not fool-proof. Large numbers of scientists thinking hard can land on entirely false conclusions, as exemplified by the recent collapse of the social priming literature (Chivers, 2019; Shanks & Vadillo, 2021).

Turning to online labor markets such as Amazon Mechanical Turk, Prolific, or Qualtrics Panels does increase the size and diversity of samples, but obtaining large samples can be prohibitively expensive and the available subjects are still not that diverse (Difallah et al., 2018; Moss et al., 2020; Turner et al., 2020). Of particular concern to developmental psychologists, participants in these pools are required to be at least 18 years old. This can be circumvented by paying *parents* to have their children participate, but only a small fraction of the pool consists of parents with children of the right age.

One option — the "many labs" approach — is for a large number of laboratories collaborate on collecting data for a single study (Byers-Heinlein et al., 2020; Ebersole et al., 2016; Ebersole et al., 2020; Frank, Bergelson, et al., 2017; Jones et al., 2021; Klein et al., 2018; ManyBabies Consortium, 2020; Van Essen et al., 2012). In most cases, the datasets are much larger than typical, usually numbering in a few thousand, which is an improvement but still far short of ideal. Moreover, the "many labs" approach does not so much speed progress but rather concentrate activity (and progress) on a smaller number of methods and questions.

An older, more top-down approach is large-scale government data-collection efforts, such as the Early Childhood Longitudinal Study (N=14,000; West, 2000) or the Danish National Birth Cohort (N=100,000; Olsen et al., 2001), or similar efforts run by non-governmental organizations, such as the National Longitudinal Study of Adolescent Health (N>90,000; Harris et al., 2019) or the Twins Early Development Study (N=15,000; Trouton et al., 2002) (for review, see Gilmore, 2016). While these efforts show that it is in principle possible to run individual studies at the necessary scale, they also illustrate the difficulty of doing so for every research question. These efforts require enormous dedicated resources, often over the span of decades. They may speed discovery by providing higher-quality datasets, but — like the many labs approach — they do so by concentrating efforts on a small number of projects. Moreover, they are often (but not always) limited to survey data, which is historically easier to collect at scale.

There are also *ex post facto* "many labs" studies, where researchers aggregate data collected in many locations and for many different purposes into a single large database. Examples include the National Database of Autism Research (NDAR) (N>85,000; Hall et al., 2012), the Child Language Data Exchange System (CHILDES; N=7,085; MacWhinney, 2000)[7], and WordBank (N=75,114; Frank, Braginsky, et al., 2017). Some such efforts have been enormously productive: CHILDES has been the backbone of language acquisition research for decades and supplied critical data for many hundreds of papers and has been cited nearly 9,000 times. However, such projects depend on large numbers of labs happening to collect compatible data, often using the same instrument (in the case of WordBank, the Communicative Development Inventory). In short, while these aggregation projects have outsized value and there can and should be more of them, they are usually not possible — particularly when the questions have not been studied previously or the methods are new.

[7]While the absolute numbers for CHILDES are small relative to some of the other examples we list, the amount of data collected per child is often staggering.

In summary, each of the standard methods of obtaining larger datasets has significant limitations. The most successful methods succeed in part by vastly curtailing the range of questions studied at any given time. Perhaps curtailing the number of studies and pooling our efforts is what we must do. Luckily, however, there is another option: citizen science.

Citizen Science

In citizen science, participants act as "volunteer researchers" to benefit science and society (Bonney et al., 2014). Citizen scientists can vary in how directly involved they are in the research, from being the primary movers such as in the Provincetown COVID oubreak (Simmons-Duffin, 2021) to "merely" helping with labor-intensive data-processing such as conducting migratory bird surveys or collecting water samples (Bonney et al., 2014; Chari et al., 2017; Cooper et al., 2014; Lintott, 2019; Von Ahn, 2006).

A Method for Obtaining Large Samples

While even small studies involving volunteers are citizen science, we are particularly interested in the fact that projects with tens of thousands of subjects or even millions of subjects are increasingly common (e.g., Awad et al., 2018; Brysbaert et al., 2016; Chen & Hartshorne, 2021; Coutrot et al., 2022; Gebauer et al., 2016; Hartshorne et al., 2014; Liu et al., 2021; Mehr et al., 2019; Nosek et al., 2002; Reinecke & Gajos, 2015; Robins et al., 2001; Westgate et al., 2015; Youyou et al., 2017). Moreover, samples are often strikingly diverse. Systematic comparisons show that citizen science samples are far more diverse than those of typical lab-based studies (Gosling et al., 2004; Reinecke & Gajos, 2015; but see Strange et al., 2019). It is not uncommon for citizen science studies to report data from dozens of countries, a wide range of socioeconomic statuses, and from across much of the lifespan (Bleidorn et al., 2013; Dodell-Feder & Germine, 2018; Hartshorne & Germine, 2015; Hartshorne et al., 2018; Maylor & Logie, 2010; Riley et al., 2016) — something that is otherwise rarely seen. Indeed, although not all citizen science projects involve large, diverse subject samples, nearly all studies involving large, diverse subject samples are citizen science projects.

Ironically, what makes such large, diverse samples possible is that citizen science projects do not offer cash or course credit as compensation. The first thing to notice is that this eliminates many practical constraints on participation. The vast majority of humanity cannot be induced to participate in your study through cash compensation or course credit because you have no way to get the money to them and they are not enrolled in an introductory psychology course at your university. Resorting to labor markets such as Amazon Mechanical Turk or Prolific help only so much: less than 0.01% of humanity in enrolled in these platforms.[8]

In contrast, if subjects do not need compensation and participation is possible remotely over the Internet, then over half the world's population — more than 4 billion individuals — are

[8]As of writing, the research tools on Prolific list just over 120,000 subjects active in the last 90 days. There is some debate about the exact number of subjects available through Amazon Mechanical Turk, but it is likely no more than a quarter million and perhaps fewer than 10,000 (Robinson et al., 2019; Stewart et al., 2015).

in principle available (ITU Telecommunication Development Sector, 2019). Critically, while access is higher in developed countries (87% of individuals), it is substantial in developing countries (44%) and even in even designated Least Developed Countries (LDCs; 19%).[9] Moreover — and critically for developmental psychology — Internet access skews young. In the United States, for instance, 95% of 3–18 year-olds have Internet access, including 83% of American Indians and Alaska Natives (Irwin et al., 2021). In short, while not everybody is reachable online, far more are at lest in principle reachable by this method than by any other.

The second factor is that while only some people can be induced to participate in a study by offers of nominal monetary compensation or course credit, nearly anyone will participate in a project they find intrinsically rewarding. Citizen science projects succeed by attracting participants to the project.[10] Birders contribute to bird surveys and astronomy enthusiasts categorize images of galaxies (Raddick et al., 2009). Still others participate because the activity has been specifically designed to be fun. One common paradigm of particular relevance to psychology is the "Game With A Purpose", in which the data-collection or data-processing task is gameified, making it more interesting and easier to understand (Von Ahn, 2006). A somewhat more common variant is the "Viral Experiment", where participants engage in some experimental task and get personalized feedback at the end. (These are "viral" in that netizens spontaneously promote the project on social media, through web videos, or by simply emailing the link to friends — all because they find participation fun and something they want to share with others.)

These paradigms are particularly relevant for developmental psychology, since while young children may not be highly motivated to contribute to science, they are enthusiastic players or video games, watchers of videos,[11] and participants in other online activities that (in principle) can generate highly valuable psychological data. In essence, participants in citizen science projects are not truly uncompensated; in fact, they are compensated with something they find far more valuable than what in-lab studies offer.

Instrumentation & Data Quality

Large, diverse samples would be meaningless if one could not measure the behaviors of interest. There was certainly a point in time in which the machinery needed to measure human behavior was available only in laboratories. These days, most laboratory experiments are conducted using widely-available, off-the-shelf consumer technology such as laptops and tablets, much of which subjects already own. Taking into account the proliferation of computers, mobile devices, and wearables, it is possible — using the subjects' own equipment and without requiring any travel on their part — to present subjects with a wide range of stimuli, including audio, video, and even rudimentary virtual reality, and to collect such measures as button-presses (with reaction time), mouse and track-pad tracking,

[9]These numbers are smaller for the percentage of individuals who have internet at home (cf. UNICEF et al., 2020). While home internet access is critical for remote schoolwork, it is probably less important for participation in citizen science.

[10]Anecdotally, some researchers question suggest that it is more ethical to induce people into participating in research by paying them than by inviting people to do something they enjoy and will do of their own accord. Ethics is inherently subjective, but this seems backwards to us.

