The New England Journal of Statistics in Data Science Volume 1, 39–45 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS6REJ

A Not-so-radical Rejoinder: Habituate Systems Thinking and Data (Science) Confession for Quality Enhancement[☆]

Xiao-Li MENG

1. PRINCIPLED SYSTEMS THINKING

Among the four "radical" proposals I ventured in my article, the last one—promoting kidstograms—is endorsed universally by the discussants, with the first one—doubling the variance—receiving the most criticisms. I value both kinds of feedback, because they enrich my understanding of the challenges and opportunities our profession faces, and I am deeply grateful to all discussants for their inspiring and candid comments. Clearly we all recognize the positive impact of an earlier or the earliest statistical education, from producing maturer and wiser citizens to building adroiter and astuter leaders for the digital age. It is particularly heartwarming to receive strong endorsements from Christine Franklin, an eminent leader in K-12 statistical education, and from Erik Kolaczyk, a pioneer in reforming statistical practicums at the graduate level. Franklin's thought-provoking list of questions is also action-prompting, but to address them fully would take generations of effort, because they will require *institutional change*, as emphasized by Wasserstein, Schirm, and Lazar (WSL). However, this is the very reason that we need to act as urgently as possible, especially for the aims of broad data science education, while the data science is still young but rapidly evolving, as Kolaczyk reminds all of us. Developing a native appreciation of variations is the first essential step in cultivating life-long habits of statistical thinking, just as learning a dialect in a native environment has lasting impact on one's everlasting fluency in that language.

Kolaczyk's mantra that "Theory informs principle; principle informs practice" constructs the missing bridge and archway for the usual emphasis on training students with both theoretical insights and practical skills. My proposal on principled corner cutting (PC2) necessarily predicates on a good appreciation of the required principles, without which insights and skills do not automatically translate into statistical adroitness, astuteness or data acumen [12]. Whereas insights and skills can be developed well asynchronously, adroitness, astuteness and acumen all require principled systems thinking [8, 28], which is critical for making sensible trade-offs or more generally for handling wicked problems [5]. Systems thinking refers to contemplation—factually or counter-factually—and appreciation for how components of a complex system (problem) and their individual considerations (solutions) may interact with each other, and the overall consequences of these interactions on the internal stability and external impact of the system (the overall solution). Principled systems thinking invokes domain principles to guild this contemplation, investigates the incompatibility among such principles, and ultimately establishes overarching principles for sustaining a system or solution.

Virtually all systems and problems involving humans are complex, perhaps a sine qua non consequence of humans' emotionally-infused intelligence (which defeats machines' imitation so far and perhaps forever [15]). The increasingly debated issue of concomitantly preserving data privacy and utility is just one example of many. While it illustrates the criticality of principled systems thinking for the broader data science, allow me to briefly digress before invoking statistical systems thinking in responding to various criticisms from the discussants. We humans are insatiable consumers of information and data collected from fellow homo sapiens, whether we admit or realize it or not. Yet we rarely would be willing to supply the same for others' benefit, and for good reasons; personal information in the wrong hands can do much harm. This differential attitude towards the information demand and supply can be only more salient as the world becomes increasingly digital. However, it is impossible to retain both information and privacy to our desired levels respectively as a consumer and supplier, and indeed this impossibility can be proved mathematically [14]. Compromises have to be made.

But how, and by whom? Statisticians are trained to extract as much information as possible from data, whereas computer scientists have pioneered many methods for data security and privacy, perhaps the most widely known of which is differential privacy [9, 10]. A key quantity in implementing differential privacy is ϵ, known as the (log of) privacy loss budget, governing the amount of noise to be injected into the data for privacy protection. But how large or small ϵ should be is not a question that can be handled by statistical or computer science principles alone or even together. As an example, what is the appropriate value ϵ for the United States census is a question for collective contemplation by its users and stakeholders. Principled systems

[☆] Main article: https://doi.org/10.51387/22-NEJSDS6.

thinking, for which statistical systems thinking is an integral part, is indispensible to make headway for such complex, large, and on-going issues; see the special issue in *Harvard Data Science Review* on differential privacy for the 2020 US census, introduced by a thought-provoking editorial [11].

