Journal of Cognition and Development

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hjcd20

Broadening Convenience Samples to Advance Theoretical Progress and Avoid Bias in Developmental Science

Sabine Doebel & Michael C. Frank

To cite this article: Sabine Doebel & Michael C. Frank (12 Nov 2023): Broadening Convenience Samples to Advance Theoretical Progress and Avoid Bias in Developmental Science, Journal of Cognition and Development, DOI: 10.1080/15248372.2023.2270055

To link to this article: https://doi.org/10.1080/15248372.2023.2270055

JOURNAL OF COGNITION AND DEVELOPMENT
https://doi.org/10.1080/15248372.2023.2270055

Routledge
Taylor & Francis Group

Broadening Convenience Samples to Advance Theoretical Progress and Avoid Bias in Developmental Science

Sabine Doebel[a] and Michael C. Frank[b]

[a]George Mason University, Fairfax, US; [b]Stanford University, Stanford, US

Abstract

Diverse samples are valuable to the study of development, and to psychology more broadly. But convenience samples—typically recruited from local populations close to universities—are still the most widely used in developmental science, despite the fact that their use leads to a vast over-representation of Western, White, and high socio-economic status participants in our studies. Do convenience samples still have a place in our research? While diverse samples are crucial to advancing developmental science, policies designed to encourage recruitment of such samples may not always succeed in improving sample diversity in ways that will benefit our theories and reduce bias. Further, convenience samples are just, well, convenient – and because their costs are lower, they allow for faster and more precise research. We suggest three paths forward to resolve this tension: 1) use theory and observed variation to choose aspects of diversity to prioritize in a particular study; 2) use online methods as an important route to broaden which samples are convenient; and 3) work in teams to achieve "inconvenient" samples using many different convenience populations.

Developmental science, and psychology more broadly, has been grappling with the over reliance on convenience samples in which White, Western, educated, rich and industrialized individuals are overrepresented (so-called "W.E.I.R.D" samples, Henrich, Heine, & Norenzayan, 2010; Kidd & Garcia, 2022; Nielsen, Haun, Kärtner, & Legare, 2017). There is widespread agreement that it is critical to reduce reliance on this "narrow slice of humanity" in developmental research. First, diverse samples are necessary for generating good theories. Such theories pertain to humans in general rather than a single population. Generic language used in many research articles reflects this aspiration (DeJesus, Callanan, Solis, & Gelman, 2019). Second, reducing reliance on convenience samples will reduce bias. Use of convenience samples reinforces assumptions about who represents the prototypical person (Roberts, Bareket-Shavit, Dollins, Goldie, & Mortenson, 2020). It also biases our beliefs about human nature toward tendencies observed in convenience samples (Sears, 1986). A well-documented negative consequence of this over-reliance is the use of deficit models, in which observed differences between majority participants in convenience samples and others are construed as reflecting deficits (Akhtar & Jaswal, 2013; Gaskins, 2013; Miller-Cotto, Smith, Wang, & Ribner, 2022; Rogoff et al., 2017; Singh, 2022).

CONTACT Sabine Doebel sdoebel@gmu.edu Department of Psychology, George Mason University, Fairfax, US

One example of how using convenience samples can lead to incomplete or inaccurate theories and deficit thinking can be seen in research on pretend play, which has focused almost exclusively on Western middle-class families, with overrepresentation of US families. This focus has contributed to theoretical accounts of pretend play that emphasize its role in the development of cognitive capacities like imagination, creativity, and executive function (Weisberg, 2015; White & Carlson, 2016; see; Gaskins, 2013 for discussion). Ideas about what "mature" play looks like are also heavily influenced by the population in which these ideas have been studied (Haight, Wang, Fung, Williams, & Mintz, 1999; Thompson & Goldstein, 2019). As a result, children in non-US, non-Western or non-industrialized societies, whose pretend play can be less fantasy-oriented and less likely to involve caregivers (e.g., Chessa et al., 2013; Farver & Shin, 1997; Gaskins, 2013; Haight, Wang, Fung, Williams, & Mintz, 1999; Thibodeau-Nielsen, Gilpin, Nancarrow, Pierucci, & Brown, 2020), may be perceived as deficient in their pretend play and associated cognitive capacities (see Doebel & Lillard, 2023; Gaskins, 2013 for discussion).

