5

Taking Goodhart's Law Meta: Gaming, Meta-Gaming, and Hacking Academic Performance Metrics

James Griesemer

This chapter offers two thoughts to the meta-conversation about academic performance metrics and misconduct. One is that Goodhart's law (1975) concerns more than simply the idea of individual responsiveness to pressures from societal policies, for example, central bank monetary policies employ economic performance measures as standards of regulation and control in banking. The other concerns how we might exploit what *more* there is to Goodhart's law to probe the character of "mis"-conduct, as individuals and organizations adapt to, and comply with, academic performance metrics institutionalized as standards. Contrast this with "bad" conduct, as individuals and organizations cynically attempt to "game" or "exploit" the system to achieve a better evaluation *than their performance warrants*. Along with other chapters in this volume (Csiszar, this volume, chapter 1; Power, this volume, chapter 3), I suggest Goodhart's law describes conditions that not only undermine the representational success in modeling causal order in human social systems, but also the operation of the law *inverts* the causal order. Conversions of metrics into standards not only *invite* "gaming the system," they also practically construct "gaming" as the new form of practice, rendering the original product or practice to be measured as a "side effect" in the new causal order. Or, as Wouters (this volume, chapter 4) urges, we must distinguish gaming the system from properly functioning in an inverted system. It is thus problematic to moralize and shame so-called "predatory" practices as if it were clear what constitutes ethical, nonpredatory practice in social worlds where Goodhart's law operates.

Goodhart's lesson was that such measures are self-defeating because they invite "mis"-conduct. If people respond to standards as intended, the measure ceases to represent and record the primary target performance and comes to measure only compliance or conformity to the standard. The critique cuts deeper. As Lucas's (1976) critique of macro-econometric

models showed, such measures are self-defeating because the underlying causal structures of individual and organizational social behavior change when people and organizations respond to policies based on the models, so the policy causes the model to cease to represent the very thing the measure was designed to measure as it changes the system's causal structure. Goodhart's law and Lucas's critique apply to social policies built upon quantitative metrics taken as standards of quality and performance evaluation, for example, academic achievement. So the causal influence is reciprocal. This creates conditions for an arms race. The only escape from arms races is to realize their futility. Either policy-makers must stop making arms-race–producing policies, or the governed must revolt against intolerable institutions.

The second thought is that, philosophically, asking a different *question* may lead to more valuable insights than seeking *answers* to the question originally posed. Questions like "what is the nature of misconduct in response to academic metrics?" are of this sort because *posing* this question itself reinforces or facilitates circumstances in which Goodhart's law will apply. Framing the question of academic misconduct as one of gaming a system of metrics is to accept metrics as standards, if only for the sake of argument. Asking draws attention to the problem while entrenching presuppositions of the question. It is a prime mode of escalation in a metrics arms race between standards imposers and gamers. Csiszar (this volume, chapter 1) calls this an "inverse" form of Goodhart's law: "Only when a measure becomes a target is it widely taken up as a measure worth using." Practices and policies that use metrics *as* standards turn work performance, including scholarship, into a game in which the goal is to exceed the standard rather than perform the work that was to be measured. In other words, compliance with the standard becomes the goal rather than a side effect of the performance, and, in turn, performance of the work becomes a side effect of a policy imposing a standard. As Chamayou (2009) disturbingly but eloquently argues: the goal is to write articles, not do research. Other institutions are changing in parallel ways, so this is a widespread change across societies. Managing health *risk* has become the goal (and meaning) of health while curing illness has become the side effect of pharmaceutical use, rather than the other way around, as Dumit (2012) has shown.

Alternative questions are: If academic performance metrics are a game, can we hack them? If metrics-based standards and gaming the system create an arms race to nowhere (or to the decline and fall of the research system), what would it take to get rid of them? If metrics-based

academic performance standards can be hacked, then perhaps hacking would reveal their futility as unsuited to serve as *standards* to begin with. Perhaps hacking would redirect the question of "mis"-conduct back onto social policies of performance evaluation and quality judgment and away from charges of manipulated metric outcomes and undeserved gain. My aim is to propose interventions leading to better questions, whether or not they suggest good answers to old questions.