2. PRACTICING STATISTICAL SYSTEMS THINKING ...

In general, without principled systems thinking, the more wicked the problem, the more likely we succumb to strategies producing only Pyrrhic victories or instant gratifications. Wicked problems are wicked because they don't have right or wrong answers, but only better or worse answers, which themselves change over time. Worse, whenever a seemingly satisfactory solution is found to address a part of a challenge, it creates new problems for some other parts. Controlling a pandemic by strictly enforcing quarantine may be a better strategy than encouraging self-quarantine during a pandemic peak, but it can unduly create many hardships and damages otherwise. Indeed damages can be most severe especially when the strict enforcement is perceived as the best overall strategy, because then its enforcement inhibits concerns with other hardships.

My contemplation of variance-doubling proposal as a way to control false positive is nothing comparable to that for controlling a pandemic. Nevertheless, controlling pandemic reminded me of the need of systems thinking for dealing with one of the wickedest problems for academic researchers: control the misuse and abuse of a popular method. A consumer product can become popular for a host of reasons, but the most common ones appear to be that it meets a frequent demand, and that it is economical. The popularity of a statistical method follows a similar pattern, where being economical is with respect to users' time and mental investments. All statistical methods come with assumptions, explicit, subtle, or even deliberately hidden. However, the most critical but almost never stated assumption is that it will be used in "the right hands". To define precisely "the right hands" may require more hands than we can handle for covering a series of rabbit holes. But it should be clear that in general a method cannot be labelled as being popular if it is used only by "the right hands". Therefore, it is expected that the users of a popular method come with a diversity in competency, and some misuses are inevitable.

However, when misuses and abuses become pervasive, despite all the efforts to reduce them, it signals an underlying incentivizing cause. For cases where variance-doubling can make a difference, I believe it is the substantial asymmetry in our reward systems, which incentivizes false positive far more frequently than false negative [21]. For example, identifying a disease-causing gene typically takes substantial effort, time, and resources, and often there is a race among different teams for being the original discoverer; indeed, for genetics studies in general, the race could be so intense that the phrase "gene wars" has been used [6]. Hence a research team would be more likely to risk a premature announcement than losing the priority by conducting a thorough investigation. If the finding ends up being a false positive, well, it is just one of many and we could, as we usually do, all excuse false empirical studies since they never come with guarantees. Other than obvious disappointments and possible reputation reduction if a team consistently produces false positive results, few punitive measures exist. However, if the announcement turns out to be correct, then the rewards can be bountiful: fame, funding, followers, etc. Why wouldn't an investor prefer a stock with little chance of going down?

Considering this overall need of greatly reducing false positive in practice by investigators from all walks of scientific inquires and with all levels of statistical competences, I believed that the simple adjustment by doubling one's variance estimates has a better chance to be adopted and make a net-positive overall difference than more sophisticated and tailored strategies (which of course should be adopted whenever available, such as those derived from the elegant mathematical bounds discussed by Dennis Lin). Whereas any sweeping assertion has the danger of being naive or counterproductive, the practical equivalence between doubling the variance and the proposal of raising the standard for statistical significance to $\alpha = 0.005$ by [1] provides some empirical evidence for such a belief. As [1] reported, in two empirical studies, changing $\alpha = 0.05$ to $\alpha = 0.005$ resulted in approximately doubling the rates for replicating reported significant experimental results. (Convincing scientific evidences of course require more than $n = 2$; these studies merely raise the awareness of the sizable potential of variance-doubling.)

3. ... BUT I NEED MORE PRACTICE

However, reading the criticisms and concerns from discussants Thomas Junk, Dennis Lin, and WSL, made it clearer that my overall arguments for doubling the variance are unconvincing (or the discussants disagree with the premise that false positive occurs far more frequently in practice than false negative). For specific applications, I agree with all the cautions regarding its complications and negative consequences, except for the concern that variance doubling can lead to implausible confidence intervals, because such occurrences should serve as a reminder of the inappropriate constructions of the confidence intervals in the first place.[1] For systems thinking, the discussants' criticisms reveal several issues that I overlooked or did not think through. I'm

[1] For example, a confidence interval for voting percentage p can go outside the unit interval because of the common but inappropriate practice of applying a normal approximation directly to $\hat{p}$ when one should use a normal approximation to its logit transformation $\hat{\lambda} = \log(\hat{p}/(1 - \hat{p}))$, which typically yields a more accurate normal approximation. Doubling the variance for $\hat{\lambda}$ will not lead to implausible values for p, since any confidence interval for p obtained by converting a confidence interval for λ is mathematically constrained to be in $[0, 1]$.

therefore grateful to the opportunity to rethink, though by no means that I'd get everything right this time or ever, as my judgments are necessarily limited by my experiences and knowledge.