Yet despite these two major negatives (i.e., inaccurate theories and bias), convenience samples are just that – convenient. And the more convenient recruitment is, the easier it is to achieve larger samples, increasing the precision of measurements and decreasing the chance of statistical false positives (Bergmann et al., 2018; Button et al., 2013; Frank et al., 2017). Thus, there is a tradeoff between recommendations to avoid convenience populations and the desire to make speedy progress toward completing studies with large enough samples to draw precise, well-powered inferences. Following the strong consensus in the field, we both believe in the need to increase diversity to improve theories and reduce bias. Yet current recommendations for increasing sample diversity neither acknowledge this tradeoff nor accomplish the broader goals. In the following section, we discuss some of these recommendations before turning to describing our proposed solutions.

Current recommendations that miss the mark or do too little

Recruit representative samples

One recommendation to increase diversity in research samples has been to recruit samples that reflect the sociodemographic makeup of the region – or for large scale studies, the country – in which the study is performed (Bornstein, Jager, & Putnick, 2013). This approach allows researchers to generalize their findings to the population in a particular area, which is important for policy-relevant research (Tipton, 2013; Tipton & Olsen, 2018). Yet, population representativeness is not typically the goal of basic research in developmental science. Rather, basic research aims to identify findings that generalize to all people, not regions, whether local or national. Moreover, basic research increasingly aims to analyze heterogeneity (Bryan, Tipton, & Yeager, 2021). A typical sample in experimental developmental psychology (e.g., Bergmann et al., 2018) would be far too small to analyze heterogeneity, even if it were population representative. Instead, a better way to navigate the cost tradeoffs associated with collecting diverse samples would be to collect a sample that targeted the hypothesized source of variation, as we argue below.

Further, insofar as geographically representative samples are lacking in theoretically-relevant diversity that can be analyzed, they will do nothing to dislodge deficit models. For example, the finding that socioeconomic status is negatively related to performance on

laboratory measures of executive functions has been explained in terms of low-quality environments contributing to brain-based executive function deficits in children from lower socioeconomic status backgrounds (e.g., Rosen et al., 2020). To challenge this interpretation, we need to investigate the specific population of interest in more depth, for example by studying variation in background knowledge and the task expectations of children from lower socio-economic status backgrounds (Doebel, 2020). Because their goal is generalizability of the global estimate, representative samples are not optimally designed for addressing deficit models.

Increase diversity of within-lab samples

Another popular recommendation has been for researchers to tailor their recruitment to increase diversity within individual studies by oversampling minority populations. While this recommendation may produce data that yield theoretically relevant insights, expecting researchers to substantially increase various kinds of diversity in their samples also presents significant challenges that require resources that researchers may not have at their disposal.

These challenges may be especially severe for researchers at institutions that have relatively limited financial and social resources to support their productivity (Way, Morgan, Larremore, & Clauset, 2019). The problem is compounded for researchers who conduct their work in a geographical region that has limited racial, ethnic, and/or socio-economic diversity. Similarly, some researchers may already expend their limited resources to study specific populations of interest (e.g., Black children, bilingual children), and requiring them to further diversify their samples may make it difficult for them to sustain their research agenda.

Importantly, there is a cost-benefit trade off here: even if one can recruit more diverse samples within a lab, it may come at the cost of fewer participants. That means either less precise estimates and more false positives, or else publishing fewer studies. In other words, this policy recommendation asks researchers to act directly against their own interests, which may be especially hard for lower-resourced researchers, who themselves tend to be from more diverse backgrounds (Morgan et al., 2022). Part of the issue here is that funding and promotion norms incentivize publication quantity over quality (Frank, 2019), and addressing these norms could reduce the professional risks or costs associated with pursuing diverse samples. While this is an important and complex long-term goal, increasing diversity in individual studies may not be immediately feasible for many.