[11]57% of American infants under the age of three who live in an Internet-connected household watch YouTube (Auxier et al., 2020). This number rises to 81% of children ages 3 & 4.

drawing, voice responses, video, (coarse-grained) eyetracking, GPS, physical position, heart rate, and skin conductance, to just name a few (Gjoreski et al., 2018; Gosling & Mason, 2015; Harari et al., 2016; Hartshorne, de Leeuw, et al., 2019; Huber & Gajos, 2020; Miller, 2012; Mottelson & Hornbæk, 2017; Papoutsaki et al., 2016; Yang & Krajbich, 2021). While most neuroscience methods are not available, the advent of EEG headsets for gaming means this may change (Badcock et al., 2013; Duvinage et al., 2013).

Large, diverse samples would also be meaningless if data quality was poor. However, data quality is typically quite high (Germine et al., 2012; Gosling et al., 2004; Hartshorne, de Leeuw, et al., 2019; Reinecke & Gajos, 2015; Ye et al., 2017). Though it might seem *a priori* that decreased experimenter control over the procedure would lower quality, the citizen science approach also offers a key benefit often overlooked by researchers: subjects actually *want* to participate (Jun et al., 2017; Ye et al., 2017). This is exemplified by the fact that most subjects are referred by other subjects,[12] and popular studies occasion a great deal of online discussion (for examples, see bit.ly/3BYxq8o, bit.ly/3ob6lKi, and bit.ly/3Lvwqhj). As a result, data-quality is high and shirking and dishonestly is rare (Germine et al., 2012; Jun et al., 2017; Liu et al., 2021; Ye et al., 2017).

This is in contrast to studies that pay participants, where motivated lying and shirking is understandably common (Berinsky et al., 2014; Chandler & Paolacci, 2017; Chmielewski & Kucker, 2020; Kan & Drummey, 2018; Maniaci & Rogge, 2014; Marjanovic et al., 2014; Meade & Craig, 2012; Oppenheimer et al., 2009). As noted by Chandler and Paolacci (2017), participants on Amazon Mechanical Turk are routinely paid more for giving specific answers whereas payment is unaffected by whether those answers are the participant's *true* answers, and so a reasonable percentage respond accordingly. Even where this isn't the case, all paid subjects are effectively paid more for finishing sooner, which is often antithetical to producing high-quality data.

From the discussion above, there are two obvious constraints on using citizen science to collect large, diverse data sets. The first is that many common research paradigms, having been designed for a captive audience, are not attractive to participants. Designing a project that will attract large numbers of subjects is not trivial. The second is that large samples are only possible if data-collection is automated; no laboratory could do live Zoom interviews with a million subjects. We will return to these and other issues in the section "Challenges." However, the sheer range of published studies suggests that these challenges are primarily a limit on experimental paradigms, not on scientific questions. We review some of these in the next section.

What Has Been Done with Citizen Science?

The most compelling argument for using citizen science to study cognition and behavior is the wide range of fundamental discoveries made using this paradigm so far (e.g., Bleidorn et al., 2016; Brysbaert et al., 2016; Gebauer et al., 2014; Gebauer et al., 2016; Germine et al., 2011; Halberda et al., 2012; Hampshire et al., 2012; Hartshorne et al., 2018; Killingsworth

[12]Since 2008, 53% of traffic to gameswithwords with known origins has been from social media.

& Gilbert, 2010a; Kumar et al., 2014a; Mehr et al., 2019; Riley et al., 2016; Salganik et al., 2006).

With regards to developmental question, the most prominent example is studies of cognitive and social development over the lifespan, which overturned the then-dominant consensus theory of lifespan cognitive development. In particular, for much of the 20th Century, the general consensus was that some cognitive abilities (dubbed 'fluid intelligence') depended heavily on raw thinking speed and peaked in late adolescence before declining rapidly, whereas other cognitive capacities (dubbed 'crystalized intelligence') depended more on accumulated knowledge and continued to develop into middle age before declining (Cattell, 1963). This consensus was based on sparse data, however, often comparing only 2 or 3 coarsely-defined age groups. Over the last 15 years, esearchers have used popular online quizzes to track changes across much of the lifespan (usually roughly 8 to 80 years old) in attention (Fortenbaugh et al., 2015), memory (Maylor & Logie, 2010), vocabulary (Hartshorne & Germine, 2015), grammar (Hartshorne et al., 2018), numerical processing (Halberda et al., 2012), emotional perception (Olderbak et al., 2019), face perception (Germine et al., 2011), and personality (Bleidorn et al., 2013; Srivastava et al., 2003), to name just a few. These new, more finer-grained measurements showed a dizzying array of lifespan trajectories that cannot be explained by the fluid/crystalized distinction. The field is only beginning to process the findings and develop new, alternative theories that can account for the findings (Hampshire et al., 2012; Hartshorne & Germine, 2015). Similar studies of personality and social development have revealed similarly striking, unexpected findings (Nosek et al., 2002; Robins et al., 2002; Soto et al., 2011; Srivastava et al., 2003).

These lifespan data sets have also allowed researchers to delve into individual differences across development in ways not previously possible (Halberda et al., 2012; Johnson et al., 2010; Wilmer et al., 2012). For instance, Halberda and colleagues (2012) found that while number sense abilities change across the lifespan and peak in the late 30s, individual differences remain very large at every age group. Johnson and colleagues (2010) found that the underlying factor structure for working memory changes with age, indicating a need for more sophisticated theories of working memory, as well as more precise, theory-informed statistical methods.

Large lifespan data sets have also changed the debate about specific domains. For instance, debates about critical periods in language acquisition center on the age at which the ability to learn language declines and how quickly it does so. However, prior to the advent of Citizen Science, this proved impossible to measure: attempts to measure age-related changes in language-learning in real time in the lab have failed (meaningful language learning takes too many months), and retrospective cross-sectional studies require hundreds of thousands of subjects who began learning the target language at different ages and for different lengths of time (Hartshorne et al., 2018). Hartshorne and colleagues (2018) collected such a dataset and were able to provide the first estimate of how the ability to learn syntax changes with age, concluding that it declines sharply at around 17–18 years old (see also Chen & Hartshorne, 2021).

Outside of developmental psychology, citizen science has been applied to a wide range of questions and psychological domains, with impressive results. For instance, Riley and colleagues (2016) found that gender differences in sustained attentional control were predicted by gender disparities in employment across 41 countries (N=21,484). Personality researchers have found that friends and spouses really are more similar in terms of personality than strangers (N=897,960; Youyou et al., 2017) and that people have higher self esteem when their own personality better matches the modal personality of the city they live in (N=543,934; Bleidorn et al., 2016). Salganik and colleagues (2006) found distinct effects of music quality and perceived popularity on actual popularity by randomly assigned 14,000 participants to distinct music communities on a music streaming site. Reinecke and Gajos (2014) documented striking cross-cultural differences in aesthetic preferences. Germine and colleagues (2015) found that childhood adversity negatively impacted theory of mind and social affiliation but not face processing, suggesting the differential effects of environment on different aspects of social cognition. Finally, Awad and colleagues (2018) collected judgments on 26 million trolley problems in ten languages from more than 3 million people in 233 countries, revealing substantial systematic cultural differences in how much people value sparing women and children, blame action more than inaction, etc.

The aforementioned examples capitalize on the diversity of the samples. Other projects have used enormous sample sizes to randomize stimuli or procedures across subjects, directly addressing concerns about generalization across methods (Brysbaert et al., 2016; Hartshorne et al., 2014; Hartshorne, de Leeuw, et al., 2019; for discussion, see Hilton & Mehr, *in press*). For instance, by testing different subjects on different words, Brysbaert and colleagues (2016) were able to estimate that the average 20-year-old American native English-speaker knows around 42,000 words and 4,000 idioms (N=221,268). Hartshorne and colleagues showed that 40 years of theories about how people interpret pronouns failed to generalize beyond the stimuli used; new data sets with thousands of stimuli suggested a new theory with deep connections to theories of semantics (Hartshorne et al., 2015; Hartshorne & Snedeker, 2013).

Citizen Science for Developmental Psychology: Paradigms and Prospects

As noted above, currently the dominant paradigm for citizen science in psychology is the viral quiz. As illustrated by the work on lifespan development described above, such methods can be used to study children as young as 8 or 9. Viral quizzes can also be used to make retrospective inferences about development by studying adults now, as exemplified by the aforementioned studies of critical periods (Chen & Hartshorne, 2021; Hartshorne et al., 2018) or effects of childhood adversity on adult cognition (Germine et al., 2015). In some cases, parents can be induced to help their children participate in such quizzes. For instance, Hilton and colleagues (2021) successfully encouraged parents of nearly 5,000 children as young as 3 to have their children complete a music identification quiz.

However, the viral quiz has obvious limitations when it comes to developmental psychology. Small children are generally not motivated to participate in quizzes, and quizzes are often not a particularly effective assay of their behavior and cognition, particularly for the

youngest children. Researchers have been actively developing paradigms more suited to developmental psychology.