First, I did not explicate that "double your variance" should really be "double your variance estimate" or more precisely "double your uncertainty assessment." There is of course no reason to double a theoretical variance when it captures the correct uncertainty assessment, even just approximately. The doubling strategy is designed to combat the common tendency of under-assessing uncertainty, which occurs for a host of reasons. As I emphasized in the article, doubling the variance is about increasing the chance of quality control and keeping our promises. In most if not all applications, uncertainty assessments necessarily involve statistical asymptotics (e.g., via a large-sample approximation), estimation errors (e.g., via bootstrap), or numerical approximations (e.g., by a Monte Carlo integration), not to mention all other sources of uncertainties that are not captured by the variance (e.g., selection bias in the data collections, and model uncertainty). To increase the chance that our final uncertainty estimate delivered in any particular study is at least what it should be (e.g., leading to the appropriate confidence coverage as declared), doubling the variance is in the right direction but it may still be far from rendering the needed correction. For example, several recent studies show the effective sample sizes of some "big data" are in the range of 0.02% to 0.1% of the reported data sizes [23, 4, 3], which implies that the corrective multipliers range from $\sqrt{1/0.001} \approx 32$ to $\sqrt{1/0.0002} \approx 71$, far exceeding the proposed 2. Therefore doubling the variance should not be equated as being conservative. Nevertheless, WSL's several questions on losing power, and proper balancing Type-I and Type-II errors trade-offs certainly should be asked whenever we prioritize on controlling false positive, regardless whether or not doubling the variance estimate is involved.

Second, my premise that there are far more false positive findings than false negative ones focuses on comparing frequencies of their occurrences, not their consequences. This narrow focus weakens my arguments, even if my judgment is correct. This weakness is captured by WSL's tough question about the statistical and ethical principles to guide the considerations of consequences of false negatives. It will be the case that for some sub-systems (e.g., studies of a particular type of diseases), the false negative is more consequential than false positive, for which of course doubling the variance can do harm. However, in such cases, we should reconsider our conventional strategy—and the corresponding promise—of first controlling false positive (Type I error), and then minimizing false negative (Type II error). The matter is again a wicked one when we seek general guidelines, such as on setting α. On the one hand, we need to be adaptive, or "fit for purposes", as WSL noted. However, the practice of adaption could be abused into "α-hacking" (which of course has already happened even with only two common choices $\alpha = 0.01$ and $\alpha = 0.05$). For those who do not intentionally engage in α-hacking, the general advice is to be as transparent as possible, always reporting the results as well as the (thinking) process of arriving at them. Such guidelines, however, will have little effect on those who knowingly engage in p-hacking or α-hacking. A further systems thinking would require us to contemplate if the increased false negative rates can collectively do more damage to science and society than maintaining the current levels of false positive rates. These are dizzily difficult contemplations, starting from conceptualizing sagacious metrics for such a comparison. Nevertheless, given the practical equivalence between doubling the variance and changing the significance standard to $\alpha = 0.005$, all the responses to potential objections given in [1] are essentially applicable here.

Third, physicist Thomas Junk's point that "In many fields of study, it is impossible to determine in advance whether a result will be the 'last' one in a chain of reasoning, or if it will be used as input to another study" highlights further the kind of systems thinking that we statistician should engage ourselves in more. I certainly failed on this point until I was reminded by Junk's discussion, which I highly recommend readers to dive in because it brims with food for thought. How could I have missed this seemingly obvious point: any output of a study can be used as an input for a future study? This oversight is particularly ironic for me since I have been promoting the idea of multi-phase inference, which explicitly recognizes that the output of an earlier phase is an input for the next [19, 2, 22]. Could this be a simple glitch of an aging mind, or something that requires deeper troubleshooting?

4. TAKING A LEAD IN DATA (SCIENCE) CONFESSION

Retrospectively, I believe my contemplation of the pros and cons of the variance-doubling strategy was preconditioned by two subconscious streams, one pardonable but the other more troublesome. As primarily a methodological and theoretical (but not mathematical) statistician, my "field of study" focuses significantly more on methodological breadth and theoretical depth than "a chain of reasoning" as scientists would engage themselves in for substantive investigations, especially those of large-scale and fundamental nature. As such, when I was emphasizing *final* confidence interval or p-value in Section 2.4 of the article, I was clearly not habituated to consider serial studies for addressing an overarching scientific goal or a large societal need, such as searching for fundamental particles or reducing global poverty, for which the notion of final result/solution would be considered hopelessly naive, even though we all hope for them.