Report sample demographics prominently to achieve theoretical synthesis

Recent editorial policies recommend or require prominent reporting of sample demographics, for example in article abstracts (SRCD Sociocultural Policy, 2020). Such reporting is intended to contextualize research samples and make explicit to whom the findings might be expected to generalize (see also Bornstein, Jager, & Putnick, 2013; Simons, Shoda, & Lindsay, 2017). But demographic reporting can also carry long-term benefits. For example, standardized reporting – when coupled with transparent data sharing – can lead to opportunities for theoretical synthesis across datasets (Singh et al., 2022).

Yet in the short term, reporting requirements are insufficient to develop more general theories of cognitive development because they apply only at the last stage of the scientific process. Most researchers will simply include the demographics of the sample they have already collected, rather than collecting a different sample. Similarly, journal policies may encourage researchers to use more precise language in their claims so as not to suggest findings generalize more broadly than is warranted by the data (e.g., "Children from US middle-class backgrounds do X" instead of "Children do X"). However, without changes in the "economics" of recruiting participants, reporting policies may lead to the publication of articles that prominently feature demographic information and statements about the same convenience populations that researchers made use of before the policies came into effect.

Summary

Achieving more diverse samples is important to advance theoretical goals and avoid bias, but many proposed solutions – mandates to recruit representative samples, to increase sample diversity, or to change reporting requirements – miss the mark, come with their own challenges, or do too little. We suggest that the field can best achieve these goals by: 1) using theory and observed variation to make thoughtful choices regarding diversity versus convenience; 2) expanding and refining our use of online testing to make recruiting diverse samples more convenient; and 3) working in teams to pool resources and achieve meaningful comparisons across large groups.

Using theory and observed variation to choose dimensions of diversity

Broadening beyond convenience samples is a valuable goal for the field; however, a given participant sample need not be diverse along all dimensions. Rather, decision making about sampling should be guided by theoretical considerations and prior data. Since it is expensive to gather more diverse samples, we should prioritize cases where: 1) generalization is a key goal, and 2): variation is expected.

First, whether or not variation is observed or expected, if a claim is made that some developmental phenomenon is universal in humans generally or in specific groups, then an appropriately diverse sample should be recruited and tested to provide support for the claim. For example, if it is claimed that infants make social evaluations and that this is universal (e.g., Hamlin, Wynn, & Bloom, 2007), it is not convincing to argue for universality on the basis of results from a sample of White babies from affluent, educated families. Whether or not we expect infants' social evaluation capacities to vary by socioeconomic status and culture, scientific claims require evidence. Researchers who intend for their research to be interpreted as reflecting insights into universal developmental processes need to sample accordingly or constrain their claims explicitly (Simons, Shoda, & Lindsay, 2017).

In addition, we can prioritize cases where we expect variation, based on observation and/or theory. While our expectations are not always correct, observed or theoretically expected variation is a good place to start when resources are limited. Although no aspect of development is completely invariant to experience, some are known to vary so widely across contexts that differences are immediately apparent (e.g., caregiving environments; Keller, 2018) while others are known to be relatively more constant despite some evidence for variation in carefully designed studies (e.g., color perception; Bosten, 2022).

There are many good examples of this sort of predictable variation. One recent example from the literature is related to the development of self-control. Recent theories have shifted away from explaining developing self-control and executive function in terms of endogenous factors toward greater consideration of the influence of social and contextual factors (Doebel, 2020; Doebel & Munakata, 2018; Gaskins & Alcalá, 2023; Miller-Cotto, Smith, Wang, & Ribner, 2022; Munakata & Michaelson, 2021), in part because of observed variation in when and how children exercise control in different cultural and social contexts (Kidd, Palmeri, & Aslin, 2013; Lamm et al., 2018). This theoretical interest, informed by observation, has prompted researchers to further target samples that are known to vary in socialization practices related to delaying in order to understand how culture may influence delay of gratification skills via children's beliefs, values, and habits (Munakata et al., 2020; Yanaoka et al., 2022). In one specific recent study, researchers tested how cultural expectations regarding delay of gratification in specific contexts may shape children's ability to delay via practice doing so. Children were recruited in Japan and the US, two cultures with different patterns of expectations about delaying in two contexts – food and gifts. In Japan, children are more practiced at waiting for food and not gifts, whereas in the US the opposite is true. Consistent with the hypothesis that cultural expectations shape delay skills, Japanese children tended to wait longer for food than gifts, while US children tended to wait longer for gifts than food (Yanaoka et al., 2022). If these differences in cultural practices had been neglected (e.g., Japanese children tested on food trials only), researchers might draw incorrect inferences about the children's general delay ability. This "targeted variation" approach is not new, and there are many other examples of research that samples across different groups (e.g., culture, ethnicity, language status, social structure) in order to gain theoretical insights that challenge prevailing accounts. Our main point here is that using observed variation to guide sampling decisions can help us make immediate, efficient progress in our goals of improving our theories and reducing bias.