One promising avenue starts with the observation that every day caregivers are collectively recording truly massive amounts of cognitive and behavioral data about their children — likely more than has been collected by all developmental psychologists to date. This includes parents collecting videos of their children moving and speaking, parents tracking child behavior and milestones through apps like Baby Manager or The Wonder Weeks, and children playing tablet and phone games or choosing content to stream. In many cases, this data is being freely donated to commercial companies. A few labs have begun experimenting with recruiting parents to donate the same data to science instead. Addyman & Addyman (2013) studied the development of laughter in infancy by soliciting videos of babies laughing from more than 500 parents across 25 countries, along with information about the context. Hartshorne, Huang and colleagues developed an app (kidtalkscrapbook.org) for parents to record and transcribe linguistic data during the pandemic (Hartshorne et al., 2021; KidTalk, 2020). In a lower-tech but rapidly deployable variant on this theme, Srinivasan and colleagues asked parents to collect daily audio-recordings their babies at bathtime (providing a sample of their linguistic interactions) and fill out brief on-line surveys gauging parents' worries and mood (Ellwood-Lowe et al., *in prep*).

Another promising direction is to build on the appeal of electronic games to preschoolers. Most games are designed to probe and test human cognitive abilities (this is part of what makes them fun). Just as importantly, whereas it can be a struggle to recruit subjects into the laboratory for a half-hour study, the same people will willingly *pay money* to spend dozens of hours playing a given game. Psychologists are increasingly using performance on both commercial and custom-built games to study cognition among adults (Brändle et al., 2021; Coutrot et al., 2022; Stafford & Haasnoot, 2017; Stafford & Vaci, 2021; Steyvers et al., 2019; van Opheusden et al., 2021; Vélez, 2021). For instance, Vélez (2021) studied cultural accumulation of knowledge by studying 25,060 players in One Hour One Life. Steyvers and colleagues capitalized on semi-longitudinal data from tens of thousands of users of the "brain games" site Luminosity to inform an unprecedentedly precise account of the factor structure of learning and practice effects (Steyvers et al., 2019; Steyvers & Schafer, 2020). Van Opheusden and colleagues partnered with a different "brain games" company to produce a custom-built variant of "tic-tac-toe" in order to statistically model how planning depth changes with expertise in a strategy game (van Opheusden et al., 2021). Brändle et al., 2022 analyzed data from a popular, non-goal-directed world-exploration game, allowing the development and testing of a new model of intrinsically motivated environment exploration. Finally, Stafford and Hassnoot (2017) capitalized on data from 1.2 million players of a complex planning and perception game to precisely measure skill consolidation during wake and sleep. Note that most of these examples are not citizen science in that the data was obtained from the gaming company not donated by the subjects themselves, though direct donation of game data from subjects is sometimes possible. However, these studies do illustrate the fact that electronic games produce enormous amounts of useful data about behavior and cognition.

Given the sheer popularity of games for young children (Nofziger, 2021), experiments designed as games should be plenty attractive to young subjects and their families. A quick browse through an app store reveals a vast range of behaviors that can occur in children's games. To date, there are only a handful of preliminary examples. For instance, Long, Fan, Chai, and Frank (2019) studied the development of children's ability to draw, collecting over 13,000 drawings from children ages 2–10 by installing an electronic elicited-drawing game in a children's museum. By comparing drawing and tracing ability, they showed that improvement in drawing was not entirely explained by developing visuo-motor skills and is likely explained in part by improvement in higher-level cognition.

Another promising and underutilized opportunity is recruiting citizen scientists to help process large datasets. As reviewed above, this is widely used in astronomy, zoology, and other fields (Lintott, 2019; Von Ahn, 2006). While such projects are not common in psychology, at least a few have produced important results, for instance confirming key but controversial predictions of modern linguistic theory (Hartshorne et al., 2014) and building precise, 3D descriptions of neurons, which confirmed that space–time wiring specificity supports direction selectivity in the retina (Kim et al., 2014). Developmental researchers are starting to take note. One recent pilot study recruited online volunteers to categorize nearly four hours of infant vocalizations, finding good correspondence between the results and those of expert annotators (Semenzin et al., 2020).

The four paradigms described above (viral quizzes, naturally-occurring data, games, and annotation projects) are the most obvious avenues at the moment, and they in principle allow a very wide range of studies. Even so, they are likely only the beginning.

Challenges

In principle, Internet-based citizen science can make use of data collected by any widely-used computer, tablet, phone, or wearable, including measures such as button-presses (with reaction time), mouse and track-pad tracking, drawing, voice responses, video, (coarse-grained) eyetracking, GPS, physical position, heart rate, skin conductance, to just name a few (Gjoreski et al., 2018; Gosling & Mason, 2015; Harari et al., 2016; Hartshorne, de Leeuw, et al., 2019; Miller, 2012; Papoutsaki et al., 2016; Yang & Krajbich, 2021). Moreover, the range keeps expanding, and may soon include, for instance, EEG (Badcock et al., 2013; Duvinage et al., 2013). Similarly, stimuli can include audio, video, and even rudimentary virtual reality (Huber & Gajos, 2020; Mottelson & Hornbæk, 2017). While there are some limitations in terms of the paradigms available (e.g., fMRI), this must be balanced against the fact that citizen science also allows for paradigms that are *not* feasible in the laboratory. In particular, because people take their phones and wearables everywhere, these devices can be used to measure behavior and cognition during real-life experiences. For instance, researchers have, for instance, used experience sampling via mobile phone apps to study real-time influences of happiness during daily experience (Killingsworth & Gilbert, 2010b; Kumar et al., 2014b). Such studies have ecological face validity in a way no in-lab experiment can match.

However, not all of these methods work equally well, and sometimes the measures may not be sufficiently precise, at least for the time being. For instance, while online eyetracking methods are sufficiently accurate for preferential looking or the Visual World Paradigm, they currently lack the precision for eyetracking-while-reading paradigms (Ariel et al., 2022; Murali & Çöltekin, 2021; Slim & Hartsuiker, 2021; Yang & Krajbich, 2021). Even where the precision is available, it is not always attained. A major area of current research and development is improving instrument calibration and ensuring proper use by subjects (Kritly et al., 2021; Li et al., 2020; Woods et al., 2017). In the meantime, there is an increasingly large tool bag of tricks for designing around instrument limitations (Hartshorne, de Leeuw, et al., 2019; Krantz, 2001; Passell et al., 2021). For instance, careful design of Visual World Paradigm experiments can work around limitations in the accuracy of Webcam-based eyetrackers (Figure 1).

Relatedly, one of the complications of citizen science software stems from one of citizen science's key advantages: subjects use the devices already available to them. Unfortunately, different people use different devices – and, even worse, different versions of different operating systems on different devices – and the software needs to be compatible with everything. In fact, not only must the software be compatible with the gamut of devices, but it must correct for the biases of those different devices. For example, Passell et al. (2021) found that people who use mobile devices have a significantly slower reaction time than computer users. Even when restricting to computers, reaction times can be biased slightly but measurably in different directions depending on the device (for review, see Hartshorne, de Leeuw, et al., 2019). These issues are addressable but militate against blindly writing software and assuming that it works on every device as expected.

A more fundamental problem is experimental design. The familiar protocols of lab-based studies are the results of decades of optimization of research methods for the exigencies and opportunities of in-lab studies. Not surprisingly, these methods do not always translate well to citizen science — because the tasks are too confusing, take too long, are not sufficiently interesting, etc. It is important to remember that these familiar paradigms are not the best way to study the research question, merely the best way to study the research question in a brick-and-mortar lab. Conversely, many of the most exciting citizen science studies to date took advantage of the unique affordances of the Internet to design experiments that are impossible in the laboratory and which address otherwise intractable questions (e.g., Salganik et al., 2006).

Researchers are actively developing methods optimized for citizen science (anecdotally, citizen science in psychology only really began to take off once the viral quiz paradigm was established and mastered). As noted above, while the viral experiment paradigm is unlikely to be particularly well-suited for studies with children under the age of 8 or 9, game-based studies are extremely child-friendly and may prove a powerful format (e.g. Long et al., 2019). Projects that involve popular parental activities are also promising (Addyman & Addyman, 2013; KidTalk, 2020). Until more paradigms are developed, creativity is sometimes required. If creativity fails, it may sometimes be worthwhile to study a different aspect of the question, one that is more amenable to current methods. Finally, it is certainly the case that some questions will probably never be amenable to citizen science, such as

studies involving pre-technological societies. (However, we note that this does not absolve the researcher from finding a method of ensuring replicability and generalizability.)