Performance and publication metrics are of epistemological interest concerning how scientific practices become transformed through institutional change, not only of ethical concern about the "conduct" of scientists and their publishers. Studies of gaming metrics may provide insight to philosophy of science as well as to research ethics.

Power's chapter and his book, *The Audit Society*, urge us to think about the kinds of problems surveyed in this volume from the point of view of narrative impact stories in addition to quantitative metrics (Power, 1997). The relation between individual narratives and aggregate impacts is a productive way to articulate where and how the "audit society" may have gone off the rails into the kind of self-defeating process Goodhart warned of. Turning a creative individual enterprise into the bureaucratic one of pursuing metrics as the *meaning* as well as the measure of productivity and success, of pursuing the "CV for its own sake," as Biagioli remarked in our workshop, is the moment of translation of performance in an audit society into a potential for "mis"-conduct, that is, conduct that emerges as unethical or unbecoming in the system *as it used to be* but which no longer functions that way. Institutionalized "mis"-conduct then becomes "bad" conduct when the ethics and optics of performance are *judged* against the old system but *measured* in the transformed system. Ethical evaluation lags metrical assessment, and "mis"-conduct that used to signal unethical "bad" conduct should now merely point to this misalignment, not express condemnation. The ethical judgment becomes misplaced and unjustified because the research system is no longer correctly understood epistemically.

Fundamentally, anyone interested in scientific practice must take into account how the research experience is now shaped by academic metrics. The "metric tide" (Wilsdon et al., 2015) demands that philosophy of science consider that science is now conducted in a regime of metrics, not only for research performance evaluation, but also for judgments and decisions affecting workflows in the research process itself: what problems and projects to pursue, what grants to seek, what personnel to hire, and what schools to attend, not merely which journals to publish in.

The comprehensive discussions and critiques in *The Metric Tide* report (Wilsdon et al., 2015) and *The Leiden Manifesto* (Hicks et al., 2015) indicate that there is a kind of balancing act going on in the metrics debate about what the future holds. The war on metrics seems to be over: metrics will not go away, even if which metrics are today's favorites will face a constant churn. We can describe the central tension in various ways, but they boil down to the idea that expert judgment should play a central role in evaluating research content and that quantitative data should play a role in measuring research productivity. Judgment cannot be automated, yet productivity can in some ways be measured.

The Leiden Manifesto says the problem is that performance evaluation is now *led* by the data *rather than* by judgment. If balance between judgment and data is the goal, then the question to ask is how to rebalance, reenergize, and reimagine a role for judgment in the face of the data-driven metrics onslaught. The metric tide can only rise because data now comprises a gravitational force tugging on the digital ocean; the energy of big data makes for ever larger waves. Solving the rebalancing problem will not be achieved by trying to turn back the digital ocean.

To explore how to rebalance judgment in evaluation of research, consider a thought experiment to probe the alleged greater "objectivity" and "reliability" of measurement with data and "subjectivity" of judgment. The experiment is to "go meta" in a *strategic* gaming response to metrics-based evaluation. I take inspiration from Daston and Galison (1992, 2007; Galison, 1998): the image of objectivity is historical, not static. They historicize the concept of objectivity by showing how it flip-flops between mechanical and judgmental zeitgeists.

Thought experiment isn't enough to advance our understanding of the problem sufficiently to pose the right questions, however. I propose actually *doing* the experiment, which I will call "hacking the metrics." If we think of hacking as a form of experimentation and experimenting as a form of research, the idea is that hacking can be both a means of intervening in the metric tide and a legitimate mode of research about it.

Experimentation of this kind should challenge ethical intuition about conduct and at the same time *transparently* undermine metrics for experimental purposes while redirecting questions about performance and rebalancing evaluation toward judgments of the work.