Of course, there may be important variation that we have not *yet* observed, particularly if we are not making the effort to look beyond our typical samples and the culture within which we are immersed as researchers (Rogoff et al., 2017), and if we are not prominently reporting demographic information in our studies. We also know that literatures can be biased and are often not a great guide to where we might expect variation (Henrich, Heine, & Norenzayan, 2010). For example, a large literature may suggest a phenomenon is universal, but closer inspection may reveal that much of the literature reflects narrow sampling (Singh, 2022).

However, rather than assuming heterogeneity in all phenomena and concluding that more diverse samples are always the right way to allocate our resources, both theory and large-scale data collection can be used to examine sources of heterogeneity (e.g., ManyBabies Consortium, 2020; Coppock, Leeper, & Mullinix, 2018; Klein et al., 2018). For example, in young children's early vocabulary, children seem to learn more nouns than expected (a "noun bias"), but in some East Asian languages, this bias appears to be reduced or absent (Tardif, Gelman, & Xu, 1999). Based on empirical evidence of this type, theorists have proposed explanations highlighting both syntactic features and cultural features shared by these languages (Frank, Braginsky, Yurovsky, & Marchman, 2021), which can in turn guide targeted investigations of children's vocabulary in other languages that share either cultural or linguistic similarities. In sum, both data and theory can be used to help us allocate our limited recruitment resources most optimally.

Using online methods to make recruiting more diverse samples convenient

Although research with children has traditionally been performed face to face, cognitive developmentalists were recently forced to embrace online data collection during the COVID-19 pandemic that paused in-person data collection (Sheskin et al., 2020). Many eagerly awaited a return to normalcy where research could resume in person, but the convenience of online testing has continued to accelerate uptake of online testing platforms (e.g., Scott & Schulz, 2017; Scott, Chu, & Schulz, 2017). Research suggests that with a bit of care online data collection with young children can yield very comparable data to the same experiments performed in person in the lab (see e.g., Chuey et al., 2021; Chuey et al., under review, for discussion and meta-analysis; but see; Fong, Imuta, Redshaw, & Nielsen, 2021; Kirkorian, 2018). Here we argue that online data collection should become a permanent part of our developmental science methodological toolkit, in part because of its immense potential to broaden the convenience samples available to researchers to include children diverse in socioeconomic status, race, ethnicity, culture, and nationality (Sheskin et al., 2020).

Just like in-person convenience samples, online convenience samples vary in their sociodemographic makeup depending on the recruitment channels that researchers use. Some studies using US-based convenience recruiting have reported more socioeconomic diversity (e.g., Scott, Chu, & Schulz, 2017), but others still see an over-representation of high socio-economic status participants from majority groups. For example, deMayo et al. (2021) aggregated online administrations of a common parent-report instrument for measuring children's early vocabulary (the MacArthur-Bates Communicative Development Inventory) across several labs. Pooling these US convenience samples revealed an over-representation of highly-educated White families. However, they conducted systematic outreach and online recruitment within US communities of color, using survey vendors (Prolific) and targeted social media advertising (Facebook). These efforts yielded more diverse samples, although the researchers had to navigate several issues to ensure task comprehension and comparable data quality.