This difference in emphases and mind framing itself is expected from the perspective of disciplinary division of labors, and arguably it is even healthy. As WSL emphasized, genuine replications of a single study is rare in practice (even if we do not insist on the necessary differences, such as temporal and idiosyncratic variations). But to answer WSL's question on the meaning of confidence coverage or error rate for any particular study, by observing quality delivery of a procedure for a variety of studies, we gain confidence in its performance in each single one, but with the awareness that our confidence can be misplaced in any single instance. Inevitably such confidence transference requires some degree of *leap of faith*, which turns out to be a necessary part of any statistical inferential paradigm, as discussed in [7]. This notion of transferring confidence from a collection to a single instance—known as *transitional inference*—was used by philosophers as early as Galen of Roman empire, as discussed in [17], in the context of accumulating evidence for assessing the effectiveness of individualized treatments. Indeed this is how we humans make judgments in our daily lives, where we can entertain genuine replications such as *Groundhog Day* (1993) only in Hollywood style. This evaluative perspective also applies to Bayesian procedures, as their operating characteristics can be evaluated pragmatically only by assessing their performances over many different studies [29]. Nevertheless, as much as this multitudinous perspective is necessary for a tool-providing discipline like statistics, it apparently created a blind spot in my contemplation of the applicability of the variance-doubling strategy, treating each study in isolation and forgetting the sequential nature of many studies in human inquires.

This blind spot was enlarged by a second preconditioning, which is more pervasive, because it is a form of confirmation bias. Once I convinced myself that variance-doubling is generally beneficial, my contemplation moved from assessing to communicating its advantages. Whereas this is a common practice, it also comes with a common pitfall – our minds are preconditioned then to ignore inconvenient truths, or at least to avoid seeking them. The notion of doubling only the *final* variance was a critical one in convincing others (and myself), and hence anything challenging it would become rather inconvenient. Of course, once an inconvenience is identified, then any responsible researchers will address it. But it is a different matter whether one actively seeks such inconvenience, as a way to stress-test one's idea or methods. I often initiate such stress-tests when I argue with others, but I clearly had failed to argue with myself in this instance.

I am therefore grateful to Junk for reminding me of my oversight, and with examples. Junk is clearly right that doubling the variance does not guarantee the right answer, because the resulting estimate can be too small or too large in itself. Furthermore, when it is applied to an already (approximately) correct assessment, then it obviously will mislead the subsequent studies. The more interesting cases are those common ones where we have under-assessed an uncertainty for a variety of reasons, and hence variance doubling is in the right direction for *that* assessment. But does this imply that it is also in the right direction for any subsequent analysis that relies on this assessment? Would the subsequent analysis err in the same direction when the variance doubling under-corrects or over-corrects? As the example in [13] indicates, here our intuition can easily mislead us if we do not stress-test ourselves, because variation propagation under non-linear relationships defies simple summative rules, even approximately.

Junk's caution therefore is spot on. Nevertheless, it is not necessarily a call to ignore the variance-doubling adjustment, but rather a call for clear documentation of the procedure applied, a practice should be followed in any scientific study. I accept Junk's criticism that my article didn't provide a well-defined line between applying and not applying the variance-doubling strategy when the variance is not "final". I must confess that I still don't—or will ever—have one. However, after much contemplation and considering the wickedness of the problem as previously discussed, I'd still recommend doubling one's variance estimate unless one has little doubt that it does not under-assess, and then explicitly stating the use of the doubling strategy. Even if subsequent analysts are unable to undo this doubling adjustment, they will be properly informed of the uncertainties in the previous uncertainty assessments, since otherwise there would be no need for invoking any adjustment. This awareness can provide some resistance to our general tendency of rushing to conclusions, whether being preconditioned or incentivized.

