Are larger studies always better? Sample size and data pooling effects in research communities

David Waszek, Cyrille Imbert

Are larger studies always better?
Sample size and data pooling effects in research communities

David Waszek* and Cyrille Imbert*

* Archives Henri-Poincaré – Philosophie et Recherches sur les Sciences et les Technologies (CNRS & Université de Lorraine, UMR 7117).

Manuscript presented at PSA 2022: The 28th Biennial Meeting of the *Philosophy of Science Association* (Pittsburgh, PA, November 10-13, 2022)
PhilSci Archive no. 21110, http://philsci-archive.pitt.edu/id/eprint/21110

Abstract. The persistent pervasiveness of inappropriately small studies in empirical fields is regularly deplored in scientific discussions. Consensually, taken individually, higher-powered studies are more likely to be truth-conducive. However, are they also beneficial for the wider performance of truth-seeking communities? We study the impact of sample sizes on collective exploration dynamics under ordinary conditions of resource limitation. We find that large collaborative studies, because they decrease diversity, can have detrimental effects in certain realistic circumstances that we characterize precisely. We show how limited inertia mechanisms may partially solve this pooling dilemma and discuss our findings briefly in terms of editorial policies.

Keywords. Bala-Goyal-Zollman model. Sample size. Small studies. Low power. Scientific collaboration. Data pooling. Resource limitation. Individual/collective discrepancy. Scientific diversity. Rational inertia. Editorial policies.

4770 words

1. Introduction. Sample size is a central parameter at all stages of empirical investigations, from their production to their evaluation by reviewers and their reception by their readership. The issue is both epistemological (the quality of studies depends on their epistemic power) and ethical (patients should not be involved in studies for weak, if not null, epistemic gains). In this context, scientific communities tend to determine normative conventions concerning appropriate sample sizes. Unfortunately, a consensus on this issue is lacking.

Since the 60s, the literature on sample sizes has been dominated by worries about the persistent pervasiveness of "underpowered" studies in the empirical sciences, for instance in the behavioral sciences, ecology, or evolutionary biology (Cohen 1962; Sedlmeier and Gigerenzer 1989; Button et al. 2013a; Vankov, Bowers, and Munafò 2014; Smaldino and McElreath 2016). Small samples are said to contribute to the lower reliability of the empirical sciences, and they may be a core factor in the so-called "reproducibility crisis" (Button et al. 2013b). Hence repeated calls for publication norms requiring higher sample sizes, and often for deeper reforms of the incentive structure of science (Higginson and Munafò 2016). Other authors counter that, in some cases, well-conducted small studies may be intrinsically preferable (Smith and Little 2018). Further, some good data is always better than no data: even information about one single animal, if suitably interpreted, can be very valuable (Fries and Maris 2021). In this perspective, the real problem may be the uncritical and exclusive use of $p < 0.05$ as a measure of the evidence and, more broadly, of Null Hypothesis Statistical Testing (NHST) (Bacchetti 2013). Indiscriminate sample size requirements would then be inappropriate. Instead, one should encourage, e.g., editorial practices requiring the preregistration of studies and their data analysis plans, the publication of data with relevant distributional statistics (Trafimow and Marks 2015), or other statistical paradigms such as Bayesianism (Rouder et al. 2009).

While essential, these discussions primarily investigate how *individual studies* can be made reliable—and, consensually, higher-powered studies are more likely to be truth-conducive. However, individual scientists and inquiries are fallible and should also be seen as intermediate instruments within broader scientific processes, which are expected to converge towards more reliable final results (Laudan 1981; Romero 2016). Accordingly, specific scientific practices, such as norms about sample sizes, should not be evaluated merely *locally*—according to their effect on isolated studies—but also *globally*—according to their effects on the broader performance of truth-seeking communities (Kitcher 1990). Low sample sizes may then turn out to be problematic in specific collective contexts only, e.g., when combined with publication practices that "filter out" data below significance thresholds and thereby spoil meta-analyses (Romero and Sprenger 2021).

Be this as it may, before studies can be aggregated through meta-analysis, they need to be conducted in the first place, which depends on preliminary choices by scientists to investigate this or that particular hypothesis. This is the question we investigate here. A recurrent risk in exploratory contexts is that hypotheses are discarded prematurely, based on limited data. This leads to an exploration dilemma, aspects of

which have already been studied by recent papers based on a common versatile model (Zollman 2007; 2010; Rosenstock, Bruner, and O'Connor 2017; Frey and Šešelja 2020): when previous studies suggest that a research hypothesis is less promising than alternate ones, should scientists keep collecting data about it?