Despite the promise of collecting developmental data online, there remain numerous psychological and infrastructural barriers to making this a widespread practice that can achieve meaningfully diverse samples. These include lingering skepticism about the validity of data collected online, and growing pains inherent in changing established practices. Online data collection can be particularly vulnerable to fraudulent responses (Chmielewski & Kucker, 2020; Storozuk, Ashley, Delage, & Maloney, 2020; Webb & Tangney, 2022). And while unmoderated data collection without a live experimenter present may have the most promise for achieving diverse samples by making participation convenient and unintimidating for first-time participants (Buhrmester, Talaifar, & Gosling, 2018; Sheskin et al., 2020), it may also be especially vulnerable to bots and fraudulent respondents who threaten the integrity of the data. However, these challenges are surmountable. There are, for example, many strategies to limit these threats, including remote video capture and data quality checks. We believe that as online data collection becomes more widespread and data quality is evaluated and demonstrated (e.g., via projects such as ManyBabies-At Home, which will compare infant data collected at home and in the lab), the field will become more comfortable with data collected online.

It is not a given that online data collection means more diversity (Lourenco & Tasimi, 2020). Attracting diverse participants to devote time to developmental research requires questioning assumptions about what participants understand about and seek from the research experience, and the amount of time they are willing to commit to it. Historically, the prototypical family who participates in cognitive developmental research is highly motivated and available to contribute to research and to learn about how their child thinks. Correspondingly, studies are often pitched as providing insights into how children think, and tokens for participating often include "Child scientist" certificates or t-shirts to commemorate children's participation. Pitching studies to a more diverse audience requires learning more about the families that are sought and adapting to them. For example, some families may be put off by the emphasis on science and understanding children's thinking, and may be more interested in studies that are fun for the child. Researchers may also need to reconsider the kinds of questions they seek to ask in a single study if they seek more diverse samples. Online testing that limits the duration of studies may attract the largest and most diverse samples, although there is not yet good empirical data on what is the most desirable duration for a given age group or population.

Another issue deserving of more consideration in the field is that of digital access and the challenge of including and representing those who do not have computer, tablet or smart phone access. Online testing excludes many in the US, in the global south, in non-industrialized societies, and elsewhere who do not have digital access, and this is a problem. This persistent issue is nevertheless not specific to online testing, as laboratory-based testing also requires access to technology to make contact with potential participants. We thus do not mean to suggest that online testing is a panacea and that it does not come without its own representational challenges, but rather that we are hopeful that it can open up new opportunities for increasing sample diversity compared to the status quo.

It is possible that continuing to pursue online testing without making substantial advances in appealing to more diverse families could worsen rather than ameliorate the field's reliance on convenience samples; however, there are already some efforts that suggest that this may not be the case (e.g., Rizzo, Britton, & Rhodes, 2022). We suggest that by addressing the foregoing issues and other potential barriers to online participation, we can make inconvenient samples more convenient.

Working in teams allows pooling of resources and samples to achieve meaningful comparisons across large groups

Teamwork among labs, despite its own unique challenges, can meaningfully broaden convenience samples. Given the trade-off between recruiting more diverse samples and convenience (cost), pooling samples from different labs that vary on dimensions of theoretical interest (e.g., culture, language status) can be a helpful alternative to achieve theoretically-relevant variation that can be analyzed.

For example, consortia like ManyBabies allow pooling of many local convenience samples across cultures. Despite vast linguistic variation, ManyBabies 1—a 67-lab study of infant-directed speech preferences in infants – found limited heterogeneity in experimental effects. Specifically, data were collected from labs in North America, Europe, Australia, and Asia using three common methods for measuring infants' discrimination (head-turn preference, central fixation, and eye tracking), and results indicated a small but

robust preference for infant directed speech across culture and paradigms, with some modulation of these effects by age, native language, and testing procedure (The ManyBabies Consortium, 2020). A follow-up ManyBabies project examined whether these findings generalized to bilingual infants, finding that infants with less exposure to North American English showed a weaker preference for infant directed speech (Byers-Heinlein et al., 2021). Another in-progress ManyBabies project will extend this work to examine preference for infant directed speech in African infants via a large-scale, multisite study of African infants (Tsui et al., 2022,accepted pending data collection).

The Child Language Data Exchange System (CHILDES; MacWhinney, 2014) hosts data from dozens of languages from around the world, allowing researchers to take advantage of diverse data collected by researchers over the past fifty years. Open collaborative enterprises like CHILDES as well as newer efforts like Wordbank (Frank et al., 2017) and Databrary (Gilmore & Adolph, 2017) allow fields to pool their efforts to create more diverse and representative samples. An embrace of data sharing – especially around specific, developmentally-relevant data types and constructs – can bear substantial fruit in efforts to understand the shape of development around the world. For example, in work with the Wordbank dataset, we were able to quantify some of the vast cross-cultural variation in the "noun bias" in early language, a phenomenon once thought to be universal (Frank, Braginsky, Yurovsky, & Marchman, 2021).

