Developmental psychologists should adopt citizen science to improve generalization and reproducibility

Wei Li | Laura Thi Germine | Samuel A. Mehr | Mahesh Srinivasan | Joshua Hartshorne

1Department of Psychology and Neuroscience, Boston College, Chestnut Hill, Massachusetts, USA
2McLean Hospital, Belmont, Massachusetts, USA
3Department of Psychiatry, Harvard Medical School, Cambridge, Massachusetts, USA
4Data Science Initiative, Harvard University, Cambridge, Massachusetts, USA
5School of Psychology, Victoria University of Wellington, Wellington, New Zealand
6Department of Psychology, University of California, Berkeley, California, USA

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 a greater scale and from more diverse populations. Luckily, this mechanism already exists as follows: 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.
Widespread failures of replication and generalization pose an 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?

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 (Frank, Braginsky, et al., 2017; Byers-Heinlein et al., 2020; Ebersole et al., 2016; Ebersole et al., 2020; 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 the 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.

1 | THE CRISSES OF REPLICATION AND GENERALIZATION

Systematic investigations in psychology and neuroscience typically report replication rates between 30% and 70% (Hartshorne, Skorb, et al., 2019; Camerer et al., 2018; Cova et al., 2021; Ebersole et al., 2020; 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).
1.1 | 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 an average-sized effect—an issue first identified in the 1960s and reconfirmed 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). 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. 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 an effect larger than 1/20 of a standard deviation, resulting in 20,794 subjects for a two-sample t-test. (Smaller effects are not only of limited practical 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 7000 subjects. The problem is actually worse if one uses Bayesian statistics, where 7000 subjects only give 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).

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 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.

1.2 | 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 the 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 4× 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.

1.3 | 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 be 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 & Lesniewska, 2021), the relative importance of pretend to play in children’s development (Lillard et al., 2011; Lillard et al., 2013; Weisberg, 2015), whether bilingualism affects executive function (Dick et al., 2019; Paap, 2019), and whether toddlers can succeed at ‘implicit’ theory of mind tasks (Kulke, Rei, et al., 2018; Kulke, von Duhn, et al., 2018; Baillargeon et al., 2018; Burnside et al., 2018; Dörrenberg et al., 2018; Poulin-Dubois & Yott, 2018; Powell et al., 2018; Wiesmann et al., 2018).

1.4 | Reconciling abysmal math with the actuality of progress

The gloomy assessment above might seem to preclude psychology from 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, and 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 the classic Gestalt perception effect or the fact that babies do not know the 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 dialogue 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).

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.

1.5 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 video conferencing era; Janghorban et al., 2014; Reñosa et al., 2021; Sheskin & Keil, 2018; Su & Ceci, 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 to collaborate on collecting data for a single study (Frank, Bergelson, et al., 2017; Byers-Heinlein et al., 2020; Ebersole et al., 2016; Ebersole et al., 2020; 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 the ideal. Moreover, the ‘many labs’ approach does not so much speed progress but rather concentrates 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’ approaches—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 = 7085; MacWhinney, 2000), 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 9000 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.

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.

2 | 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 outbreak (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).

2.1 | 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., 2015b; 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 et al., 2018; Hartshorne & Germine, 2015; 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 is enrolled in these platforms.

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 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%).

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 least 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. 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 gamified, 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.

### 2.2 Instrumentation and data quality

Large, diverse samples would be meaningless if one could not measure the behaviours of interest. There was certainly a point in time in which the machinery needed to measure human behaviour 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, drawing, voice responses, video, (coarse-grained) eye tracking, GPS, physical position, heart rate, and skin conductance, to just name a few (Hartshorne, de Leeuw, et al., 2019; Gjoreski et al., 2018; Gosling & Mason, 2015; Harari et al., 2016; 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 (Hartshorne, de Leeuw, et al., 2019; Germine et al., 2012; Gosling et al., 2004; 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/3Ob6IKI, and bit.ly/3Lvwqhj). As a result, data quality is high and shirking and dishonesty 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 are 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 is not 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...
The most compelling argument for using citizen science to study cognition and behaviour 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 & Gilbert, 2010; Kumar et al., 2014; Mehr et al., 2019; Riley et al., 2016; Salganik et al., 2006).