Here, we pursue this investigation with the additional idea that empirical resources are limited both at the individual and communal level in terms of available data, number of published studies, and research time. After reviewing relevant results in the literature (Section 2), we show that the obvious suggestion of pooling empirical resources can sometimes be detrimental (Section 3), which creates a mismatch between what is good for individual studies and for scientific communities. After highlighting the importance of timescales in this discussion, we emphasize how inertia mechanisms can slacken these tensions (Sections 4-5) and discuss the scope of these results and their implications for science policy (Section 6).

2. The BGZ Model and Sample Sizes. To study the impact of sample sizes on hypotheses exploration, we turn to the Bala-Goyal-Zollman (hereafter BGZ) model (Zollman 2007; 2010). A number N of Bayesian agents, arranged in a communication network, compare two hypotheses, A and B, corresponding to two actions (typically, administering drug A or B to a patient) with success probabilities $p_A=.5$ and $p_B=p_A+\varepsilon$. We suppose that agents know p_A (as in Zollman, 2007) and start off with random (non-extreme) priors about p_B. In each round, those who believe $p_B > p_A$ perform action B n times (our "study size") and communicate their number of successes and failures to their neighbors, while the others do nothing, perhaps using their resources elsewhere. At the end of each round, agents use the data about B they and their neighbors gathered to update their belief distributions over p_B, which, following (Zollman 2010), we model by beta distributions. The question—typical of exploration dilemmas—is whether the community will devote enough resources to B to discover that $p_B > p_A$.

The communication structure of the model can be seen as a publication and readership network. As scientists usually read major publications in their fields, but not everything else beyond these, we focus here on two idealized structures, complete graphs (all agents are connected together) and wheel-shaped graphs (agents are connected on a circle and to an extra central agent), taking them as limiting cases between which actual communities lie.

The BGZ model is usually studied by running simulations for a large number of rounds and computing their success rate (i.e., the proportion of communities that finally agree that $p_B > p_A$) and their average convergence speed.[1] One then explores how these metrics depend on other parameters. For instance, well-connected communities can be misled by unlucky early runs of data, whereas less connected ones are *more* accurate but slower because they preserve diversity for longer and thereby

1 Here, we say that a community has "converged" when either (1) all agents believe A is better, and B is no longer investigated), or (2) all are confident that B is better, in the sense that their mean estimate for p_B is more than two standard deviations above p_A.

secure successful convergence (Zollman 2007, 584; Rosenstock, Bruner, and O'Connor 2017, §3).

The literature about the BGZ model does not systematically discuss the issue of sample sizes. However, it appears that *everything else being equal*, when the number *n* of trials per agent per round increases, so do success rates (Rosenstock, Bruner, and O'Connor 2017, fig. 3), an effect we reproduced across a wide range of parameter values. This is unsurprising: larger studies are individually less misleading and therefore less likely to lead communities astray.

However, should we jump to the conclusion that, individually *and* collectively, higher-powered studies are always preferable? Existing investigations are potentially uninformative here. First, they vary the size of studies indiscriminately, whereas resource limitations imply that large studies may come at the expense of smaller ones. We investigate this "pooling dilemma" below. Second, success rates after large numbers of rounds are irrelevant. Thus, we investigate success rates in reasonable time frames to assess *when* large studies may be preferable.

Importantly, our exploration below integrates a prominent criticism addressed to the BGZ model, namely that its results hold only in a limited (and possibly unrealistic) parameter range "in which learning is especially difficult" (Rosenstock, Bruner, and O'Connor 2017, 235); the focus on small effect sizes (typically, $\varepsilon = p_B - p_A = .001$), especially, is seen as a problem (Frey and Šešelja 2020, 1416). We present our results for an effect size of 2% ($\varepsilon = .02$) and sample sizes *n* in the 20–100 range, which we believe are reasonable.[2] (All our results are averaged over 10,000 simulations.) While space is missing for details, note that these results *are* robust for ε between .0001 and .1, as long as *n* is scaled appropriately. If such effect sizes still seem low, consider the following. First, the only source of errors in the standard BGZ model is the random variation of experimental trials, without any additional noise or systematic study biases (as in, e.g., (Holman and Bruner 2017)). Adding real-life noise would increase both the effects and the sample sizes required to detect them. Second, small effects are often investigated by scientists, from physics and genetics to rare side-effects in pharmacology.