Merely developing a good experimental design, however, is only the first part. It also must be implemented. While there are an increasing number of software platforms for running developmental studies online (Lo et al., 2021; Rhodes et al., 2020; Scott & Schulz, 2017), they primarily support running compensated subjects through familiar in-lab paradigms. Moreover, they often fail to take advantage of the opportunities presented by citizen science, such as capturing semi-longitudinal data from participants who play a game a variable numbers of times with variable spacing between sessions. Thus, the overlap between what that software supports and what is required for citizen science is limited. While there are ongoing efforts to build more robust software for citizen science (Hartshorne, de Leeuw, et al., 2019; Trouille et al., 2020), usability remains sufficiently limited such that most psychology researchers conducting citizen science projects write their own software from scratch. This is particularly challenging for research teams that lack programming experience. (Note there is a similar challenge in analysis: the larger the dataset, the more valuable it is to process the data using code — e.g., in R or Python — from start to finish. Fortunately, at least in this case, there is an abundance of not just software but tutorials, classes, and help forums.)

A common concern about Internet-enabled citizen science studies — or, really, any Internet-enabled studies — is security. In principle, data stored on any computer connected to the Internet — which is to say, nearly all research data, whether collected online or in the lab — is at risk from hackers, but online databases of data collected through public-facing apps may be a more tempting target for hackers. The simplest option is to not collect any identifying information, rendering participation if anything more secure than an in-lab study, which can never be truly anonymous. Even where audio, video, or wearable sensor data is involved, it is sometimes possible to process the data immediately on the subjects' own device, not retaining anything identifiable (i.e., WebGazer uses a video camera for eyetracking, but processing is immediate and no images are stored; Papoutsaki et al., 2016). However, sometimes identifiable data is required, raising not just security issues but sometimes laws and regulations about data collection, storage, and sharing in different countries. These issues are not insurmountable: indeed, they affect nearly every segment of society, and there are robust methods for handling them (Majeed & Lee, 2020; Stopczynski et al., 2014). Addressing these issues can, however, require active effort on the part of the researcher. Even when security has been ensured, there may still be effort required to reassure the participants (Lo et al., 2021).

Another issue is sample bias. citizen science samples are far more diverse than in-lab samples, but they do not include everybody. Not only are certain populations systematically excluded (e.g., pre-technological societies), but citizen science studies attract subjects through intrinsic interest, and different people are intrinsically interested in different things (Jun et al., 2017). To be clear, this is not a reason to prefer the typical in-lab convenience sample – if everyone adopted citizen science, development psychology samples would be far more diverse both individually and in aggregate – but it does mean that citizen science is not sufficient to completely solve the problem of sample diversity (Lourenco & Tasimi, 2020).

Relatedly, care must be taken in designing studies for diverse populations. It is not by accident that many psychology studies involving adult subjects look like classroom exams. Exams are a familiar paradigm for researchers — many of whom are educators — and our traditional undergraduate subjects are by definition elite test-takers. Anecdotally, researchers often run into trouble when applying these same methods to individuals who are less familiar with Western-style classroom exams. Developmental psychologists are, by definition, more adept at working with a subject population that has less robust cultural expectations and norms, and thus we tend to rely somewhat less on their expectations about what to do in an experiment setting. However, while children's cultural expectations and knowledge are less developed than that of adults, they do have expectations. We strongly recommend piloting studies with populations of interest to get feedback, including manipulation checks (Hoewe, 2017) and other methods of confirming that subjects understood what they were supposed to do, and when possible consulting with researchers who have experience with each of the populations one hopes to include.

Lest these challenges seem insurmountable for the majority of researchers, we note that not that long ago, fMRI was similarly beyond the reach of most psychologists. Now it is a fairly normal part of being a cognitive psychologist and increasingly common even among developmental psychologists. Some decades earlier, hardly anyone had the expertise or resources to run computerized experiments. If a method is sufficiently powerful, both individual researchers and the field itself will adapt. Our central thesis in this paper is that citizen science is every bit as transformative as computerization or fMRI.

Concluding Remarks

Numerous mathematical and empirical studies show that enormous amounts of data are required to characterize any aspect of human psychology, but on reflection this is common sense. What makes human psychology so remarkable and so fascinating is its sheer complexity and flexibility. Unlike, say, electrons, each of us is different and behaves differently. This need for data is only magnified in development, which unfolds over the course of decades; characterizing this trajectory thus requires massive amounts of data collected every step of the way. Collecting data based on a few stimuli and a few dozen subjects at a time was always going to be a painfully slow way of making progress.

Citizen Science promises to accelerate progress by orders of magnitude. While certain kinds of questions remain out of reach — particularly those involving neural data or pre-technological societies — most behavioral questions can be addressed better and faster through Citizen Science than through traditional methods, at least in principle. The primary limitation is that we have limited experience with Citizen Science, both individually (most of us have never tried it) and collectively (as a field, we are only scratching the surface of what can be done). This means that each individual study is slower than we are used to, in part because the studies are much larger, but also because we are often creating key aspects of the method for the first time.

Thus, using traditional methods will generally lead to faster progress at first. Certainly, it is easier. But ultimately, it takes decades to accomplish what could be managed in a

single Citizen Science study. That payoff makes the investment in Citizen Science not just worthwhile, but necessary.

Acknowledgments

Work on this manuscript was supported by NSF 2030106, awarded to Joshua Hartshorne and NIH DP5OD024566, awarded to Samuel A. Mehr.

References

Addyman C, & Addyman I (2013). The science of baby laughter. Comedy Studies, 4 (2), 143–153.

Ariel J, Ryskin RA, Hartshorne JK, Backs H, Bala N, Barcenas-Meade L, Bhattarai S, Charles T, Copoulos G, Coss C, Eisert A, Furuhashi E, Ginel K, Guttman-McCabe A, Chaz H, Hoba L, Hwang W, Iannetta C, Koenig K, … de Leeuw JR What paradigms can webcam eye-tracking be used for? attempted replications of 5 “classic” cognitive science experiments. Symposium conducted at the meeting of Cognitive Development Society, Madison, WI. 2022, April.

Auxier B, Anderson M, Perrin A, & Turner E (2020). Parenting children in the age of screens. https://www.pewresearch.org/internet/2020/07/28/parenting-children-inthe-age-of-screens/

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon J-F, & Rahwan I (2018). The moral machine experiment. Nature, 563 (7729), 59–64. [PubMed: 30356211]

Baayen RH, Davidson DJ, & Bates DM (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language, 59 (4), 390–412.

Badcock NA, Mousikou P, Mahajan Y, De Lissa P, Thie J, & McArthur G (2013). Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps. PeerJ, 1, e38. [PubMed: 23638374]

Baillargeon R, Buttelmann D, & Southgate V (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. Cognitive Development, 46, 112–124.

Bakker M, Hartgerink CH, Wicherts JM, & van der Maas HL (2016). Researchers’ intuitions about power in psychological research. Psychological science, 27 (8), 1069–1077. [PubMed: 27354203]

Bergmann C, Tsuji S, Piccinini PE, Lewis ML, Braginsky M, Frank MC, & Cristia A (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. Child development, 89 (6), 1996–2009. [PubMed: 29736962]

Berinsky AJ, Margolis MF, & Sances MW (2014). Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. American Journal of Political Science, 58 (3), 739–753.

Bezeau S, & Graves R (2001). Statistical power and effect sizes of clinical neuropsychology research. Journal of clinical and experimental neuropsychology, 23 (3), 399–406. [PubMed: 11419453]

Bialystok E, & Kroll JF (2018). Can the critical period be saved? a bilingual perspective. Bilingualism: Language and Cognition, 21 (5), 908–910.

Birdsong D, & Molis M (2001). On the evidence for maturational constraints in second-language acquisition. Journal of Memory and language, 44 (2), 235–249.

Black A, & Bergmann C (2017). Quantifying infants’ statistical word segmentation: A meta-analysis. 39th Annual Meeting of the Cognitive Science Society, 124–129.

Blake KR, & Gangestad S (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. Personality and Social Psychology Bulletin, 46 (12), 1702–1711. [PubMed: 32208875]

Bleidorn W, Klimstra TA, Denissen JJ, Rentfrow PJ, Potter J, & Gosling SD (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. Psychological science, 24 (12), 2530–2540. [PubMed: 24142813]

Bleidorn W, Schönbrodt F, Gebauer JE, Rentfrow PJ, Potter J, & Gosling SD (2016). To live among like-minded others: Exploring the links between person-city personality fit and self-esteem. Psychological Science, 27 (3), 419–427. [PubMed: 26842317]

Bonney R, Shirk JL, Phillips TB, Wiggins A, Ballard HL, Miller-Rushing AJ, & Parrish JK (2014). Next steps for citizen science. Science, 343 (6178), 1436–1437. [PubMed: 24675940]

Brändle F, Allen KR, Tenenbaum J, & Schulz E (2021). Using games to understand intelligence. Proceedings of the Annual Meeting of the Cognitive Science Society, 43 (43).

Brändle F, Stocks LJ, Tenenbaum JB, Gershman SJ, & Schulz E (2022). Intrinsically motivated exploration as empowerment.