Recently, I coined the term "data confession" [26] to encourage more disclosures in research publications about defects in data conceptualization, collection or pre-processing, as another component in enhancing the replicability and ultimately the reliability of published scientific studies, since data quality matters far more than data quantity [23, 4, 27, 3]. The retrospective introspection summarized above suggests a more general *data science confession* (DSC) in our publications, where we can benefit from each others' mistakes and lessons learned, especially how we reason with ourselves, where we can engage in a pure intellectual dialogue without being distracted by suspicions of impure motivations. I therefore appreciate the editors' patience and flexibility in giving me sufficient time (and space) to go through and document this retrospective introspection. I certainly hope other journals will follow the lead of *NEJSDS* in encouraging, or at least permitting, such data science confessions. I use the term DSC instead of *statistical confession* to avoid suggesting that statistics is the same as data science, which is a much broader ecosystem [24]; indeed, the same consideration led to the journal title *NEJSDS* with SDS abbreviating "Statistics *in* Data Science" instead of "Statistics *and* Data Science."[2] Rather, I see even more a need for self-disclosing defects, mistakes, and oversights in the broader data science community, and statisticians, being a vital part of this community, can and should lead.

[2] I was given the honor to serve as the founding president of NESS, and hence the opportunity to suggest its journal title.

5. IT IS ALL ABOUT "QUALITY AT EVERY STEP"

Much of this rejoinder has been devoted to the variance-doubling issue because the discussants' criticisms has encouraged me to explore further the need of systems thinking [25] and of data confession [26]. However, both notions' raison d'être is quality enhancement in everything we do. This emphasis on quality is shared by all discussants, from Franklin's call on enhancing training on problem solving to include quality control, to Lin's emphasis on moving from "validity of inference" to "quality of inference", and to Junk's stress to improve the quality of statistical analyses rather than merely adjusting the labeling after the fact. Indeed, Lin reminded us that the quality control should start earlier, that is, at the data collection stage, "addressing the problem before analysis even began." Although this was not a topic for my panel presentation in 2017 and hence it was not included in the article, it has been a central topic of my research since [22], which led to the proposal of "data minding" [26]. The main finding, as reported in [23], supports Lin's emphasis to its core—data quality matters far more than data quantity. In particular, if we fail to take into account the data quality, then we may become victims of the big data paradox: the bigger the data, the surer we fool ourselves because of the misleadingly narrower confidence interval centered at a biased estimator. Such lessons and pertaining principles should and can be taught to precollege students without much mathematics or technical jargon beyond the most basic statistical terms such as mean, variance, and correlations, but with engaging and effectual case studies, e.g., the vaccination surveys investigated in [4]. This will help to enhance our teaching on data quality, as Junk stressed. It can also help to improve the quality of our data science teaching, because it reflects Kolaczyk's mantra on bridging theory and practice with principles.

Although WSL's ATOMIC is not a bombshell among this group of discussions, its composing principles summarized well the essences of my proposals, collectively emphasizing "quality at every step" to enhance replicability and reliability of scientific studies. WSL's accentuation on "quality requires transparency" underscores precisely the aforementioned confessions. Furthermore, I agree with Junk that a well-kept research dairy is important, not just for transparency and keeping others informed, but also for reducing our research mistakes or inefficiency. I'm also very pleased to see Junk's and WSL's strong endorsement of the proposed quality introspection list. I wholeheartedly agree with Junk that the list should be considered before starting a study (and regret that I failed to emphasize this point in the first place), not merely after, because by then it could just be a postmortem checklist, to paraphrase R. A. Fisher's famous quote on consulting statisticians after an experiment.[3]

Ultimately, *quality at every step* requires institutional changes (IC), and hence I'm delighted to see that WSL has enhanced their ATOM to ATOMIC. Change is hard, and institutional change is particularly hard, even just in our contemplation. My salary-reduction *imaginary*—to borrow a (philosophical) term from [16]—evidently has generated some uneasy feelings. These are expected, because the imaginary was designed to be a provocative thought experiment to push us to think harder about more realistic incentives for autonomous quality controls. Nevertheless, I fully accept Lin's criticism that my proposal would be far more forceful if it came with practical schemes that "can incentivize self-quality controls in intuitions." Here is another confession: I did try, but to paraphrase the concluding line of Peter McCullagh's discussion[4] of [18], the parts of my ideas that are practical are not new, and the parts that are new are not practical. Indeed, there are also a number of questions, especially those posted by WSL, for which I do not have anything intelligent to add to the existent literature.