Collaborative team science efforts such as consortia and repositories can be a powerful way to pool effort to create more diverse samples, but they are not without their challenges (Coles, Hamlin, Sullivan, Parker, & Altschul, 2022). As more such collaborations emerge, new norms will be necessary for ensuring that contributors – especially junior collaborators and collaborators from less-resourced parts of the world – get credit for their contributions. Further, some of the concerns we raised above still apply: any individual dataset or consortium project may include a series of convenience samples that do not represent the most theoretically relevant aspects of sociodemographic diversity. Nevertheless, we believe these approaches have the potential to allow individual investigators to help navigate tradeoffs between convenience and diversity.

Conclusion

Convenience samples were initially embraced in psychology because they were, well, convenient. Yet convenience and narrow sampling are two different things. There is strong consensus that the narrow sampling that persists in our field should be addressed in order to advance theories and reduce bias. But the solution is not simply that every lab should now seek broadly diverse samples or samples that represent the specific region from which they are drawn. Rather, to improve our theories and reduce bias, we suggest the field should pursue diverse samples in light of theory and observation, and through online methods and multi-lab collaborations.

While our focus has been on how we can address our goals by making inconvenient samples more convenient, it is also important to recognize that there are systemic reasons for the overreliance on convenience samples that may be challenging to address but would make it less professionally costly to pursue inconvenient samples. Feedback loops between historical reliance on convenience samples and professional expectations around publication have contributed to the notion that convenience samples are indispensable if one is to

be hired, promoted, and funded. Such loops can be disrupted if universities, hiring committees, and funders rely less on crude metrics like number of publication and more on evaluation of the quality of scientific research – including how it contributes to our understanding of human diversity.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Akhtar, N., & Jaswal, V. K. (2013). Deficit or difference? Interpreting diverse developmental paths: An introduction to the special section. *Developmental Psychology*, *49*(1), 1–3. doi:10.1037/a0029851

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. doi:10.1111/cdev.13079

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357–370. doi:10.1016/j.dr.2013.08.003

Bosten, J. M. (2022). Do you see what I see? Diversity in human color perception. *Annual Review of Vision Science*, *8*(1), 101–133. doi:10.1146/annurev-vision-093020-112820

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. doi:10.1038/s41562-021-01143-3

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149–154. doi:10.1177/1745691617706516

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J. … Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920974622. doi:10.1177/2515245920974622

Chessa, D., Lis, A., Riso, D. D., Delvecchio, E., Mazzeschi, C., Russ, S. W., & Dillon, J. (2013). A cross-cultural comparison of pretend play in U.S. and Italian children. *Journal of Cross-Cultural Psychology*, *44*(4), 640–656. doi:10.1177/0022022112461853

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. doi:10.1177/1948550619875149

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T. & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, *4968*. doi:10.3389/fpsyg.2021.734398

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. Conducting developmental research online vs. in-person: A meta-analysis. Manuscript under review. 10.31234/osf.io/qc6fw

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. doi:10.1038/d41586-022-00150-2

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, *115*(49), 12441–12446. doi:10.1073/pnas.1808083115

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, *116*(37), 18370–18377. doi:10.1073/pnas.1817706116

deMayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C. F. . . . Marchman, V. (2021). *Web-CDI: A system for online administration of the MacArthur-bates communicative development inventories*. Language Development Research.

