Understanding *Noise* in the Context of Noise[1]

Gaurav Sood and Andrew Gelman

12 Mar 2024

Noise

It can be natural to think of judgment in terms of mathematical functions in which the same inputs map to the same output. It turns out that this isn't even remotely true in many human decision-making systems. Take insurance underwriting, for instance. Given the same data (realistic but made-up information about cases), the median percentage difference between quotes between any pair of underwriters is a stunningly large 55% (which means that for half of the cases, it is worse than 55%), a difference that is about five times as large as expected by the executives asked about this scenario in a survey.

Several points flow from this result. First, if you are a customer, your optimal strategy is to get multiple quotes. Second, what explains ignorance about the disagreement? There could be a few reasons. First, when people come across a quote from another underwriter, they may "anchor" their estimate on the number they see, reducing the gap between the number and the counterfactual. Second, colleagues plausibly read to agree—less effort and optimizing for collegiality, asking, "Could this make sense?", than read to evaluate, "Does this make sense?"

This—the consequences of the mismatch between naive deterministic models and real-world individual and social decision processes—is the topic of the recent bestselling book, *Noise: A Flaw in Human Judgment,* by psychologist Daniel Kahneman, business strategist Olivier Sibony, and law professor Cass Sunstein. *Noise* is an appropriate follow-up to earlier books by Kahneman on heuristics and biases in human judgment and by Sunstein on legal and social institutions.

For another example, Kahneman et al. discuss a study from asylum reviews (Schoenholtz et al., 2009), writing, "A study of cases that were randomly allotted to different judges found that one judge admitted 5% of applicants, while another admitted 88%."

Variability can stem from two things. It could be that the data don't allow for a unique judgment (irreducible error). But even here, the final judgment should reflect the uncertainty in the data. Or that at least one person is disagreeing with the consensus, which can arise from: (1) variation in skill, e.g., how to assess visa applications; (2) variation in effort, e.g., some judges put more

---

[1] Review, to appear in *Chance* magazine, of *Noise: A Flaw in Human Judgment,* by Daniel Kahneman, Olivier Sibony, and Cass Sunstein. Much of this review is taken from posts by Sood in 2022 (http://gojiberries.io/2022/07/10/bias-in-noise/) and Gelman in 2021 (https://statmodeling.stat.columbia.edu/2021/05/23/thinking-fast-slow-and-not-at-all-system-3-jumps-the-shark/). We thank John Bullock and Don Green for helpful comments

effort than others; (3) agency and preferences, e.g., I am a conservative judge, and I can deny an asylum application because I have the power to do so; (4) biases induced by cognitive errors or the use of irrelevant information, e.g., weather, hypoglycemia, etc.

A lack of variability doesn't mean we are on to the right answer, but the existence of variability puts the lie to implicit deterministic models of the social world.

As quantitative social scientists, we appreciate the message of Kahneman et al. in this book. The concept of noise—in statistical terms, unexplained variation—is central to our work and to our understanding of empirical research, but we do not always think about how this interacts with what might be called the folk psychology of human decision making. Their point is not just that noise exists but that we need to directly confront our deterministic intuitions.

Mismatch between claims and research method

Ironically, *Noise* makes some of the errors that it warns about. Here are two patterns we noticed, which follow a general practice of not questioning results that are congenial to the main story:

1. *Extremely small n studies cited without qualification*. For example, "when the same software developers were asked on two separate days to estimate the completion time for the same task, the hours they projected differed by 71%, on average." The cited study (Grimstad and Jørgensen, 2007) was based on only seven developers. In a discussion of hypoglycemia and judgment, Kahneman et al. cite a paper on "what the judge ate for breakfast" (Danzinger et al., 2011) that used data from only eight judges.
2. *Surprising but likely unreplicable results*. For example, "When calories are on the left, consumers receive that information first and evidently think 'a lot of calories!' or 'not so many calories!' before they see the item. Their initial positive or negative reaction greatly affects their choices. By contrast, when people see the food item first, they apparently think 'delicious!' or 'not so great' before they see the calorie label. Here again, their initial reaction greatly affects their choices. This hypothesis is supported by the authors' finding that for Hebrew speakers, who read right to left, the calorie label has a significantly larger impact." As explained by Francis and Thunell (2020), this research (Dallas et al., 2019) has several problems which make us doubt that its claims would replicate or apply in the real world.

Earlier books of the authors have also cited published findings that were weakly supported. *Nudge* (Thaler and Sunstein, 2008) had enthusiastically cited the work of later-discredited behavioral researcher Brian Wansink (see Gelman, 2022), and in response to criticisms of claims about implicit priming in his 2011 book *Thinking Fast and Slow*, Kahneman (2017) wrote that he had "placed too much faith in underpowered studies."

We see a fundamental incoherence between the *Noise* book's key *substantive* point—that, contrary to intuition, individual and group decisions are noisy, that is, fundamentally unpredictable—and its *research method,* which is to interpret studies as implying something close to universal truths. To put it another way, if software developers' judgments are noisy and if judges' decisions are easily swayed by irrelevant factors, then how could we expect to extract general insights from a small study of seven or eight people at one place and time? This is related to the probabilistic arguments by Tosh et al. (2024) that it is not realistic to expect the coexistence of many large and consistent effects, as they would tend to interfere with each other.

A useful implication we have drawn from the general thesis of *Noise* is that our awareness of the unpredictability of individual and social decisions should make us skeptical of naive expectations of predictability even in a setting such as a study of insurance underwriting where there would seem to be clear incentives for economic efficiency, and also skeptical of simple solutions backed by unreplicated research. Noise can be measured and possibly reduced in level but, by its nature, it cannot be easily hacked.