6. WHAT'S RADICAL TODAY MAY BE OPTIMALLY RATIONAL TOMORROW

Of course, it would be rather depressing to end my rejoinder with merely confessions. I'm therefore grateful to Lin for questioning the absence of any discussion in my article about the value of the peer review process for ensuring quality of research. This is because improving peer review processes is a great example of how institutional changes can be achieved by collective and persistent effort,[5] considering the significantly reduced review time compared to when I started my statistical career—I was one of many who were frightened by review processes that could take more than a year to provide merely the initial review [20]. But more excitingly, Lin's question reminded me of a recent proposal by an innovative young scholar, who used statistical insights to design an apparently radical scheme to address an unprecedented challenge and threat to review quality created by the rapid evolution of the data science itself, providing an inspiring demonstration of how future generations can and will necessarily do better than their predecessors.

The unprecedented challenge is vividly highlighted by the fact that NeurIPS 2020, a machine learning conference, received nearly 9500 submissions. An apparent reason for such staggering numbers is that there are increasingly more prolific researchers or research teams in data science. For example, it was reported that for ICLR 2020, there were 133 authors whose names appeared on at least five submissions,

[3] "To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination. He can perhaps say what the experiment died of." (Fisher, First Session of the Indian Statistical Conference, Calcutta, 1938).

[4] See page 35 of [18], where McCullagh wrote, "In the discussion of foundational matters, however, the parts of the paper that are true are not new, and parts that are new are not true."

[5] See, for example, the panel discussion summarized in [30] and the special issue of IMS Bulletin on improving the review process, available at https://imstat.org/wp-content/uploads/Bulletin37_2.pdf.

with the highest number being thirty two. How can we ensure review quality when the number of available qualified reviewers is much smaller than the number of submissions?[6]

A recent proposal by Weijie Su [31] comes with an apparently radical title "You are the best reviewer of your own papers ...". However, the theoretical results justifying the proposal [32] establish that when authors submit multiple articles, it is in the authors' interest to report their honest ranking of their submissions in order to maximize their expected utility under the proposed owner-assisted scoring system. This scheme optimally projects the external reviewers' scores to the space that obeys authors' self-ranking, and such projected scores are guaranteed to be more accurate estimates of the underlying true scores than the external reviewers' scores. In other words, an apparently extreme radical idea by conventional wisdom regarding peer review (how could one involve authors in reviewing their own submissions?) provides an optimally rational strategy to address new challenges. Whereas the optimality results are (necessarily) established under several assumptions, all of which can be questioned, it is a great demonstration of how challenges can yield innovation, and how previous unthinkable or unacceptable mechanisms can become optimal strategies with fresh thinking.

Making institutional or system changes is often a frustrating experience, and dealing with wicked problems can be downright depressing. Yet such changes are necessary for every generation. It is therefore fitting to conclude this rejoinder by showcasing that what's radical today may be optimally rational tomorrow. Indeed, *NEJSDS*'s two-track review system permitting an author-led track, building on another young-scholar driven radical experiment (https://researchers.one), indicates that tomorrow may come sooner than we realize.

Accepted 7 September 2022

REFERENCES

[1] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. Redefine statistical significance. *Nature Human Behaviour* **2**. 6–10 (2018).

[2] Blocker, A. W. and Meng, X.-L. The potential and perils of preprocessing: Building new foundations. *Bernoulli* **19**(4) 1176–1211 (2013). https://doi.org/10.3150/13-BEJSP16. MR3102548

[3] Boyd, R. J., Powney, G. D. and Pescott, O. L. We need to talk about nonprobability samples (2022). arXiv preprint. arXiv:2210.07298.

[4] Bradley, Valerie C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. and Flaxman, S. Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* **600**(7890) 695–700 (2021).

[5] Conklin, J. *Dialogue mapping: Building shared understanding of wicked problems*. John Wiley & Sons, Inc., (2005).

[6] Cook-Deegan, R. M. *The gene wars: Science, politics, and the human genome*. WW Norton & Company, (1994).

[7] Craiu, R. V., Gong, R. and Meng, X.-L. Six statistical senses. *Annual Review of Statistics and Its Application*. in press. https://doi.org/10.1146/annurevstatistics-040220-015348.

[8] Dahleh, M. A. Data Science in Domains: Interaction of Physical, Social, and Institutional Systems. *Harvard Data Science Review* **3**(2) (2021). https://hdsr.mitpress.mit.edu/pub/nv9xq9yu.

[9] Dwork, C. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation* 1–19. Springer, (2008). https://doi.org/10.1007/978-3-540-79228-4_1. MR2472670.