3. The Pooling Dilemma. As mentioned above, larger studies may *prima facie* seem beneficial, both individually *and* collectively, if they are less likely to lead community-wide exploratory dynamics in the wrong direction. However, this result is obtained by looking at the impact of a single varying parameter (here, the number *n* of trials per agent per round) while keeping everything else fixed. In essence, the result means that if one could just increase everyone's resources, the community would do better overall. Though helpful in furthering our understanding of the model, this result is of limited relevance to the real world. In practice, larger sample sizes come at a cost —in terms of funding, of the time of competent experts, etc.—and often entail a decrease in the overall number of independent inquiries because of community-wide constraints in terms of available experts, data costs and scarcity (e.g., patients suffering

2 The parameters α and β determining agents' initial belief distributions about p_B are drawn at random in (0,.01], which means that agents are very responsive to evidence.

from a particular disease), or financial resources available for some area of research. This leads to a "pooling dilemma": should agents produce several small studies by themselves or a single large collaborative one? For example, should researchers in hospitals conduct their own studies with local patients, or should they try to join forces to contribute to a large multicenter trial? This alternative can be studied with respect to the interests of individuals (Boyer-Kassem and Imbert 2015) or that of communities. We focus here on the latter.

We explore this dilemma in a straightforward way: we revisit the model above while keeping the total amount of data per round, D, fixed. Under such conditions, conducting larger studies (increasing *n*) means having less of them (decreasing the number of agents *N*) because D = n·N. Success rates are then subject to two competing mechanisms. On the one hand, larger sample sizes reduce the risk of statistical noise and misleading runs of data, thus increasing success rates. On the other hand, a smaller number of agents potentially means less scientific diversity—when unlucky studies do come out, it is more likely that the *entire* community will fall for them, hence *de*creasing success rates. The question, then, is whether the increase in reliability afforded by higher-powered studies will make up for the corresponding decrease in diversity.

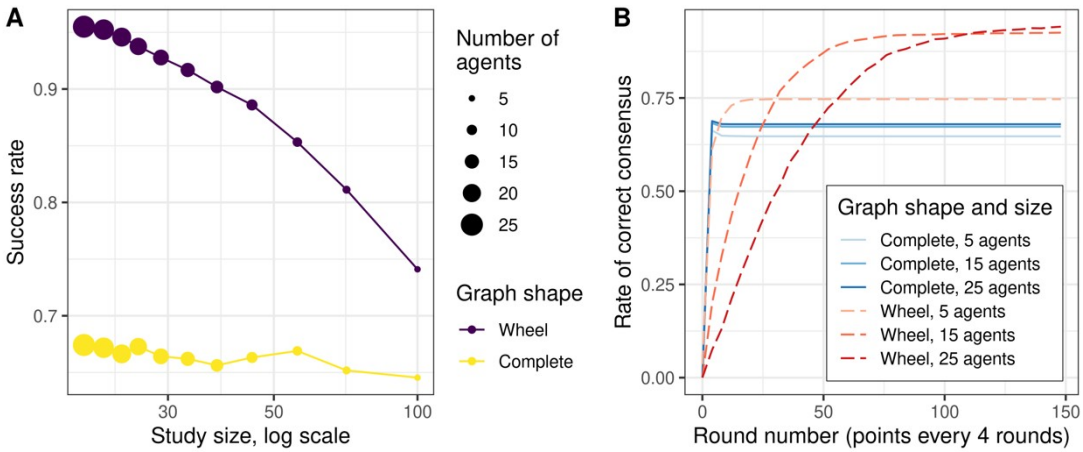

Fig. 1. Impact of sample size on community outcomes, with constant global amount of data per round (D = 500) and numbers of studies and study sizes varying accordingly. **1A:** Success rates at the end of the evolution. **1B:** Temporal evolution of correct consensus rates.

The results plotted in Fig. 1A are clear, if counterintuitive: from the point of view of community-wide research dynamics, when the global amount of data per round is kept constant, "non-pooling" communities producing more less powered studies do better than "pooling" communities producing fewer more powered studies. This is markedly true for less connected graphs; complete graphs—whose agents always take into account all data, whether packaged in smaller or larger batches—should be unaffected, yet even they display a small reliability decrease when there are only a few agents, presumably because of a loss of diversity in initial priors.