The conception of the book

The study of noise and uncontrolled variation has a long history in statistics (Shewhart, 1939, Deming, 1981) and economics (Black, 1981), but the authors of this new book are neither statisticians nor economists. Indeed, they came at the topic as outsiders. For example, one of the authors of *Noise* said, "Unlike bias, noise isn't intuitive, which is why we think we've discovered a new continent" (quoted in Rusoff, 2021). From the standpoint of the fields of statistics, economics, or quality engineering, noise is not a new idea at all—so, at the very least, this author's previous unfamiliarity with the topic suggests that statisticians and economists have failed to fully communicate this idea to the general educated public. When one of the authors of a celebrated book says that he discovered something, he's saying that he recently became aware of something that was already well known, just not known in his social circle, in the same way that millions of people in 1492 already knew about the Americas, just not the people in the society where Columbus and other explorers had been living.

The *Noise* author also said, "One [of the things] I learned in this [book] collaboration is not to think in terms of [for instance], 'Will this stock go up'? 'Is this the right investment strategy?' but instead to think: 'What's the probability that this stock will go up?' 'What's the probability that this is a good investment strategy?' So rather than asking, 'Is it good to invest in international stocks [versus] domestic stocks?', it's better to ask, 'What probability do you assign to the proposition that international stocks will outperform domestic stocks in 2022?'"

We agree that it is a good idea to think probabilistically. But this also seems like common sense to us. Don't financial advisers tell you this all the time, that we can't know the future, we can only guess and at best assign probabilities?

We are reminded of that scene in one of the books by academic satirist David Lodge where a group of professors of English are sitting in a circle, playing a game where they take turns listing famous books that, embarrassingly, they've never read. And one of them lists Hamlet. A bit too embarrassing, it turns out! Similarly, it's kind of admirable how open the *Noise* author is about his former cluelessness, but it makes you wonder whether he was really the most qualified person to write a book about a topic that lots of people know about, but which until five years ago he'd never thought about.

Also, just a minor point, but we don't think it's quite right to ask questions like "What's the probability that this stock will go up?" Sure, you can ask the question just to check that your investment advisor is on the ball, but it does not make sense to think of the stock price going up or down as a binary outcome. The investment advisor should be thinking of things like expectation and tail risk. Anyway, not a big deal, but perhaps revealing of the authors' continuing discomfort with the concept of noise.

In summary, *Noise* discusses important issues connecting cognitive heuristics and biases to our understanding of decision making. It is an interesting book that relates to, but goes beyond, its authors' earlier influential work on cognitive psychology and social processes. The book also has a weakness in that its interpretation of research follows a fallacy that it warns us about in other contexts, which is to interpret unreplicable—noisy—research claims as implying an unrealistic predictability about the social world. This mistake may stem from the authors' unfamiliarity with the existing understanding of noise in statistics and economics. The authors do not seem to have fully incorporated the concept of noise into their understanding of statistical evidence. It can be challenging to explore a new continent without local guides who can show you the territory. That said, we appreciate the way the authors connect the statistical concept of variation to intuitions about cognition and decision making, and we hope this book spurs further work in this direction.

References

Black, F. (1986). Noise. *Journal of Finance* 41, 528-543.

Dallas, S. K., Liu, P. J., and Ubel, P. L. (2019). Don't count calorie labeling out: Calorie counts on the left side of menu items lead to lower calorie food choices. *Journal of Consumer Psychology* 29, 60-69.

Danziger, S., Levav, J. and Avnaim-Pesso (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 6889-6992.

Deming, W. E. (1981). Improvement of quality and productivity through action by management. *National Productivity Review* 1, 12-22.

Francis, G., and Thunell, E. (2020). Excess success in "Don't count calorie labeling out: Calorie counts on the left side of menu items lead to lower calorie food choices." *Meta-Psychology 4,* MP.2019.2266.

Gelman, A. (2022). Pizzagate and "Nudge": An opportunity lost. Statistical Modeling, Causal Inference, and Social Science blog, 4 June. https://statmodeling.stat.columbia.edu/2022/06/04/pizzagate-and-nudge-an-opportunity-lost/

Grimstad, S., and Jørgensen, M. (2007). Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software* 80, 1770-1777.

Kahenman, D. (2011). *Thinking Fast and Slow.* New York: Farrar, Straus and Giroux.

Kahneman, D. (2017). Comment on "Reconstruction of a Train Wreck: How Priming Research Went off the Rails," by Ulrich Schimmack. Replicability-Index blog, 14 Feb. https://replicationindex.com/2017/02/02/reconstruction-of-a-train-wreck-how-priming-research-went-of-the-rails/comment-page-1/#comment-1454

Kahneman, D., Sibony, O., and Sunstein, C. (2021). *Noise: A Flaw in Human Judgment.* Little, Brown.

Rusoff, J. W. (2021). Errors in economic predictions. Think Advisor, 4 May. https://www.thinkadvisor.com/2021/05/04/this-hidden-flaw-can-lead-to-grave-errors-in-economic-predictions/

Schoenholtz, A. I., Ramji-Nogales, J., and Schrag, P. G. (2007). Refugee roulette: Disparities in asylum adjudication. *Stanford Law Review* 60, 295-411.

Shewhart, W. A. (1939). *Statistical Method from the Viewpoint of Quality Control*. U.S. Department of Agriculture.

Thaler, R., and Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press.

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., and Hsu, D. (2024). The piranha problem: Large effects swimming in a small pond. *Notices of the American Mathematical Society.*