[10] Dwork, C. and Smith, A. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* **1**(2) 135–154 (2009). https://doi.org/10.1007/978-3-540-79228-4_1. MR2472670.

[11] Gong, R., Groshen, E. L. and Vadhan, S. Harnessing the Known Unknowns: Differential Privacy and the 2020 Census. *Harvard Data Science Review*, (Special Issue 2), (2022). https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

[12] Haas, L., Hero, A. and Lue, R. A. Highlights of the National Academies Report on "Undergraduate Data Science: Opportunities and Options? *Harvard Data Science Review* **1**(1) (2019). https://doi.org/10.1162/99608f92.38f16b68.

[13] Junk, T. R. and Lyons, L. Reproducibility and replication of experimental particle physics results. *Harvard Data Science Review* **2**(4) (2020). https://doi.org/10.1162/99608f92.250f995b.

[14] Kifer, D. and Machanavajjhala, A. No free lunch in data privacy. In *Proceedings of the 2011 International Conference on Management of Data – SIGMOD '11*, Athens, Greece 193–204 (2011). ACM Press. https://doi.org/10.1145/1989323.1989345.

[15] Larson, E. J. *The Myth of Artificial Intelligence*. Harvard University Press, (2021).

[16] Leonelli, S. Data Science in Times of Pan(dem)ic. *Harvard Data Science Review* **3**(1) (2021). https://hdsr.mitpress.mit.edu/pub/fi1rol2i.

[17] Li, X. and Meng, X.-L. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association* **116**(533) 353–367 (2021). https://doi.org/10.1080/01621459.2020.1844210. MR4227699.

[18] Lindsey, J. K. Some statistical heresies (with discussion). *Journal of the Royal Statistical Society: Series D (The Statistician)* **48**(1) 1–40 (1999).

[19] Meng, X.-L. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **9**(4) 538–558 (1994). http://www.jstor.org/stable/10.2307/2246252.

[20] Meng, X.-L. Double effort, not double blind! Technical Report 382, Department of Statistics, University of Chicago (1994).

[21] Meng, X.-L. Desired and feared — what do we do now and over the next 50 years? *The American Statistician* **63**(3) 202–210 (2009). https://doi.org/10.1198/tast.2009.09045. MR2750343.

[22] Meng, X.-L. A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science* (Eds: Lin et. al.) CRC Press, (2014). MR2049935.

[23] Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics* **12**(2) 685–726 (2018). https://doi.org/10.1214/18-AOAS1161SF. MR3834282.

[24] Meng, X.-L. Data science: An artificial ecosystem. *Harvard Data Science Review* **1**(1), 7 (2019). https://doi.org/10.1162/99608f92.ba20f892.

[25] Meng, X.-L. Building Data Science Infrastructures and Infrastructural Data Science. *Harvard Data Science Review* **3**(2) (2021). https://hdsr.mitpress.mit.edu/pub/kdqoo5ax.

[26] Meng, X.-L. Enhancing (publications on) data quality: Deeper

[6] The even more pressing question is how can we ensure the quality of research in a culture that inevitably incentivizes quantity more than quality? Because this rejoinder is already at the risk of failing to justify its length by its content, I will have to leave this discussion to another rejoinder.

data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184**(4) 1161–1175 (2021). https://doi.org/10.1111/rssa.12762.

[27] Msaouel, P. The big data paradox in clinical practice. *Cancer Investigation* **40**(7) 567–576 (2022). https://doi.org/10.1080/07357907.2022.2084621.

[28] Qin, S. J. Data Science Education With Domain Knowledge and System Principles. *Harvard Data Science Review* **3**(2) (2021). https://hdsr.mitpress.mit.edu/pub/8si074w9.

[29] Rubin, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12** 1151–1172 (1984). https://doi.org/10.1214/aos/1176346785. MR0760681.

[30] Stein, M. and Meng, X.-L. Report on 1995 IMS-ASA invited panel on Speeding the referee process. *IMS Bulletin* **24**. 607–608 (1995).

[31] Su, W. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems* **34**. 27929–27939 (2021).

[32] Su, W. A truthful owner-assisted scoring mechanism (2022). arXiv preprint arXiv:2206.08149.

Xiao-Li Meng. Department of Statistics, Harvard University, Cambridge, USA, MA 02138.
E-mail address: meng@stat.harvard.edu