Importantly, these results merely exhibit one mechanism making large studies potentially detrimental to the community. As evidence against increasing sample sizes in the real world, these results should be taken with a grain of salt since other empirical effects may make small studies unattractive (see also section 6). For example, low-powered studies typically display larger vibration effects (i.e., results are more sensitive to methodological choices); their protocols, which are less intensively and publicly scrutinized, are often less reliable; and when conducted within NHST frameworks, they are prone to publication biases and selective reporting (Button et al. 2013a, 367–68). Finally, at a broader level, a scientific literature comprising many contradictory, small-powered, and unreliable studies is harder to make sense of, both for scientists and for decision-makers, and it may provide "extra fodder" for industrial merchants of doubt willing to cherry-pick results to defend their agendas (Weatherall, O'Connor, and Bruner 2020, 1179–80).

Furthermore, these results (like others in the BGZ-model literature) need to be refined by exploring more precisely their temporal interpretation. The reason why larger studies eventually do worse when resources are limited is that they reduce diversity: the longer diversity is preserved, the more communities accumulate data on B, and the less likely they are to be misled by unlucky data, as in (Zollman 2007; Rosenstock, Bruner, and O'Connor 2017 fig. 5). Thus, a speed/accuracy tradeoff is essentially built into the model; the real question is to characterize its precise features, e.g., by describing the timescale over which it occurs.

4. The Significance of Timescales in Investigations about Research Communities. Investigations of the BGZ model mostly focus on *asymptotic* success rates, i.e., on the proportion of simulated communities that eventually reach a correct consensus. However, time too is a limited resource: asymptotic values are meaningless in practice unless attained over realistic timeframes. The "small-study advantage" discussed above, in particular, comes at a temporal cost, whose magnitude and real-world relevance require investigation.

Unfortunately, discussions concerning what should count as a realistic timescale are conspicuously absent from the literature. Let us attempt some such first-order calibration. Scientific results are shared by publishing studies; thus, our model's "rounds" should be interpreted as publication cycles. Assuming that one publication requires somewhere between four months and three years, 100 rounds would correspond to some 30 to 300 years—incidentally raising concerns about the average convergence times described in the literature, which are usually well beyond 100 rounds (Zollman 2007, fig. 3; Rosenstock, Bruner, and O'Connor 2017, figs. 5 and 8). For the purposes of this paper, we stipulate that one round takes six months—which is on the short side—and limit ourselves to the first 150 rounds, i.e., 75 years. Furthermore, the literature usually discusses convergence times only through their averages (if at all). This is too coarse a metric for our purposes. Instead, we plot the time evolution of the proportion of simulated communities that have reached (correct) consensus (Fig. 1B).

This temporal perspective shows the sample size–diversity tradeoff in a new light: while pooling data to conduct larger studies reduces reliability in the long run, it remains beneficial for communities for some time. As an extreme case, compare 5-agent wheels (e.g., "pooling" communities in which 25 agents join forces in groups of 5 to produce, in each round, 5 studies of size 100) with 25-agent wheels ("non-pooling" communities producing 25 individual studies of size 20). Ultimately, 5-agent wheels are about 20% *less* reliable. Nevertheless, they perform better *for roughly 30 years* (60 rounds)—only then does a small-study advantage materialize (Fig. 1B). Over shorter time intervals, communities that produce larger, collaborative studies, or are more connected (the two features being related in practice, as larger studies usually reach a larger audience), are more reliable.

These results cut both ways. They show that the accuracy gains afforded by publishing small studies come at non-negligible temporal costs. However, the time scales at which large-studies-producing communities become less reliable do fall within ordinary research timeframes. Ideally, one would like "pooling" communities to perform better *over all meaningful research timescales.* We now explore how to achieve this within the model.

5. Inertia as a Way to Forestall the Pooling Dilemma. Pooling data to obtain larger samples doubtlessly produces more reliable individual studies. However, whether it is *collectively* preferable appears, at this stage, to depend on the context. Eliminating this contextual discrepancy between the individual and collective epistemic good seems well-advised. First, context-dependence makes individual agents aiming at the collective good likely to miss it for lack of reliable information about the precise context they are in. Second, to be effective, scientific norms (like "use large samples if possible") ought to be general in scope and free of complicated restricting conditions. Overall, whenever statistical power is important and small studies are associated with important defects, it would be best if data pooling could be recommended blindly. Thus, to resolve this individual-social dilemma, we now explore how to make data-pooling communities more reliable even in the short run.