One way to make metrics hacking into a research enterprise is to exploit causal aspects of Goodhart's law. The idea is to reframe the question of gaming as a problem of causality rather than representation, one that can be tested by experimental intervention in ways that might actually change

the use of metrics as standards by disrupting the policy debate. If metrics are subjected to hacking, then perhaps the end game of finding an unhackable metric will come to seem a hopeless task to policy makers bent on automating judgment or replacing trust with "objective" quantitative measures—a fool's errand as ridiculous as an uncrackable cryptographic scheme or a winnable nuclear war.

The solution *du jour* to the problem of gaming performance metrics like citation counts, h-indexes, or impact factors is to multiply metrics and form a "basket" of them, each metric serving a particular component job well rather than hoping an all-purpose metric will emerge from the arms race. I am on board with the "alt-metrics" or basket approach as methodological antidote to the idea that some single, best all-purpose metric will be found to replace citation count or impact factor, but I am skeptical that it is going to slacken the metric tide because we are already in a positively reinforcing arms race with the metrics. Goodhart's law should apply to baskets of metrics just as much as to any single one. This should be so because the only way to *deploy* any metric as an evaluation standard without violating Goodhart's law is to keep the metric or standard secret. The basket approach only does this by being too complex to understand, or proprietary and thus secret, so that it is *de facto* hidden from the day-to-day practice of individuals subject to performance evaluation. That strategy will have only temporary success because academics are clever and are paid to solve puzzles of this kind, hence the inevitability of an arms race. Moreover, research production systems and research evaluation systems are inextricably linked by processes such as peer review for publication and grant award, which must be transparent (or would be extremely unethical, like not telling assistant professors what is required for tenure). So I don't see an end game in the basket of metrics approach, other than mutually assured destruction. Now that may not be a bad thing. Creative destruction of an old-fashioned, biased, elite research system might be justified if it brings down an ill-suited, ill-fitting, conservative system of research evaluation with it.

However, there is more than one way to creatively reimagine a role for judgment alongside metrics in performance evaluation. I return to Goodhart's law as a way of talking about how to reframe the question. A reframed question may lead to different kinds of solutions than a basket approach or *Leiden's* policy demand that inappropriate measures such as impact factors be dropped as standards for individual authors.

Goodhart's law teaches that to understand what goes on in the world of metrics-based academic performance evaluation, we should look

beyond how metrics might be gamed and beyond the question whether, because the metrics are used as standards, gaming the metrics is a form of misconduct in the bad sense of gaining something undeserved through the evaluation process. What we need to understand is how what individuals *do* in their research production processes in response to the standard changes in relation to the metrics as a consequence of their use as standards. That is not merely a matter of assessing whether, how, and to what extent researchers from this or that part of the globe decide to submit their research findings, data, figures, or code to a "predatory" or "junk" journal rather than to a "legitimate" one—nor whether they decide to analyze other people's hard-won published data rather than go to the trouble and expense of generating their own data, nor to Photoshop old figures rather than produce new ones based on new data. It is a matter of understanding deeply, and on a fine-grained scale, how research production—the whole content of research activity, and not just "publishing"—is changing in an environment where performance metrics function as standards.

Assessing the content of research *activity* is not the kind of problem that can be solved with more data captured in metrics with the digital discovery methods discussed in this volume. Transformations of research production processes at the level of the conduct and decision making of individual scientists and small teams in their day-to-day workflows can only be captured by good old-fashioned social science *research* that social scientists use to figure out what, how, and why people do what they do: interviews, participant observation, close readings of unpublished and published work, surveys, and now distance communication through the internet.

The strategic intervention I propose is this: Let's assume Goodhart's law is true. If it's true, we should *expect* "gaming" and "mis-"conduct in the system as a normal or typical part of the workflow of any well-adapted research production system whenever metrics are used as standards of research performance evaluation. So, it may be more fruitful to look for situations in which research performance appears *not* to adapt as a means of revealing *breakdown* of the kinds of behaviors one should expect to find when metrics are made into standards.