With regards to the 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, researchers have used popular online quizzes to track changes across much of the lifespan (usually roughly 8–80 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., 2014), 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 et al. (2012) found that while number sense abilities change across the lifespan and peak in the late 30s, individual differences remain very large in every age group. Johnson et al. (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 et al. (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 et al. (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 et al. (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 et al. (2015) found that childhood adversity negatively impacted the theory of mind and social affiliation but not face processing, suggesting the differential effects of the environment on different aspects of social cognition. Finally, Awad et al. (2018) collected judgements on 26 million trolley problems in 10 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, n.d.). For instance, by testing different subjects on different words, Brysbaert et al. (2016) were able to estimate that the average 20-year-old American native English speaker knows around 42,000 words and 4000 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).

4 | 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 et al. (2021) successfully encouraged parents of nearly 5000 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 behaviour 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 everyday caregivers are collectively recording truly massive amounts of cognitive and behavioural 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 behaviour 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 and 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 a 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 online surveys gauging parents' worries and mood (Ellwood-Lowe et al., n.d.).

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; Westgate et al., 2015a). For instance, Westgate et al. (2015a) studied the cultural accumulation of knowledge by studying 25,060 players in 1 h One Life. Steyvers and colleagues capitalized on semi-longitudinal data from tens of thousands of users of the ‘brain game’ 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) analysed 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 Haasnoot (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 a 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 behaviour 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 behaviours that can occur in children’s games. To date, there are only a handful of preliminary examples. For instance, Long et al. (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 4 h 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.

5 | 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) eye tracking, GPS, physical position, heart rate, skin conductance, to just name a few (Hartshorne, de Leeuw, et al., 2019; Gjoreski et al., 2018; Gosling & Mason, 2015; Harari et al., 2016; 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 behaviour 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, 2010; Kumar et al., 2014). 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 eye tracking methods are sufficiently accurate for preferential looking or the Visual World Paradigm, they currently lack the precision for eye tracking-while-reading paradigms (Ariel et al., 2022; Murali & Çöltekin, 2021; Slim & Hartsuiker, 2021; Yang & Krajbich, 2021). Even where 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 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 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 of 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 eye tracking, 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.

6 | 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 behavioural 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.

6.1 | 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; www.movingresearchonline.info/)
- CogSci 2020 Workshop on Scaling Cognitive Science (1-day seminar; bit.ly/3Ae1fX)
- 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/)
AUTHOR CONTRIBUTIONS
Wei Li: Writing – original draft; writing – review and editing. Laura Thi Germine: Writing – review and editing. Samuel Mehr: Writing – review and editing. Mahesh Srinivasan: Writing – review and editing. Joshua Hartshorne: Conceptualization; writing – original draft; writing – review and editing.

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

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2348.

ORCID
Joshua Hartshorne https://orcid.org/0000-0003-1240-3598

ENDNOTES
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 \[ P(H_0|\text{sig}) = \frac{P(H_0)P(\text{sig}|H_0)}{P(H_0)P(\text{sig}|H_0) + P(H_1)P(\text{sig}|H_1)} \]

\[ P(H_0|\text{sig}) \] 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 prior to the effect. \[ P(\text{sig}|H_0) \] is the Type I error. \[ P(\text{sig}|H_1) \] is the statistical power.

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 analysed with a t-test. For each sample size being considered, we conducted 5000 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 7000 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 to 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 than the typical researcher can obtain. It follows that studies with realistic power analyses are unlikely to be accepted as registered reports.

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).

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.
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).

These numbers are smaller for the percentage of individuals who have the 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.

Anecdotally, some researchers question suggests 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 backward to us.

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 and 4.

Since 2008, 53% of traffic to games with words with known origins has been from social media.

REFERENCES


Cohen, J. (1994). The earth is round (p < 0.05). American Psychologist, 49(12), 997–1003.


Maddock, J. E. (1999). *Statistical power and effect size in the field of health psychology*. University of Rhode Island.


How to cite this article: Li, W., Germine, L. T., Mehr, S. A., Srinivasan, M., & Hartshorne, J. (2022). Developmental psychologists should adopt citizen science to improve generalization and reproducibility. Infant and Child Development, e2348. https://doi.org/10.1002/icd.2348