In empirical inquiries, even well-prepared data may be temporarily misleading and deceive scientific communities that follow them blindly. Thus, to guard against strokes of bad luck, scientists that face unfavorable data should ignore them *up to a certain point*. Accordingly, we investigate here the effects of a simple mathematical inertia mechanism (due to (Frey and Šešelja 2020) in the context of a more complex model): when their data suddenly show that they are not investigating the most promising hypothesis, agents wait for some rounds before switching. As above, we investigate cases where the global amount of data per round is limited and explore reasonable timescales. Implications in terms of research policy are discussed in the next section.

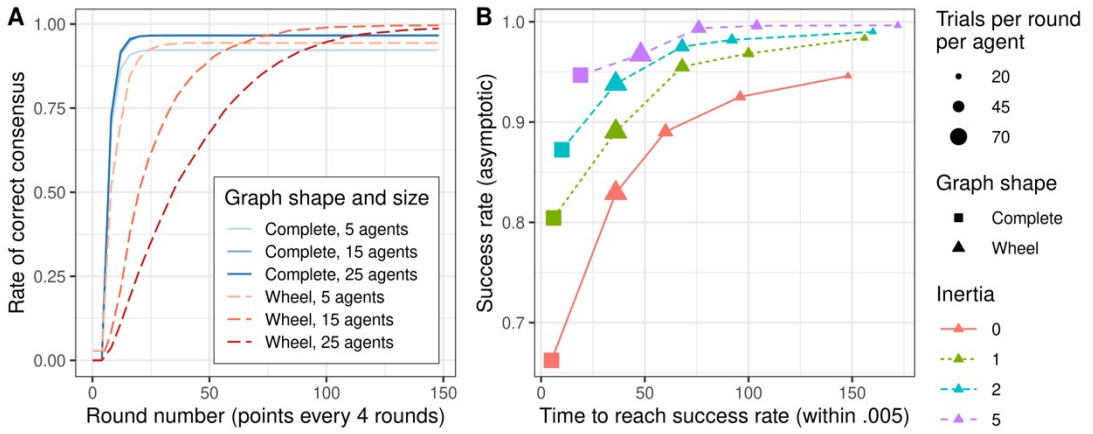

Fig. 2. Effects of inertia on the performance of communities, with constant global amount of data per round (D = 500). **2A:** Temporal evolution of the rate of correct consensus, with an inertia of 5 rounds (compare with Fig. 1B). **2B:** Scatterplot of the time-accuracy tradeoff, for wheel-shaped graphs of 7, 11, 15, and 25 agents and complete graphs of 7 agents, and inertias between 0 and 5 rounds.

As Fig. 2 shows, inertia increases long-term reliability across the board; while "non-pooling" and less connected networks still do better asymptotically, long-term reliability differences between communities become much smaller. Moreover, inertia does not significantly impact the timescale over which each community converges, usually increasing it by only a few years. Finally, and more to the point, inertia substantially increases the timeframe over which well-connected and "pooling" communities are more reliable than "non-pooling" ones (fig. 2A). Indeed, over most of the research time of a community, conducting higher-powered studies is now preferable; low-power "non-pooling" communities take the upper hand only by a small margin and over large timescales, typically above 100 rounds (50 years). Our findings are summarized in Fig. 2B: without inertia, we had to choose between fast but ultimately inaccurate "pooling communities" (bottom-left quadrant) and more accurate but slow "non-pooling" ones (top-right quadrant). An excellent compromise is reached with only five rounds of inertia (top-left quadrant). In sum, mechanisms ensuring inertia seem capable of solving the pooling problem.

6. Discussion. Insights from (idealized) models must always be taken with care (in this context, see especially (Rosenstock, Bruner, and O'Connor 2017; Frey and Šešelja 2018; 2020). We now discuss the scope and interpretation of our results.

First, this version of the BGZ model describes situations where, perhaps because of data or funding shortages, communities cannot at the same time produce high-powered studies *and* preserve scientific diversity. Though particular, such cases are by no means rare. However, when no such tradeoff exists and power matters, larger studies are unequivocally preferable, provided some error-preserving diversity is maintained.