Doebel, S. (2020). Rethinking executive function and its development. *Perspectives on Psychological Science*, *15*(4), 942–956. doi:10.1177/1745691620904771

Doebel, S., & Lillard, A. S. (2023). How does play foster development? A new executive function perspective. *Developmental Review*, *67*, 101064. doi:10.1016/j.dr.2022.101064

Doebel, S., & Munakata, Y. (2018). Group influences on engaging self-control: Children delay gratification and value it more when their in-group delays and their out-group doesn't. *Psychological Science*, *29*(5), 738–748. doi:10.1177/0956797617747367

Farver, J. A. M., & Shin, Y. L. (1997). Social pretend play in Korean- and Anglo-American preschoolers. *Child Development*, *68*(3), 544–556. doi:10.2307/1131677

Fong, F. T., Imuta, K., Redshaw, J., & Nielsen, M. (2021). The digital social partner: Preschool children display stronger imitative tendency in screen-based than live learning. *Human Behavior and Emerging Technologies*, *3*(4), 585–594. doi:10.1002/hbe2.280

Frank, M. C. (2019). N-best evaluation for academic hiring and promotion. *Trends in Cognitive Sciences*, *23*(12), 983–985. doi:10.1016/j.tics.2019.09.010

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J. & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. doi:10.1111/infa.12182

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. Cambridge, MA: MIT Press.

Gaskins, S. (2013). Pretend play as culturally constructed activity. *The Oxford Handbook of the Development of Imagination*, 224–247.

Gaskins, S., & Alcalá, L. (2023). Studying executive function in culturally meaningful ways. *Journal of Cognition and Development*, *24*(2), 260–279. doi:10.1080/15248372.2022.2160722

Gilmore, R. O., & Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nature Human Behaviour*, *1*(7), 1–2. doi:10.1038/s41562-017-0128

Haight, W. L., Wang, X. L., Fung, H. H. T., Williams, K., & Mintz, J. (1999). Universal, developmental, and variable aspects of young children's play: A cross-cultural comparison of pretending at home. *Child Development*, *70*(6), 1477–1488. doi:10.1111/1467-8624.00107

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450* (7169), 557–559. doi:10.1038/nature06288

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. doi:10.1017/S0140525X0999152X

Keller, H. (2018). Universality claim of attachment theory: Children's socioemotional development across cultures. *Proceedings of the National Academy of Sciences*, *115*(45), 11414–11419. doi:10.1073/pnas.1720325115

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42* (6), 01427237211066405.

Kidd, C., Palmeri, H., & Aslin, R. N. (2013). Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, *126*(1), 109–114. doi:10.1016/j.cognition.2012.08.004

Kirkorian, H. L. (2018). When and how do interactive digital media help children connect what they see on and off the screen? *Child Development Perspectives*, *12*(3), 210–214. doi:10.1111/cdep.12290

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr, Alper, S. & Sowden, W. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. doi:10.1177/2515245918810225

Lamm, B., Keller, H., Teiser, J., Gudi, H., Yovsi, R. D., Freitag, C. & Lohaus, A. (2018). Waiting for the second treat: Developing culture-specific modes of self-regulation. *Child Development*, *89*(3), e261–e277. doi:10.1111/cdev.12847

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584. doi:10.1016/j.tics.2020.05.003

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk*. Volume II: The database. Psychology Press.

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. doi:10.1177/2515245919900809

Miller-Cotto, D., Smith, L. V., Wang, A. H., & Ribner, A. D. (2022). Changing the conversation: A culturally responsive perspective on executive functions, minoritized children and their families. *Infant and Child Development*, *31*(1), e2286. doi:10.1002/icd.2286

Morgan, A. C., LaBerge, N., Larremore, D. B., Galesic, M., Brand, J. E., & Clauset, A. (2022). Socioeconomic roots of academic faculty. *Nature Human Behaviour*, *6*(12), 1625–1633. doi:10.1038/s41562-022-01425-4

Munakata, Y., & Michaelson, L. E. (2021). Executive functions in social context: Implications for conceptualizing, measuring, and supporting developmental trajectories. *Annual Review of Developmental Psychology*, *3*(1), 139–163. doi:10.1146/annurev-devpsych-121318-085005

Munakata, Y., Yanaoka, K., Doebel, S., Guild, R. M., Michaelson, L. E. . . . Moses, L. (2020). Group influences on children's delay of gratification: Testing the roles of culture and personal connections. *Collabra: Psychology*, *6*(1). doi:10.1525/collabra.265

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. doi:10.1016/j.jecp.2017.04.017

