Building Less Flawed Metrics: Dodging Goodhart and Campbell's Laws

Manheim, David

University of Haifa School of Public Health

26 November 2018

Building Less Flawed Metrics
Dodging Goodhart and Campbell's Laws

David Manheim

January 23, 2020

ABSTRACT

Metrics are useful for measuring systems and motivating behaviors. Unfortunately, naive application of metrics to a system can distort the system in ways that undermine the original goal. The problem was noted independently by first Campbell, then Goodhart, and in some forms it is not only common, but unavoidable due to the nature of metrics. (Campbell, 1979; Goodhart, 1975; Rodamar, 2017; Manheim & Garrabrant, 2018) There are two distinct but interrelated problems that must be overcome in building better metrics; first, specifying metrics more closely related to the true goals, and second, preventing the recipients from gaming the difference between the reward system and the true goal. This paper describes several approaches to designing metrics, beginning with design considerations and processes, then discussing specific strategies including secrecy, randomization, diversification, and post-hoc specification. The discussion will then address important desiderata and the trade-offs involved in each approach, and examples of how they differ, and how the issues can be addressed. Finally, the paper outlines a process for metric design for practicioners who need to design metrics, and as a basis for further elaboration in specific domains.

Keywords: Metrics, Measurement, Complex Systems, Perverse Incentives, Cobra Effect, Goodhart's Law, Campbell's Law

WHAT IS THE PROBLEM, EXACTLY?

Metrics, key performance indicators (KPIs), targets, quantifiable goals, measurable results, and objective assessments are a few of the terms that get used to refer to the modern obsession with numerical and therefore seemingly scientific ways to understand human systems. These trends have led to improvements in business processes, in medicine, in public safety, and in both primary and higher education. In part as a result of this success, there have been highly publicized failures of the ever-more commonly applied paradigm. These occur when the measure isn't aligned well enough with the true goal, when the system promotes cheating, or when a formerly useful measure is applied despite underlying changes that make it no longer relevant.

Both Campbell and Goodhart identified an important failure mode for measurement, which was later paraphrased by Mary Strathern as "When a measure becomes a target, it ceases to be a good measure." (Campbell, 1979; Goodhart, 1975; Strathern, 1997) Campbell, who seems to have discovered the concept first, was a social scientist looking at how metrics distort behavior and lead participants in a system to attempt to exploit the metrics. (Rodamar, 2017) Goodhart, on the other hand, was an economist noting a structural breakdown in inference about a system which occurs when rules change - a precursor to the now-famous Lucas critique in economics. The dynamics involved in these failures, however, are more complex than either discussed at the time, and several distinct failure modes and underlying dynamics have been identified (Manheim & Garrabrant, 2018), which can be simplified into a few cases.

Delineating the Problems with Metrics

There are four main issues with metrics that lead to the Goodhart-Campbell failure modes. The first relates to imperfect correlation, the second to misusing correlation to cause perverse incentives, the third to relative difficulty of good metrics, and the last is confusion about the goal to be measured, or worse, fundamental incoherence.

In the first case, a metric that is currently statistically correlated with the goal will inevitably be less closely correlated once the metric is used, for example when conditioning on high values of the metric. As an intuitive example, height and basketball skill are correlated, but among the tallest people, it is unlikely that the best few basketball players are also the tallest. An additional well-understood but still common problem is ignoring the difference between causation and correlation - a cardinal sin when attempting to improve a system. For example, high school grades correlate with college success, and all else equal a student who takes easier classes in high school will receive higher grades. Even if this does not change student behavior, by selecting for high grades, colleges may inadvertently select for students who care about their grades more than about learning or challenging themselves. A related problem occurs when a metric is correlated with an intermediate measure which itself correlates with a goal. This has the added issue of ignoring the simple mathematical fact that correlation isn't commutative. As an example of this non-commutativity of correlation, taller people are better at basketball, and coordinated people are better at basketball, but (ceteris paribus,) taller people tend to be less well coordinated.