Second, the deliberate simplicity of our model, together with our choice of realistic parameter values, suggests that the mechanism we analyze is general, and that the highlighted tension may be at work in actual communities: large collaborative studies,

though more reliable individually, may be globally detrimental when they decrease diversity significantly.

Further, there are reasons to suspect that de-idealizing the model would *magnify* the results. First, the only source of error in the model is statistical noise, whereas in practice, factors like methodological variations, biases in study design, cognitive and social biases, if not deliberate doubt manufacturing, often skew scientific processes (see, e.g., (Holman and Bruner 2017). In such contexts, large studies, because they reduce diversity by drying up the research pool, may be especially misleading to communities whenever they happen to be affected by such problems. Second, our model's Bayesian agents are careful not to give studies more weight than their sample size warrants. However, real-world scientists may give disproportionate weight to both small and large studies, thus increasing the potential for incorrect convergence. Similarly, it would be interesting to analyze whether statistical filters for publication selection, such as the NHST paradigm, create additional worries, as investigated in (Romero and Sprenger 2021) for meta-analyses.

Finally, one may wonder whether, when collaboration reduces the number of teams working on a problem, scientific diversity is really threatened. Is it not the case that countervailing real-world mechanisms—virtuous or otherwise—already shield scientific diversity? Perhaps. For instance, the cost of changing one's research program, the effects of conservative cognitive biases, the constraints coming with scientific grants or private funding, or even sheer stubbornness,[3] might mimic the effects of inertia and push scientists to ignore empirical data, at least temporarily. However, assuming that such mechanisms play a diversity-preserving role in just the right way would be overoptimistic, especially since, as is well-known, powerful forces also promote conformity. Counting on mechanisms that push scientists to ignore data amounts to playing with fire. It seems preferable to develop scientific norms that promote diversity explicitly by controlled and transparent means.

Space is missing to discuss in depth how inertia-promoting mechanisms could be institutionalized. However, here are suggestions. At the editorial and reviewing level, methodologically sound studies should not be disqualified based on recent evidence contradicting them, especially when large collaborative teams structure communities and dissenting investigations tend to be rare. Similarly, studies investigating seemingly disconfirmed but insufficiently explored hypotheses should be encouraged—provided, naturally, they state all available evidence and do not shun scientific method. While guaranteeing inertia may seem complicated in practice, note that very little of it already produces significant effects. Exactly how much inertia would be needed in practice depends on how long typical errors actually persist—on this, more work is required.

7. Conclusion. The recent literature on sample sizes depicts low-powered studies as an endemic problem that may be solved—among other things—by working

3 Zollman (2007) also investigates the effects of extreme agents, i.e., agents having extreme initial priors.

collaboratively and sharing data to develop higher-powered, more reliable studies (see, e.g., Button et al. 2013a, 374). However, individual studies are by no means the *terminus ad quem* of science; they are fallible intermediate instruments within broader scientific processes, and they should be judged, not just individually, but according to their broader impact on the course of science. Our results illustrate that discrepancies may arise between these two levels, just like there can be a mismatch between what is rational individually and collectively in scientific communities (Kitcher 1990). Indeed, while one might expect scientific communities to be epistemically better off when individual studies are high-powered and reliable, and worse off when they are small, we have shown that, counterintuitively, this need not always be true. We studied the impact of sample sizes on *collective exploration dynamics* under conditions of *resource limitation*—a context in which researchers use published studies not to determine conclusively whether a hypothesis is true, but to decide whether it is worth investigating further. In such cases, setting up a few higher-powered studies instead of many smaller ones may be detrimental: when large studies happen—for some reason or other—to be misleading, they are more likely to lead the *entire* community astray.

Our argument should not be misconstrued as an apology for low-powered studies, which have many drawbacks. Rather, it highlights the need to be mindful of the indirect epistemic costs that large studies can occasion through diversity loss, and the importance of mitigating these through thoughtful institutional practices to preserve both individual and collective reliability. Overall, our results support the idea that the question of sample sizes admits of no unequivocal answer; overly general, context-independent norms may be harmful. Which sample sizes are more truth-conducive for scientific communities seemingly depends on the amount of data available, the community structure, and the timeframe under consideration. Moreover, studies can serve different purposes at different stages of an investigation, and different sample sizes may be appropriate in each case. Science is a remarkably complex collective enterprise. We would do well to take heed of this complexity before issuing one-size-fits-all methodological decrees on sample sizes.