Brysbaert M, Stevens M, Mandera P, & Keuleers E (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. Frontiers in Psychology, 7 (1116), 1–11. [PubMed: 26858668]

Burnside K, Ruel A, Azar N, & Poulin-Dubois D (2018). Implicit false belief across the lifespan: Non-replication of an anticipatory looking task. Cognitive Development, 46, 4–11.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, & Munafò MR (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature reviews neuroscience, 14 (5), 365–376. [PubMed: 23571845]

Byers-Heinlein K, Bergmann C, Davies C, Frank MC, Hamlin JK, Kline M, Kominsky JF, Kosie JE, Lew-Williams C, Liu L, et al. (2020). Building a collaborative psychological science: Lessons learned from manybabies 1. Canadian Psychology/Psychologie canadienne, 61 (4), 349. [PubMed: 34219905]

Cafri G, Kromrey JD, & Brannick MT (2010). A meta-meta-analysis: Empirical review of statistical power, type i error rates, effect sizes, and model selection of meta-analyses published in psychology. Multivariate Behavioral Research, 45 (2), 239–270. [PubMed: 26760285]

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T, et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nature Human Behaviour, 2 (9), 637–644.

Cattell RB (1963). Theory of fluid and crystallized intelligence: A critical experiment. Journal of educational psychology, 54 (1), 1.

Chambers C (2019). What's next for registered reports? Nature, 573 (7773), 187–189. [PubMed: 31506624]

Chandler JJ, & Paolacci G (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. Social Psychological and Personality Science, 8 (5), 500–508.

Chari R, Matthews LJ, Blumenthal M, Edelman AF, & Jones T (2017). The promise of community citizen science. RAND.

Chase LJ, & Chase RB (1976). A statistical power analysis of applied psychological research. Journal of Applied Psychology, 61 (2), 234.

Chen T, & Hartshorne JK (2021). More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence. Cognition, 214, 104706. [PubMed: 34052616]

Chivers T (2019). What's next for psychology's embattled field of social priming. Nature, 576 (7786), 200–203. [PubMed: 31827289]

Chmielewski M, & Kucker SC (2020). An mturk crisis? shifts in data quality and the impact on study results. Social Psychological and Personality Science, 11 (4), 464–473.

Clark HH (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of verbal learning and verbal behavior, 12 (4), 335–359.

Clark-Carter D (1997). The account taken of statistical power in research published in the british journal of psychology. British Journal of Psychology, 88 (1), 71–83.

Cohen J (1962). The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology, 65 (3), 145. [PubMed: 13880271]

Cohen J (1994). The earth is round (p<. 05). American psychologist, 49 (12), 997.

Combs JG (2010). Big samples and small effects: Let's not trade relevance and rigor for power.

Cooper CB, Shirk J, & Zuckerberg B (2014). The invisible prevalence of citizen science in global research: Migratory birds and climate change. PloS one, 9 (9), e106508. [PubMed: 25184755]

Coutrot A, Manley E, Goodroe S, Gahnstrom C, Filomena G, Yesiltepe D, Dalton R, Wiener JM, Hölscher C, Hornberger M, et al. (2022). Entropy of city street networks linked to future spatial navigation ability. Nature, 1–7.

Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, Beebe J, Berniūnas R, Boudesseul J, Colombo M, et al. (2021). Estimating the reproducibility of experimental philosophy. Review of Philosophy and Psychology, 12 (1), 9–44.

D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, Chen C, Deaton J, Eisenstein J, Hoffman MD, et al. (2020). Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395.

Dick AS, Garcia NL, Pruden SM, Thompson WK, Hawes SW, Sutherland MT, Riedel MC, Laird AR, & Gonzalez R (2019). No evidence for a bilingual executive function advantage in the abcd study. Nature human behaviour, 3 (7), 692–701.

Difallah D, Filatova E, & Ipeirotis P (2018). Demographics and dynamics of mechanical turk workers. Proceedings of the eleventh ACM international conference on web search and data mining, 135–143.

Dodell-Feder D, & Germine LT (2018). Epidemiological dimensions of social anhedonia. Clinical Psychological Science, 6 (5), 735–743.

Dörrenberg S, Rakoczy H, & Liszkowski U (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. Cognitive Development, 46, 12–30.

Duvinage M, Castermans T, Petieau M, Hoellinger T, Cheron G, & Dutoit T (2013). Performance of the emotiv epoc headset for p300-based applications. Biomedical engineering online, 12 (1), 1–15. [PubMed: 23289769]

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, Baranski E, Bernstein MJ, Bonfiglio DB, Boucher L, et al. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67, 68–82.

Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR,Chartier CR, Corker KS, Corley M, Hartshorne JK, IJzerman H, et al. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. Advances in Methods and Practices in Psychological Science, 3 (3), 309–331.

Ellwood-Lowe ME, Foushee R, Horton G, Wehry J, & Srinivasan M (in prep). Exploring the impact of adversity on parents' child-directed speech: Day-to-day variability during the covid-19 pandemic.

Embretson SE, & Reise SP (2013). Item response theory. Psychology Press.

Evans N, & Levinson SC (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and brain sciences, 32 (5), 429–448. [PubMed: 19857320]

Fortenbaugh FC, DeGutis J, Germine LT, Wilmer JB, Grosso M, Russo K, & Esterman M (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. Psychological science, 26 (9), 1497–1510. [PubMed: 26253551]

Fraley RC, & Vazire S (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. PloS one, 9 (10), e109019. [PubMed: 25296159]

Francis G (2012). The psychology of replication and replication in psychology. Perspectives on Psychological Science, 7 (6), 585–594. [PubMed: 26168115]

Frank MC, Bergelson E, Bergmann C, Cristia A, Floccia C, Gervain J, Hamlin JK, Hannon EE, Kline M, Levelt C, et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Infancy, 22 (4), 421–435. [PubMed: 31772509]

Frank MC, Braginsky M, Yurovsky D, & Marchman VA (2017). Wordbank: An open repository for developmental vocabulary data. Journal of child language, 44 (3), 677–694. [PubMed: 27189114]

Funder DC, & Ozer DJ (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2 (2), 156–168.

Gebauer JE, Bleidorn W, Gosling SD, Rentfrow PJ, Lamb ME, & Potter J (2014). Cross-cultural variations in Big Five relationships with religiosity: A sociocultural motives perspective. Journal of personality and social psychology, 107 (6), 1064–1091. [PubMed: 25180757]

Gebauer JE, Sedikides C, Schonbrodt FD, Bleidorn W, Rentfrow PJ, Potter J, & Gosling SD (2016). Religiosity as social value: Replication and extension. Journal of Personality and Social Psychology, 1–74.

Germine LT, Duchaine B, & Nakayama K (2011). Where cognitive development and aging meet: Face learning ability peaks after age 30. Cognition, 118 (2), 201–210. [PubMed: 21130422]

Germine LT, Dunn EC, McLaughlin KA, & Smoller JW (2015). Childhood adversity is associated with adult theory of mind and social affiliation, but not face processing. PloS one, 10 (6), e0129612. [PubMed: 26068107]

Germine LT, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, & Wilmer JB (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. Psychonomic Bulletin & Review, 19 (5), 847–857. 10.3758/s13423-012-0296-9 [PubMed: 22829343]

Germine LT, Reinecke K, & Chaytor NS (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. The Clinical Neuropsychologist, 33 (2), 271–286. [PubMed: 30614374]

Gilmore RO (2016). From big data to deep insight in developmental science. Wiley Interdisciplinary Reviews: Cognitive Science, 7 (2), 112–126. [PubMed: 26805777]

Gjoreski M, Luštrek M, & Pejović V (2018). My watch says i'm busy: Inferring cognitive load with low-cost wearables. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 1234–1240.

Gosling SD, & Mason W (2015). Internet research in psychology. Annual review of psychology, 66, 877–902.

Gosling SD, Vazire S, Srivastava S, & John OP (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. American psychologist, 59 (2), 93. [PubMed: 14992636]

Halberda J, Ly R, Wilmer JB, Naiman DQ, & Germine LT (2012). Number sense across the lifespan as revealed by a massive internet-based sample. Proceedings of the National Academy of Sciences, 109 (28), 11116–11120.

Hall D, Huerta MF, McAuliffe MJ, & Farber GK (2012). Sharing heterogeneous data: The national database for autism research. Neuroinformatics, 10 (4), 331–339. [PubMed: 22622767]

Hampshire A, Highfield RR, Parkin BL, & Owen AM (2012). Fractionating human intelligence. Neuron, 76 (6), 1225–1237. [PubMed: 23259956]

Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, & Gosling SD (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. Perspectives on Psychological Science, 11 (6), 838–854. [PubMed: 27899727]

Harris KM, Halpern CT, Whitsel EA, Hussey JM, Killeya-Jones LA, Tabor J, & Dean SC (2019). Cohort profile: The national longitudinal study of adolescent to adult health (add health). International Journal of Epidemiology, 48 (5), 1415–1415k. [PubMed: 31257425]

Hartshorne JK, Bonial C, & Palmer M (2014). The verbcorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 397–402.