Adaptation is the most plausible state of affairs in response to the imposition of the social forces represented by metrics used as standards. In that light, the phenomenon of "predatory" or "junk" journals is one kind of response we might consider interpreting as "mis-" rather than "bad" conduct. Maladaptation or nonadaptation, in the sense of nonconformity or noncompliance to a standard, *should* appear anomalous or "bad" to an institutional policy regime that expects compliance. More

radically, hacking metrics as a form of experimental, *transparent* malor nonadaptation might serve as a means of understanding the mechanisms and dynamics of adaptive responsiveness to imposition of metrics as standards. The further point is that we can think about two ways of understanding what Goodhart's law tells us about the variety of kinds of behavior subject to lumping under "gaming the system." One is the sort of gaming discussed in this volume, and it is quite interesting.

The other kind of possible response to Goodhart's law is to *embrace* gaming the system as a tool for experimental intervention into research *production* systems, extending the traditional observational tools of social science research. The idea is to make hacking the research system part of a research *program* for understanding what causal consequences, at the micro-level of research production, follow from the imposition of social forces at the macro-level of social organization. (In a sense, "predatory" or "junk" journals can be viewed as leading the way in hacking, provided their interventions are interpreted according to the proper causal-experimental framework. We can learn much from their tactics even if we eschew their profit motives.) To use hacking as a research tool, we need to adapt research production work *on science metrics* to a new goal: experimental intervention into research performance systems that are subjected to metrics-based performance evaluation. To do that, we would need to create an artificial (i.e., experimental) publication system in which such research work could be published and an artificial (i.e., experimental) research specialty that organizes it.

We need not only to multiply the *metrics* as in the basket approach, but also to actually hack gaming *behaviors* in order to find out how the causal structure of individual researcher behavior is changed by the imposition of metrics. Traditionalists might wonder why we couldn't just interview people who may have been subjected to behavior change in the face of the metric tide or do longitudinal studies of people experiencing different metrics environments. We could do this of course. But it seems unlikely we would observe appropriate contrasts among scientists in their experiences of metrics-as-standards to develop much insight into causes. Even comparison of researchers in closely related specialties or in different national contexts or across historical periods would have so many confounding variables at work as to render them of limited utility for discovering causal impact.

What I propose instead would be a program of intervention with individual researchers and research groups, while historians, sociologists, anthropologists, and philosophers—such as science studies researchers—study

them, in an enhanced environment where science studies takes science metrics into *account* descriptively while also manipulating the research environment in which performance is evaluated. Such experiments are taking place within the current research system, for example, by Labbé's experiments with his fictional author, Ike Antkare (this volume, chapter 14). I propose parallel experimentation with the research publication system itself, using methods inspired by Labbé and others.

The tension between measurement and standards deriving from Goodhart's law, as I've noted, is that because humans are reflexive, metrics used as standards have to be kept secret, otherwise the people who are measured will change their behavior in ways that defeat the value of the metric as a measure. On the other hand, a standard has to be transparent: it is unfair and unproductive to hold people to a standard they cannot strive to meet. The problem is that secrecy and transparency don't work so well together.

I propose we use our own reflexivity as technoscience researchers and scholars of the operation of academic metrics, in the tradition of medical self-experimentation, to manipulate experimentally the conduct of science studies research to find out what kind of changes can be brought about by exploiting and manipulating metrics-based measurement of performance.

In other words, a way Goodhart's law could turn out to be *true* is not merely that people are responsive to these forces in ways that change the causal structure of their behavior—Goodhart's law could be true because people respond reflexively to satisfy Goodhart's law *on purpose* as a way of playing a game. Just as video game hackers improve game play by intentionally violating the designs of the game designers to make the game play differently, thereby inventing a new game, science studies research might engage in experimental manipulation of their own metrics as a means of understanding contemporary science in the age of metric tides.