8. References

Bacchetti, Peter. 2013. “Small Sample Size Is Not the Real Problem.” *Nature Reviews Neuroscience* 14 (8): 585–585.

Boyer-Kassem, Thomas, and Cyrille Imbert. 2015. “Scientific Collaboration: Do Two Heads Need to Be More than Twice Better than One?” *Philosophy of Science* 82 (4): 667–88.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013a. “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience.” *Nature Reviews Neuroscience* 14 (5): 365–76.

———. 2013b. “Empirical Evidence for Low Reproducibility Indicates Low Pre-

Study Odds.” *Nature Reviews Neuroscience* 14 (12): 877–877.

Cohen, Jacob. 1962. “The Statistical Power of Abnormal-Social Psychological Research: A Review.” *The Journal of Abnormal and Social Psychology* 65 (3): 145–53. https://doi.org/10.1037/h0045186.

Frey, Daniel, and Dunja Šešelja. 2018. “What Is the Epistemic Function of Highly Idealized Agent-Based Models of Scientific Inquiry?” *Philosophy of the Social Sciences* 48 (4): 407–33.

———. 2020. “Robustness and Idealizations in Agent-Based Models of Scientific Interaction.” *The British Journal for the Philosophy of Science* 71 (4): 1411–37.

Fries, Pascal, and Eric Maris. 2021. “What to Do If N Is Two?” *ArXiv:2106.14562 [Stat]*, June. http://arxiv.org/abs/2106.14562.

Higginson, Andrew D., and Marcus R. Munafò. 2016. “Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions.” *PLOS Biology* 14 (11): e2000995.

Holman, Bennett, and Justin Bruner. 2017. “Experimentation by Industrial Selection.” *Philosophy of Science* 84 (5): 1008–19.

Kitcher, Philip. 1990. “The Division of Cognitive Labor.” *The Journal of Philosophy* 87 (1): 5–22.

Laudan, Larry. 1981. “Peirce and the Trivialization of the Self-Corrective Thesis.” In *Science and Hypothesis*, 19:226–51. The University of Western Ontario.

Romero, Felipe. 2016. “Can the Behavioral Sciences Self-Correct? A Social Epistemic Study.” *Studies in History and Philosophy of Science Part A* 60 (December): 55–69.

Romero, Felipe, and Jan Sprenger. 2021. “Scientific Self-Correction: The Bayesian Way.” *Synthese* 198 (23): 5803–23.

Rosenstock, Sarita, Justin Bruner, and Cailin O’Connor. 2017. “In Epistemic Networks, Is Less Really More?” *Philosophy of Science* 84 (2): 234–52.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. “Bayesian t Tests for Accepting and Rejecting the Null Hypothesis.” *Psychonomic Bulletin & Review* 16 (2): 225–37.

Sedlmeier, Peter, and Gerd Gigerenzer. 1989. “Do Studies of Statistical Power Have an Effect on the Power of Studies?” *Psychological Bulletin* 105 (2): 309–16.

Smaldino, Paul E., and Richard McElreath. 2016. “The Natural Selection of Bad Science.” *Royal Society Open Science* 3 (9): 160384.

Smith, Philip L., and Daniel R. Little. 2018. “Small Is Beautiful: In Defense of the Small-N Design.” *Psychonomic Bulletin & Review* 25 (6): 2083–2101.

Trafimow, David, and Michael Marks. 2015. “Editorial.” *Basic and Applied Social*

Psychology 37 (1): 1–2.

Vankov, Ivan, Jeffrey Bowers, and Marcus R. Munafò. 2014. “On the Persistence of Low Power in Psychological Science.” *Quarterly Journal of Experimental Psychology (2006)* 67 (5): 1037–40.

Weatherall, James Owen, Cailin O’Connor, and Justin P. Bruner. 2020. “How to Beat Science and Influence People: Policymakers and Propaganda in Epistemic Networks.” *The British Journal for the Philosophy of Science* 71 (4): 1157–86.

Zollman, Kevin J. S. 2007. “The Communication Structure of Epistemic Communities.” *Philosophy of Science* 74 (5): 574–87.

———. 2010. “The Epistemic Benefit of Transient Diversity.” *Erkenntnis*, 2010, sec. 72 (1).