In the second case, not only is a metric without a causal relationship invalid, but using it can be pernicious. Validity of a measure is a critical ontological and epistemological necessity in research, and validity requires a causal relationship. (Borsboom, Mellenbergh, & van Heerden, 2004) This is not just correlation being confused with causation, and is not nitpicking limited to the philosophy of science. It is instead a fundamental issue with perverse effects of incentives that are causally disconnected from the goal. When explicitly optimizing a system using a metric, the optimization can change the system to make the metric not only invalid because participants react to the new rules, as Campbell noted, but actively harmful. Creating incentives for metrics that correlate with but do not cause the eventual goal will, unsurprisingly, be pursued

in ways that may not cause the goal. If a teacher notes that students who ask questions learn more, they might announce that they will assign a portion of the grade based on number of questions asked in class. The new incentives are likely to have the perverse effect of incentivising questions that detract from student learning, and by carelessly incentivising part of the system, the metric not only failed to capture the important feature, but actually harmed the intended goal.

In the third case, easy to measure is rarely the same as important (Hubbard, 2007), and easy to understand isn't the same as relevant. In nutrition research, self-reported diet and energy intake is a relatively easy quantity to measure, but is inaccurate (Schoeller, 1990), and is obviously easy for a respondent to falsify. In the same realm, fat intake is easy to understand, and at one point dietary fat intake was considered bad, but closer examination found that the specific class of fat was critical; eating a lot of trans fats has negative impacts (Liska, Cook, Wang, Gaine, & Baer, 2016), while it seems polyunsaturated fats have positive effects (Clifton & Keogh, 2017), and saturated fats have unclear effects(Szajewska & Szajewski, 2016). For these reasons, using "fat intake" as a metric can be very misleading - but since it was easy to measure, and the information is easily available, it can still be a default driver of behavior.

In the fourth and final case, the goal is incoherent or conflicting. The three cases above make an implicit assumption shared by both Goodhart and Campbell, that the goal is coherent and understood. A simple example of where this assumption fails is a committee composed of individuals with differing values and goals. If the differences are not understood, the goals are often incoherent. Even if they are understood, however, the individual goals may be diverse or even incompatible. If so, there may be no way to assign a coherent metric that will align (or even correlate positively) with all of them. For this reason, if the choice of metric is a compromise that doesn't address the conflict, the metric chosen and resulting incentives may be incoherent. A similar problem occurs when the desired outcomes are unclear to the people setting goals. Finally, when the outcomes do not occur within a time frame that can be captured, the intermediate outcomes may have unknown relationships with the final goals, making any metric potentially incoherent.

An example of all of the issues in this final case occurs in the education system. The desired outcomes of education include life-satisfaction, fitness for the future job market, fostering the intellectual curiosity of students, and/or creating informed citizens. These are all long-term, and thus hard to measure or discuss concretely, are not often discussed by those setting priorities, and are often conflicting. Unsurprisingly, various intermediate metrics like GPA, even at the college level, or college completion, are poorly correlated with the desired long-term outcomes (Caplan, 2018) - and the difference is subject to gamification. (Hess, 2018) This is unsurprising - the degree to which incoherent, conflicting, or poorly defined goals can be achieved is intrinsically limited. Worse, as Deresiewicz argues (Deresiewicz, 2015), imposing simplistic metrics distorts education in a way that defeats the original goals.

Addressing the Problem

The problem statements above seem to suggest solutions. Unfortunately, these solutions are not always simple or practical, and as we will explore later, the approaches are viable and acceptable in different areas. Still, concrete examples of how the problems are typically addressed can be helpful in understanding what viable and non-viable solutions look like.

To address the problem of collapsing correlation, it is often possible to build metrics that more closely relate to the actual goal. In our first example, instead of using height as a proxy for basketball ability, we can use a weighted sum of height, athleticism, mastery of basketball skills, and experience. This will improve the model, but unless a clear causal model for basketball ability is found, it will be only a partial solution. In the second example of college use of grades, they can measure the relationship between student behavior like choosing easier classes and college success, instead of making the mistaken assumption that correlation is transitive. For this reason, colleges might specifically consider advance placement and international baccalaureate classes as a marker. Unfortunately, investigating all the potential confounding interactions between high-school choices and college success (which itself must be measured in ways that are fallible,) is a much larger project, and it still does not ensure that causal mistakes would not allow other forms of collapse. For example, perhaps hours of studying is caused in large part by interest in academic subjects, which causes later success. Selecting students who participate in study groups would seem to help, but perhaps attendance at such groups is itself due to poor grades and disinterest in the subject, so that it will anti-correlate with the actual cause of later success.