Hartshorne JK, de Leeuw JR, Goodman ND, Jennings M, & O'Donnell TJ (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. Behavior research methods, 51 (4), 1782–1803. [PubMed: 30746644]

Hartshorne JK, & Germine LT (2015). When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span. Psychological science, 26 (4), 433–443. [PubMed: 25770099]

Hartshorne JK, Huang YT, Paredes PML, Oppenheimer K, Robbins PT, & Velasco MD (2021). Screen time as an index of family distress. Current Research in Behavioral Sciences, 2, 100023.

Hartshorne JK, O'Donnell TJ, & Tenenbaum JB (2015). The causes and consequences explicit in verbs. Language, cognition and neuroscience, 30 (6), 716–734. [PubMed: 26052518]

Hartshorne JK, & Schachner A (2012). Tracking replicability as a method of post-publication open evaluation. Frontiers in computational neuroscience, 6, 8. [PubMed: 22403538]

Hartshorne JK, Skorb L, Dietz SL, Garcia CR, Iozzo GL, Lamirato KE, Ledoux JR, Mu J, Murdock KN, Ravid J, et al. (2019). The meta-science of adult statistical word segmentation: Part 1. Collabra: Psychology, 5 (1).

Hartshorne JK, & Snedeker J (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. Language and Cognitive Processes, 28 (10), 1474–1508.

Hartshorne JK, Tenenbaum JB, & Pinker S (2018). A critical period for second language acquisition: Evidence from 2/3 million english speakers. Cognition, 177, 263–277. [PubMed: 29729947]

Henrich J (2020). The weirdest people in the world: How the west became psychologically peculiar and particularly prosperous. Penguin UK.

Henrich J, Heine SJ, & Norenzayan A (2010). Most people are not weird. Nature, 466 (7302), 29–29. [PubMed: 20595995]

Hilton C, Crowley L, Yan R, Martin A, & Mehr S (2021). Children infer the behavioral contexts of unfamiliar foreign songs.

Hilton C, & Mehr S (in press). Citizen science can help to alleviate the generalizability crisis. Behavioral and Brain Sciences.

Hoewe J (2017). Manipulation check. The international encyclopedia of communication research methods, 1–5.

Huber B, & Gajos KZ (2020). Conducting online virtual environment experiments with uncompensated, unsupervised samples. Plos one, 15 (1), e0227629. [PubMed: 31999696]

Ioannidis JP (2012). Why science is not necessarily self-correcting. Perspectives on Psychological Science, 7 (6), 645–654. [PubMed: 26168125]

Irwin V, Zhang J, Wang X, Hein S, Wang K, Roberts A, York C, Barmer A, Mann FB, Dilig R, Parker S, & Nachazel T (2021). Report on the condition of education 2021. United States Department of Education.

ITU Telecommunication Development Sector. (2019). Measuring digital development: Facts and figures. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2020.pdf

Janghorban R, Roudsari RL, & Taghipour A (2014). Skype interviewing: The new generation of online synchronous interview in qualitative research. International journal of qualitative studies on health and well-being, 9 (1), 24152. [PubMed: 24746247]

John LK, Loewenstein G, & Prelec D (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science, 23 (5), 524–532. [PubMed: 22508865]

Johnson W, Logie RH, & Brockmole JR (2010). Working memory tasks differ in factor structure across age cohorts: Implications for dedifferentiation. Intelligence, 38 (5), 513–528.

Jones BC, DeBruine LM, Flake JK, Liuzza MT, Antfolk J, Arinze NC, Ndukaihe IL, Bloxsom NG, Lewis SC, Foroni F, et al. (2021). To which world regions does the valence–dominance model of social perception apply? Nature human behaviour, 5 (1), 159–169.

Judd CM, Westfall J, & Kenny DA (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. Journal of personality and social psychology, 103 (1), 54. [PubMed: 22612667]

Jun E, Hsieh G, & Reinecke K (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. Proceedings of the ACM on Human-Computer Interaction, 1 (CSCW), 1–15.

Kan IP, & Drummey AB (2018). Do imposters threaten data quality? an examination of worker misrepresentation and downstream consequences in amazon's mechanical turk workforce. Computers in Human Behavior, 83, 243–253.

Kidd E, & Garcia R (2021). How diverse is child language acquisition?

KidTalk. (2020). Kidtalkscrapbook.org [Accessed: 2021-09-30].

Killingsworth MA, & Gilbert DT (2010a). A wandering mind is an unhappy mind. Science, 330 (6006), 932–932. [PubMed: 21071660]

Killingsworth MA, & Gilbert DT (2010b). A wandering mind is an unhappy mind. Science, 330 (6006), 932–932. [PubMed: 21071660]

Kim JS, Greene MJ, Zlateski A, Lee K, Richardson M, Turaga SC, Purcaro M, Balkam M, Robinson A, Behabadi BF, Campos M, Denk W, Seung SH, & the EyeWires. (2014). Space–time wiring specificity supports direction selectivity in the retina. Nature, 509 (7500), 331–336. [PubMed: 24805243]

Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB Jr, Alper S, Aveyard M, Axt JR, Babalola MT, Bahnık Š, et al. (2018). Many labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science, 1 (4), 443–490.

Krantz JH (2001). Stimulus delivery on the web: What can be presented when calibration isn't possible. Dimensions of Internet science, 113–130.

Kritly L, Basecq V, Glorieux C, & Rychtáriková M (2021). Challenges on level calibration of online listening test: A proposed subjective method. EuroNoise 2021, Madeira, Portugal.

Kulke L, Reiß M, Krist H, & Rakoczy H (2018). How robust are anticipatory looking measures of theory of mind? replication attempts across the life span. Cognitive Development, 46, 97–111.

Kulke L, von Duhn B, Schneider D, & Rakoczy H (2018). Is implicit theory of mind a real and robust phenomenon? results from a systematic replication study. Psychological science, 29 (6), 888–900. [PubMed: 29659340]

Kumar A, Killingsworth MA, & Gilovich T (2014a). Waiting for Merlot: Anticipatory consumption of experiential and material purchases. Psychological Science, 25 (10), 1924–1931. [PubMed: 25147143]

Kumar A, Killingsworth MA, & Gilovich T (2014b). Waiting for merlot: Anticipatory consumption of experiential and material purchases. Psychological science, 25 (10), 1924–1931. [PubMed: 25147143]

LeBel EP, McCarthy RJ, Earp BD, Elson M, & Vanpaemel W (2018). A unified framework to quantify the credibility of scientific findings. Advances in Methods and Practices in Psychological Science, 1 (3), 389–402.

Li Q, Joo SJ, Yeatman JD, & Reinecke K (2020). Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. Scientific reports, 10 (1), 1–11. [PubMed: 31913322]

Lillard AS, Lerner MD, Hopkins EJ, Dore RA, Smith ED, & Palmquist CM (2013). The impact of pretend play on children's development: A review of the evidence. Psychological bulletin, 139 (1), 1. [PubMed: 22905949]

Lillard AS, Pinkham AM, & Smith E (2011). Pretend play and cognitive development.

Lintott C (2019). The crowd and the cosmos: Adventures in the zooniverse. Oxford University Press.

Liu J, Hilton CB, Bergelson E, & Mehr SA (2021). Language experience shapes music processing across 40 tonal, pitch-accented, and non-tonal languages. bioRxiv. 10.1101/2021.10.18.464888

Lo CH, Mani N, Kartushina N, Mayor J, & Hermes J (2021). E-babylab: An open-source browser-based tool for unmoderated online developmental studies.

Long B, Fan J, Chai Z, & Frank MC (2019). Developmental changes in the ability to draw distinctive features of object categories. Proceedings of the 41st Annual Conference of the Cognitive Science Society.

Lourenco SF, & Tasimi A (2020). No participant left behind: Conducting science during covid-19. Trends in Cognitive Sciences, 24 (8), 583–584. [PubMed: 32451239]

MacWhinney B (2000). The childes project: The database (Vol. 2). Psychology Press.

Maddock JE (1999). Statistical power and effect size in the field of health psychology. University of Rhode Island.

Majeed A, & Lee S (2020). Anonymization techniques for privacy preserving data publishing: A comprehensive survey. IEEE Access, 9, 8512–8545.

Maniaci MR, & Rogge RD (2014). Caring about carelessness: Participant inattention and its effects on research. Journal of Research in Personality, 48, 61–83.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. Advances in Methods and Practices in Psychological Science, 3 (1), 24–52.

Marjanovic Z, Struthers CW, Cribbie R, & Greenglass ER (2014). The conscientious responders scale: A new tool for discriminating between conscientious and random responders. Sage Open, 4 (3), 2158244014545964.