Rizzo, M. T., Britton, T. C., & Rhodes, M. (2022). Developmental origins of anti-Black bias in white children in the United States: Exposure to and beliefs about racial inequality. *Proceedings of the National Academy of Sciences*, *119*(47), e2209129119. doi:10.1073/pnas.2209129119

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. doi:10.1177/1745691620927709

Rogoff, B., Coppens, A. D., Alcalá, L., Aceves-Azuara, I., Ruvalcaba, O., López, A., & Dayton, A. (2017). Noticing learners' strengths through cultural research. *Perspectives on Psychological Science*, *12*(5), 876–888. doi:10.1177/1745691617718355

Rosen, M. L., Hagen, M. P., Lurie, L. A., Miles, Z. E., Sheridan, M. A., Meltzoff, A. N., & McLaughlin, K. A. (2020). Cognitive stimulation as a mechanism linking socioeconomic status with executive function: A longitudinal investigation. *Child Development*, *91*(4), e762–e779. doi:10.1111/cdev.13315

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind*, *1*(1), 15–29. doi:10.1162/OPMI_a_00001

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14. doi:10.1162/OPMI_a_00002

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515. doi:10.1037/0022-3514.51.3.515

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S. & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678. doi:10.1016/j.tics.2020.06.004

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. doi:10.1177/1745691617708630

Singh, L. (2022). From information to action: A commentary on Kidd and Garcia (2022). *First Language, 42(*6), 814–817.

Singh, L., Barokova, M., Baumgartner, H., Lopera, D., Omane, P. O., Sheskin, M. & Frank, M. C. (2022). A unified approach to demographic data collection for research with young children across diverse cultures. Retrieved from https://psyarxiv.com/agt3d/

Singh, L., Rajendra, S. J., & Mazuka, R. (2022). Diversity and representation in studies of infant perceptual narrowing. *Child Development Perspectives*, *16*(4), 191–199. doi:10.1111/cdep.12468

Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got bots? Practical recommendations to protect online survey data from bot attacks. *The Quantitative Methods for Psychology*, *16*(5), 472–481. doi:10.20982/tqmp.16.5.p472

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and mandarin. *Child Development*, *70*(3), 620–635. doi:10.1111/1467-8624.00045

Thibodeau-Nielsen, R. B., Gilpin, A. T., Nancarrow, A. F., Pierucci, J. M., & Brown, M. M. (2020). Fantastical pretense's effects on executive function in a diverse sample of preschoolers. *Journal of Applied Developmental Psychology*, *68*, 101137. doi:10.1016/j.appdev.2020.101137

Thompson, B. N., & Goldstein, T. R. (2019). Disentangling pretend play measurement: Defining the essential elements and developmental progression of pretense. *Developmental Review*, *52*, 24–41. doi:10.1016/j.dr.2019.100867

Tipton, E. (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109–139. doi:10.1177/0193841X13516324

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524. doi:10.3102/0013189X18781522

Tsui, A. S. M., Carstensen, A., Kachergis, G., Abubakar, A., Asnake, M., & Barry, O. & Frank, M. C. (2022). Exploring variation in infants' preference for infant-directed speech: Evidence from a multi-site study in Africa. Stage 1 registered report, accepted pending data collection.

Way, S. F., Morgan, A. C., Larremore, D. B., & Clauset, A. (2019). Productivity, prominence, and the effects of academic environment. *Proceedings of the National Academy of Sciences*, *116*(22), 10729–10733. doi:10.1073/pnas.1817431116

Webb, M. A., & Tangney, J. P. (2022). Too good to be true: Bots and bad data from mechanical Turk. *Perspectives on Psychological Science*, 17456916221120027. doi:10.1177/17456916221120027

Weisberg, D. S. (2015). Pretend play. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(3), 249–261. doi:10.1002/wcs.1341

White, R. E., & Carlson, S. M. (2016). What would Batman do? self-distancing improves executive function in young children. *Developmental Science*, *19*(3), 419–426. doi:10.1111/desc.12314

Yanaoka, K., Michaelson, L. E., Guild, R. M., Dostart, G., Yonehiro, J., Saito, S., & Munakata, Y. (2022). Cultures crossing: The power of habit in delaying gratification. *Psychological Science*, *33*(7), 1172–1181. doi:10.1177/09567976221074650