To address the second problem, of metrics distorting the system, we need a two pronged approach. The first prong requires insisting on metrics robust to changes, such as ones using models of the system that represent how the measured quantity relates to or affects the goal. In the example, if the causal relationship between seating and performance is understood, the chosen metrics will properly represent the determining factors of the relationship, such as student motivation and attention paid. The exercise of thinking through the causes will hopefully make it clear that re-arranging seats will have minimal effect. While these observations are sometimes obvious, discovering causal relationships is in general complex. The second prong is ensuring metrics are not being manipulated by the participants, or at least minimizing this manipulation - via secrecy, randomization, or post-hoc choice of metrics. For example, if students are unaware that grades will be assigned based on seat position instead of work done, their actions will less severely distort the metric.

Lastly, incoherence and debated goals can sometimes be addressed with structured discussions leading to increased clarity. In such situations compromise is often needed. Abandoning the search for an optimal solution or compromising on key goals may seem unfortunate, but the alternative of using incoherent metrics based on incompatible goals is often worse than doing nothing at all. Furthermore, where clarity and compromise are possible, coherent goals can be found that (in the terminology of the late, great Herbert Simon) satisfice (Simon, 1956) - that is, the compromise goals and resulting metrics

lead to solutions that are acceptable instead of optimal. Alternatively, more complex approaches like Robust Decision Making which replace metrics and allow accounting for deep uncertainty, including disputed values, are effective. (Lempert, Groves, Popper, & Bankes, 2006; Kalra et al., 2014) Unfortunately, these usually cannot replace metrics since they require far more difficult to understand methods, as well as needing both analytic expertise, and intense management involvement.

METRICS AND INCENTIVES ACROSS DOMAINS

The "Scientific Management" movement was an early proponent of reward systems similar to those seen in use in corporations today; profit sharing, per-task payments or bonuses, and merit-based pay(Caudill & Porter, 2014). In each case, the reward is tied to a metric. On the other hand, motivators are complex, and there are important trade-offs between the various positive and negative reward factors. (Herzberg, 1968) These trade-offs are not just practical, but have significant ethical implications, leading for some to call for an ethics of quantification. (Saltelli, 2020) Rewards to motivate behavior, and punishments to prevent behavior, are much more general.

Clearly, metrics are not limited to the domain of management, and the issues in other domains can differ. Public policy often uses tax incentives such as credits or deductions to change public behavior via the 'metric' of taxes owed. Here, despite the obvious incentives involved, this type of policy intervention has limited effectiveness on public behavior due to complexity, non-immediacy, and suspicions of unfairness. In the measurement of autonomous vehicles, a recent report suggested that the measures must be "valid, feasible, reliable, and non-manipulatable," (Fraade-Blanar, Blumenthal, Anderson, & Kalra, 2018) implicating many of these same concerns.

In addition to cases above where at least a semblance of a metrics is seen, the desiderata usually extend to motivation systems in general. For example, punishment systems have many similar features - law enforcement is less effective when arbitrary, when the punishments are often avoided, or when the perpetrators of what would normally be criminal acts find technical ways to avoid culpability. Prize competitions use measurement even more directly as a motivator, but participation will be limited if potential recipients worry about unfair treatment or corruption. Lack of clarity about goals, discussed above, would be even more critical when designing a direct incentive, because without specification the people being motivated will not understand the goal, or be able to know when it has been accomplished. If it is instead specified clearly despite incoherence, rewards are likely to be either impossible to receive, or trivially accomplished in ways unrelated to the goal.

STRATEGIES AND TRADE-OFFS

The design of metrics requires both an understanding of the goals, the potential strategies available, and the trade-offs involved. To introduce these issues, we first outline a number of useful desiderata for a metric. Following this, there are a number of specific metric

design considerations and strategies that involve the process of creating and considering the metric. These are not reflected in the metric itself, but can lead to better choices of metric. Lastly, there are a set of metric features. Concluding the discussion of those features is a final point that not all problems can be effectively addressed using metrics. In such cases, rather than abandoning concrete numerical metrics altogether, we should start by reconceptualizing what they are being used for, and how.