That would be to change the *causal* relationships of the game—to make it a different game, not merely to play "the game" by explicitly engaging in "mis"-conduct or "bad" conduct with respect to the institutionalized rules of the game, but to invent a new game by hacking the old one. What can we do to investigate metrics on a par with the kind of hacking that goes on in the video game world?

One way to do it is to create a collection of journals designed to publish research on science metrics but which includes in their mission the explicit "gaming" of metrics that are used as standards in performance evaluation. Call them "PuLP" for "Public Library of Philosophy" on the

model of PLoS—Public Library of Science. (Thanks to Jonathan Eisen for suggesting the acronym PuLP.)

The mission statement for "PuLP-ONE" would include:

1. Reviews and surveys of current performance measures/metrics and which ones are used by whoever's policies as standards. As it happens, journals are already beginning to appear that have this scope, for example, *Research Integrity and Peer Review* (http://researchintegrityjournal.biomedcentral.com).
2. Success and failure impact narratives authored by individuals and groups about what metrics have done for/to them.
3. Science studies research on practices considering the role of science metrics in the conduct of research or its evaluation, such as the chapters in this volume.
4. Overt hacking of metrics by publishing work of the above three kinds and of any other kind (including machine-generated papers) in whatever nominal field of study so as to explicitly and transparently attempt to manipulate metrics and thus to game standards.

A system of PuLP journals, a public library of hacker science studies, so long as it is sufficiently amusing to indicate transparently that its goals are not business as usual, would be designed primarily to transparently and openly *intervene into* how metrics affect research or behavior. PuLP would, for example, publish science studies work with, say, five hundred authors, citing articles in other PuLP journals and also other journals the authors regularly publish in. By manipulating the number of articles published in a particular PuLP journal and the number of citations to articles in that journal, we could not only manipulate the Journal Impact Factor, as many "predatory" journals already do, we could also engineer whatever impact factor we wanted, showing just how arbitrary a measure it is and just how irrelevant to the content of research. The "h-index," as a measure of author impact, could be manipulated by engineering many citations to works published by that author from other PuLP journals as well as by seeking agreement of those publishing in PuLP to cite PuLP journals in their works published in non-PuLP ("civilian") journals. The full range of tactics discussed in other chapters in this volume would not only be available to authors in PuLP, but would also be part of the mission to use these methods and build new ones.

Because the journals would also be designed to publish scholarly research *assessments* and *interpretations* of how changes in research production behaviors undermine metrics and at the same time reorient or

redirect workflows and production of whole fields of scientific research, PuLP journals could not be dismissed as "mere" junk, especially if their papers published under missions one through three are of high quality. PuLP journals would provide a "respectable" outlet for science metrics research. If this mission is held to community standards of scholarship, it would be harder to discount the hacker work out of hand on the self-serving grounds usually supporting metrics-based assessment in the first place: appearance in journals that meet metrics-based standards.

Scholarly publications in PuLP journals that report, assess, and interpret responses to metrics-based research performance and impact narratives could serve as a basis for designing, announcing, and conducting new hacking techniques and experiments, as the project would presumably kick the arms race into a higher gear, especially when hundreds of science studies researchers begin to submit experimental CVs for personnel evaluation with dozens of publications per month.

The primary mission of PuLP would be to undermine existing metrics by embracing and exploiting Goodhart's law. PuLP-ONE would be a *transparent* journal for hacking the metrics—not designed for the sake of gain like a junk journal might be: to make money or earn prestige for authors, but for the sake of understanding experimentally how metrics manipulate social behavior, thereby showing how they are subject to gaming and to undermine their use as standards lacking a balanced involvement of judgment in evaluation. The goal would not be just to tell stories about how scientists conduct their research and the sorts of pressures they experience, but also to interrogate and ultimately change practices of research performance evaluation.

In providing a forum for *review* of research metrics and *assessment* of responses to metrics as standards, and thus the material platform needed to design hacker interventions, PuLP might help end the arms race of metrics and gaming by revealing the likely decline and fall of the research system from the inadvertent, unintended consequences of continued pursuit of metrics-based evaluation.