Maxwell SE (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. Psychological methods, 9 (2), 147. [PubMed: 15137886]

Maxwell SE, Lau MY, & Howard GS (2015). Is psychology suffering from a replication crisis? what does “failure to replicate” really mean? American Psychologist, 70 (6), 487. [PubMed: 26348332]

Maylor EA, & Logie RH (2010). A large-scale comparison of prospective and retrospective memory development from childhood to middle age. Quarterly Journal of Experimental Psychology, 63 (3), 442–451.

Meade AW, & Craig SB (2012). Identifying careless responses in survey data. Psychological methods, 17 (3), 437. [PubMed: 22506584]

Mehr SA, Singh M, Knox D, Ketter DM, Pickens-Jones D, Atwood S, Lucas C, Jacoby N, Egner AA, Hopkins EJ, et al. (2019). Universality and diversity in human song. Science, 366 (6468).

Miller G (2012). The smartphone psychology manifesto. Perspectives on psychological science, 7 (3), 221–237. [PubMed: 26168460]

Mone MA, Mueller GC, & Mauland W (1996). The perceptions and usage of statistical power in applied psychology and management research. Personnel psychology, 49 (1), 103–120.

Moriguchi Y (2021). Beyond bias to western participants, authors, and editors in developmental science. Infant and Child Development, e2256.

Moss AJ, Rosenzweig C, Robinson J, & Litman L (2020). Demographic stability on mechanical turk despite covid-19. Trends in cognitive sciences, 24 (9), 678–680. [PubMed: 32553445]

Mottelson A, & Hornbæk K (2017). Virtual reality studies outside the laboratory. Proceedings of the 23rd acm symposium on virtual reality software and technology, 1–10.

Murali M, & Çöltekin A (2021). Conducting eye tracking studies online. Proceedings of the Workshop on Adaptable Research Methods for Empirical Research with Map Users, Virtual Workshop, 6.

Newell A (1973). You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium.

Nielsen M, Haun D, Kärtner J, & Legare CH (2017). The persistent sampling bias in developmental psychology: A call to action. Journal of Experimental Child Psychology, 162, 31–38. [PubMed: 28575664]

Nisbett R (2004). The geography of thought: How asians and westerners think differently… and why. Simon; Schuster.

Nofziger H (2021). 2021: Deconstructing mobile and tablet gaming. %7Bhttps://www.npd.com/lps/pdf/npd-2021-mgr-preview.pdf%7D

Nosek BA, Banaji MR, & Greenwald AG (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dynamics: Theory, Research, and Practice, 6 (1), 101.

Nosek BA, Spies JR, & Motyl M (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science, 7 (6), 615–631. [PubMed: 26168121]

Oakes LM (2017). Sample size, statistical power, and false conclusions in infant looking-time research. Infancy, 22 (4), 436–469. [PubMed: 28966558]

Olderbak S, Wilhelm O, Hildebrandt A, & Quoidbach J (2019). Sex differences in facial emotion perception ability across the lifespan. Cognition and Emotion, 33 (3), 579–588. [PubMed: 29564958]

Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Nybo Andersen A-M, Taxbøl D, Hansen KD, Juhl M, Schow TB, et al. (2001). The danish national birth cohort-its background, structure and aim. Scandinavian journal of public health, 29 (4), 300–307. [PubMed: 11775787]

Open Science Collaboration. (2015). Psychology. estimating the reproducibility of psychological science. Science, 349 (6251), aac4716. [PubMed: 26315443]

Oppenheimer DM, Meyvis T, & Davidenko N (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of experimental social psychology, 45 (4), 867–872.

Osborne JW (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. Educational psychology, 28 (2), 151–160.

Paap K (2019). The bilingual advantage debate: Quantity and quality of the evidence. The handbook of the neuroscience of multilingualism, 701–735.

Papoutsaki A, Sangkloy P, Laskey J, Daskalova N, Huang J, & Hays J (2016). Webgazer: Scalable webcam eye tracking using user interactions. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016.

Pashler H, & Harris CR (2012). Is the replicability crisis overblown? three arguments examined. Perspectives on Psychological Science, 7 (6), 531–536. [PubMed: 26168109]

Passell E, Dillon DG, Baker JT, Vogel SC, Scheuer LS, Mirin NL, Rutter LA, Pizzagalli DA, & Germine LT (2019). Digital cognitive assessment: Results from the testmybrain nimh research domain criteria (rdoc) field test battery report.

Passell E, Strong RW, Rutter LA, Kim H, Scheuer L, Martini P, Grinspoon L, & Germine LT (2021). Cognitive test scores vary with choice of personal digital device. Behavior Research Methods, 53 (6), 2544–2557. [PubMed: 33954913]

Peterson JC, Bourgin DD, Agrawal M, Reichman D, & Griffiths TL (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. Science, 372 (6547), 1209–1214. [PubMed: 34112693]

Poulin-Dubois D, & Yott J (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. Developmental science, 21 (4), e12600. [PubMed: 28952180]

Powell LJ, Hobbs K, Bardis A, Carey S, & Saxe R (2018). Replications of implicit theory of mind tasks with varying representational demands. Cognitive Development, 46, 40–50.

Raddick MJ, Bracey G, Gay PL, Lintott CJ, Murray P, Schawinski K, Szalay AS, & Vandenberg J (2009). Galaxy zoo: Exploring the motivations of citizen science volunteers. arXiv preprint arXiv:0909.2925.

Reinecke K, & Gajos KZ (2014). Quantifying visual preferences around the world. Proceedings of the SIGCHI conference on human factors in computing systems, 11–20.

Reinecke K, & Gajos KZ (2015). Labinthewild: Conducting large-scale online experiments with uncompensated samples. Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, 1364–1378.

Reñosa MDC, Mwamba C, Meghani A, West NS, Hariyani S, Ddaaki W, Sharma A, Beres LK, & McMahon S (2021). Selfie consents, remote rapport, and zoom debriefings: Collecting qualitative data amid a pandemic in four resource-constrained settings. BMJ global health, 6 (1), e004193.

Rhodes M, Rizzo MT, Foster-Hanson E, Moty K, Leshin RA, Wang M, Benitez J, & Ocampo JD (2020). Advancing developmental science via unmoderated remote research with children. Journal of Cognition and Development, 21 (4), 477–493. [PubMed: 32982602]

Richard FD, Bond CF Jr, & Stokes-Zoota JJ (2003). One hundred years of social psychology quantitatively described. Review of general psychology, 7 (4), 331–363.

Riley E, Okabe H, Germine LT, Wilmer J, Esterman M, & DeGutis J (2016). Gender differences in sustained attentional control relate to gender inequality across countries. PloS one, 11 (11), e0165100. [PubMed: 27802294]

Robins RW, Tracy JL, Trzesniewski K, Potter J, & Gosling SD (2001). Personality correlates of self-esteem. Journal of research in personality, 35 (4), 463–482.

Robins RW, Trzesniewski KH, Tracy JL, Gosling SD, & Potter J (2002). Global self-esteem across the life span. Psychology and aging, 17 (3), 423. [PubMed: 12243384]

Robinson J, Rosenzweig C, Moss AJ, & Litman L (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. PloS one, 14 (12), e0226394. [PubMed: 31841534]

Rossi JS (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. What if there were no significance tests, 175–197.

Salganik MJ, Dodds PS, & Watts DJ (2006). Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311 (5762), 854–856. [PubMed: 16469928]

Schäfer T, & Schwarz MA (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in Psychology, 10, 813. [PubMed: 31031679]

Scheel AM, Schijen MR, & Lakens D (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. Advances in Methods and Practices in Psychological Science, 4 (2), 25152459211007467.

Scott K, & Schulz L (2017). Lookit (part 1): A new online platform for developmental research. Open Mind, 1 (1), 4–14.

Sedlmeier P, & Gigerenzer G (1992). Do studies of statistical power have an effect on the power of studies?

Semenzin C, Hamrick L, Seidl A, Kelleher B, & Cristia A (2020). Towards large-scale data annotation of audio from wearables: Validating Zooniverse annotations of infant vocalization types. %7Bwww.npd.com/lps/pdf/npd-2021-mgr-preview.pdf%7D

Shanks DR, & Vadillo MA (2021). Publication bias and low power in field studies on goal priming. Royal Society open science, 8 (10), 210544. [PubMed: 34667618]

Sheskin M, & Keil F (2018). Thechildlab. com a video chat platform for developmental research.

Simmons-Duffin S (2021). A citizen scientist gave the cdc a head start in a covid-19 outbreak investigation. https://www.npr.org/2021/08/05/1025248628/a-citizen-scientist-gave-the-cdc-a-head-start-in-a-covid-19-outbreak-investigati

Singleton D, & Leśniewska J (2021). The critical period hypothesis for l2 acquisition: An unfalsifiable embarrassment? Languages, 6 (3), 149.