Enthusiasm for metrics is reinforced by bigger and bigger data, so it is probably necessary to do this experiment and not only talk about it. The aim is not to do away with research metrics but to return the project of *evaluating the metrics* to serve researcher valuation of the content of their research and to repair the damage caused by diverting this value into a side effect of a transformed research system that mainly values the advancement of auditable knowledge. If sustaining the research enterprise requires hacking the metrics, let the games begin!

References

Chamayou, Grégoire. 2009. "Petits Conseils aux Enseignants-Chercheurs qui Voudront Réussir Leur Évaluation." *Contretemps*. February 24, 2009, http://www.contretemps.eu/petits-conseils-enseignants-chercheurs-qui-voudront-reussir-leur-evaluation/. Consulted June 16, 2017. Read from Google Translate into English.

Daston, Lorraine, and Peter Galison. 1992. "The Image of Objectivity." *Representations* 40:81–128.

Daston, Lorraine, and Peter Galison. 2007. *Objectivity*. Brooklyn, NY: Zone Books.

Dumit, Joseph. 2012. *Drugs for Life: How Pharmaceutical Companies Define Our Health*. Durham, NC: Duke University Press.

Galison, Peter. 1998. "Judgment Against Objectivity," in C.A. Jones and P. Galison (eds.). *Picturing Science, Producing Art*, 327–359, New York: Routledge.

Goodhart, C. A. E. 1975. "Monetary Relationships: A View from Threadneedle Street," in *Papers in Monetary Economics*, Volume 1, Reserve Bank of Australia. Quoted in: *Goodhart's Law: Its Origins, Meaning and Implications for Monetary Policy*, by K. Alec Chrystal and Paul D. Mizen. Prepared for the Festschrift in honor of Charles Goodhart to be held on November 15–16, 2001 at the Bank of England. November 12, 2001, http://cyberlibris.typepad.com/blog/files/Goodharts_Law.pdf.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. "The Leiden Manifesto for Research Metrics." *Nature* 520(April 23):429–431.

Lucas, Robert 1976. "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer (eds.). *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy 1, 19–46, New York: American Elsevier.

Power, Michael 1997. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

Star, Susan L., and James R. Griesemer. 1989. "Institutional Ecology, 'Translations,' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907–1939." *Social Studies of Science* 19:387–420.

Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. doi:10.13140/RG.2.1.4929.1363.

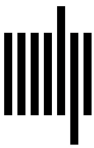
Gaming the Metrics

Misconduct and Manipulation in Academic Research

Edited by: Mario Biagioli, Alexandra Lippman

Citation:

Gaming the Metrics: Misconduct and Manipulation in Academic Research
Edited by: Mario Biagioli, Alexandra Lippman
DOI: 10.7551/mitpress/11087.001.0001
ISBN (electronic): 9780262356565
Publisher: The MIT Press
Published: 2020



The MIT Press

This title is freely available as an open access edition thanks to the TOME initiative and the generous support of the University of California, Davis. Learn more at openmonographs.org.

This book was set in Sabon by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Biagioli, Mario, 1946- editor. | Lippman, Alexandra, editor.
Title: Gaming the metrics : misconduct and manipulation in academic research / edited by Mario Biagioli and Alexandra Lippman.
Description: Cambridge, MA : MIT Press, [2020] | Series: Infrastructures | Includes bibliographical references and index.
Identifiers: LCCN 2019010150 | ISBN 9780262537933 (pbk. : alk. paper)
Subjects: LCSH: Scholarly publishing—Corrupt practices. | Learning and scholarship—Corrupt practices. | Research—Corrupt practices. | Communication in learning and scholarship—Moral and ethical aspects.
Classification: LCC Z286.S37 G36 2020 | DDC 070.5—dc23
LC record available at https://lccn.loc.gov/2019010150

10 9 8 7 6 5 4 3 2 1