Slim MS, & Hartsuiker R (2021). Visual world eyetracking using webgazer. js.

Soto CJ, John OP, Gosling SD, & Potter J (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. Journal of personality and social psychology, 100 (2), 330. [PubMed: 21171787]

Srivastava S, John OP, Gosling SD, & Potter J (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? Journal of personality and social psychology, 84 (5), 1041. [PubMed: 12757147]

Stafford T, & Haasnoot E (2017). Testing sleep consolidation in skill learning: A field study using an online game. Topics in cognitive science, 9 (2), 485–496. [PubMed: 27868362]

Stafford T, & Vaci N (2021). Digital games as a platform for understanding skill acquisition from novice to expert.

Stanley TD, Carter EC, & Doucouliagos H (2018). What meta-analyses reveal about the replicability of psychological research. Psychological bulletin, 144 (12), 1325. [PubMed: 30321017]

Stewart N, Ungemach C, Harris AJ, Bartels DM, Newell BR, Paolacci G, & Chandler J (2015). The average laboratory samples a population of 7,300 amazon mechanical turk workers. Judgment and Decision making, 10 (5), 479–491.

Steyvers M, Hawkins GE, Karayanidis F, & Brown SD (2019). A large-scale analysis of task switching practice effects across the lifespan. Proceedings of the National Academy of Sciences, 116 (36), 17735–17740.

Steyvers M, & Schafer RJ (2020). Inferring latent learning factors in large-scale cognitive training data. Nature Human Behaviour, 4 (11), 1145–1155.

Stopczynski A, Pietri R, Pentland A, Lazer D, & Lehmann S (2014). Privacy in sensor-driven human data collection: A guide for practitioners. arXiv preprint arXiv:1403.5299.

Strange AM, Enos RD, Hill M, & Lakeman A (2019). Online volunteer laboratories for human subjects research. PloS one, 14 (8), e0221676. [PubMed: 31461488]

Su I-A, & Ceci S (2021). “zoom developmentalists”: Home-based videoconferencing developmental research during covid-19.

Szucs D, & Ioannidis JP (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS biology, 15 (3), e2000797. [PubMed: 28253258]

Trouille L, Lintott C, & Fortson L (2020). Zooniverse project builder platform. American Astronomical Society Meeting Abstracts #235, 235, 287–04.

Trouton A, Spinath FM, & Plomin R (2002). Twins early development study (teds): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. Twin Research and Human Genetics, 5 (5), 444–448.

Turner AM, Engelsma T, Taylor JO, Sharma RK, & Demiris G (2020). Recruiting older adult participants through crowdsourcing platforms: Mechanical turk versus prolific academic. AMIA Annual Symposium Proceedings, 2020, 1230.

UNICEF et al. (2020). How many children and young people have internet access at home?: Estimating digital connectivity during the covid-19 pandemic (tech. rep.). UNICEF.

Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TE, Bucholz R, Chang A, Chen L, Corbetta M, Curtiss SW, et al. (2012). The human connectome project: A data acquisition perspective. Neuroimage, 62 (4), 2222–2231. [PubMed: 22366334]

Vankov I, Bowers J, & Munafò MR (2014). Article commentary: On the persistence of low power in psychological science. Quarterly journal of experimental psychology, 67 (5), 1037–1040.

van Opheusden B, Galbiati G, Kuperwajs I, Bnaya Z, Ma W-J, et al. (2021). Revealing the impact of expertise on human planning with a two-player board game.

Vélez N (2021). Multigenerational innovation and division of labor in online communities. Talk presented at the 43rd Annual Conference of the Cognitive Science Society.

Von Ahn L (2006). Games with a purpose. Computer, 39 (6), 92–94.

Ward RM (2002). Highly significant findings in psychology: A power and effect size survey. University of Rhode Island.

Weisberg DS (2015). Pretend play. Wiley Interdisciplinary Reviews: Cognitive Science, 6 (3), 249–261. [PubMed: 26263228]

West J (2000). America's kindergartners: Findings from the early childhood longitudinal study, kindergarten class of 1998–99, fall 1998. US Department of Education, Office of Educational Research; Improvement …

Westgate E, Riskind R, & Nosek B (2015). Implicit preferences for straight people over lesbian women and gay men weakened from 2006 to 2013. Collabra: Psychology, 1 (1).

Wiesmann CG, Friederici AD, Disla D, Steinbeis N, & Singer T (2018). Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false belief-related action anticipation. Cognitive Development, 46, 58–68. [PubMed: 30147231]

Wilmer JB, Germine LT, Ly R, Hartshorne JK, Kwok H, Pailian H, Williams MA, & Halberda J (2012). The heritability and specificity of change detection ability. Journal of Vision, 12 (9), 1275–1275.

Wilson BM, Harris CR, & Wixted JT (2020). Science is not a signal detection problem. Proceedings of the National Academy of Sciences, 117 (11), 5559–5567.

Woods KJ, Siegel MH, Traer J, & McDermott JH (2017). Headphone screening to facilitate web-based auditory experiments. Attention, Perception, & Psychophysics, 79 (7), 2064–2072.

Yang X, & Krajbich I (2021). Webcam-based online eye-tracking for behavioral research. Judgment and Decision Making, 16 (6), 1486.

Yarkoni T (2019). The generalizability crisis. Behavioral and Brain Sciences, 1–37.

Ye T, Reinecke K, & Robert LP (2017). Personalized feedback versus money: The effect on reliability of subjective data in online experimental platforms. Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 343–346.

Youyou W, Stillwell D, Schwartz HA, & Kosinski M (2017). Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. Psychological science, 28 (3), 276–284. [PubMed: 28059682]

Additional Resources: There is a growing ecosystem of online tutorials, textbooks, and message groups providing detailed advice on developing online citizen science projects. While none are currently geared towards developmental psychology, many will nonetheless be helpful to developmental psychologists. We list below several that are current as of writing, but we encourage readers to seek out new, more up-to-date resources as they become available.

- Moving Research Online (4-session video tutorial series from 2020; bit.ly/3EFhBGx)
- CogSci 2020 Workshop on Scaling Cognitive Science (1-day seminar; bit.ly/3MvygiC)
- Pushkin Gitbook (Tutorial/Documentation for Pushkin software for massive online psychological experiments; languagelearninglab.gitbook.io/pushkin/)
- Online Experiments Google Group (message group; groups.google.com/g/online-experiments)
- Zooniverse Blog (https://blog.zooniverse.org/)

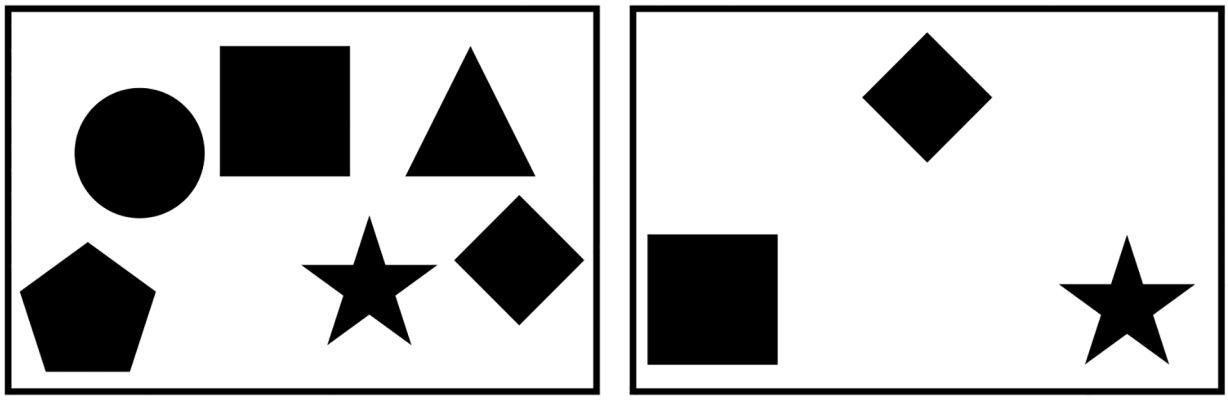

Figure 1.

Two possible layouts for a Visual World Paradigm trial. The goal of this experiment is to measure phonological cohort effects. Subjects hear auditory instructions to “click on the square.” While all subjects will ultimately look at the square, the question is whether the proportion of looks to objects whose name starts with “s” (the phonological competitors) will decline more slowly than those to objects whose name does not (the distractors). In the panel on the left, objects from the three conditions are intermixed and close to one another. Since Webcam eyetrackers are often only accurate to about 100 pixels, fixations to different objects may be difficult to distinguish. In the right panel, the display has been reduced to a single exemplar of each type, allowing them to be spaced far apart from one another, improving our ability to infer which object is being fixated.