Metric Desiderata

There are many properties of metrics that exist in tension with one another. Ideally, of course, we want metrics that give free, understandable, fair, incorruptible, and immediate insight. Unfortunately, we instead often get expensive, confusing, biased, unreliable, and out-of-date metrics that provide little insight. In addition the operational challenges like cost and availability, there are desiderata involved in choosing and using metrics for decision making and incentives. The exact trade-offs between various motivational factors are a matter of intense empirical focus, and different domains have additional critical desiderata, but stepping back from those discussion we can see that five we will discuss are often important.

Metrics generally benefit from (1) immediacy, (2) simplicity, (3) various forms of fairness, (4) Trust and transparency, and (5) non-corruptibility. Specifically, immediacy is useful for ensuring feedback can be applied quickly, and participants can learn what is expected. For example, delayed rewards like end-of-year bonuses may be less effective motivators than immediate feedback. Overly complex metrics may be less effective in motivating behavior, and impose costs on both the participants and the evaluators. Transparency is important for trust, it may be a regulatory or legal requirement, and can help avoid or mitigate principle-agent problems. Secrecy also undermines perceptions of fairness, which can create issues of trust. Fairness is also important for legal and social reasons, and even if an unfair metric is able to accomplish the intended narrow goals, it can lead to longer term issues and undermine social trust. Corruption, of course, is a more direct attack on many of these desiderata, and either the perception or the reality of manipulation can do enough harm to more than outweigh any possible benefit from the use of a metric. More central to the problems of Goodhart's and Campbell's laws, employees almost always analyze the system and are intentionally or unintentionally motivated to circumvent the intent to achieve the stated goals.

Realistically, metric design needs to accommodate the reality of what is possible, and keeping the various desiderata in minds makes it possible to make informed choices when choosing or designing the metrics and incentives for our system. The importance of a desiderata in a given domain must be weighed against the costs, the importance of preventing gaming, and the impact of gathering the data. We first define the desiderata, so that we can note where there are obvious advantages or conflicts that should be considered.

Cost: Is the extant data free? Alternatively, how expensive is it to collect the data

needed to compute the metric?
Availability: Is the data needed to compute the metric available, or does it need to be collected? Are there lags in the process?
Immediacy: Can the metric and or incentive scheme provide feedback rapidly enough? Will lags in the system create instability or uncertainty?
Simplicity: Is the metric easy or difficult to understand? Are the inputs to the metric understood? Are the implications of behavior clear? Will participants understand these factors well enough for it to influence their behavior, and/or well enough to attempt to manipulate it? Will this change over time (in good or bad ways) as participants become accustomed to the system?
Fairness: Is the metric commensurate to actual goals? Does the metric provide disproportionate benefit to some groups? Do behaviors that get influenced by the metric impose costs elsewhere in the system?
Trust: Do administrators and participants trust one another not to manipulate the metric? Can manipulation be observed by both parties? Will participants and administrators trust the system or the transparency measures enough to believe that it is not being manipulated?
Non-Corruptibility: Who has access or ability to change the data or manipulate it? Does the metric introduce exploitable information asymmetries? Can the system be used by participants to cheat? Can it be manipulated by administrators?

Design Considerations

In light of the challenges discussed at the beginning of the paper, and the desiderata listed above, we suggest five general thought processes and factors to consider which can be useful in designing better metrics, with a focus on avoiding metric over-optimization failures and corruption.

Coherence. If the goals of a system are incoherent, or are poorly understood, it will be difficult for any metric to capture them. For example, it is easier to measure lines of code written by a programmer than it is to judge how well the code performs. In some cases, the metrics in place serve simply to justify the status quo, or to act as window dressing. Promotions in companies may in theory be based on metrics, but if managers can choose to apply the metrics selectively, this can serve as a mask for justifying decisions made on a different basis.

There is a common temptation, in part driven by cost, to find easy to measure outcomes instead of choosing based on how well a measure represents the goals, or based on the value of better data(Hubbard, 2007). Unfortunately, this temptation is too-often yielded to in practice, either due to lack of thought, or too little consideration of the impacts of poorly built metrics. This is especially common given incoherent (or under-specified) goals, where the fuzziness leads to losing sight of the purpose, and not measuring what is important to the process. (Soares, 2015) This confusion is a key cause of strategy surrogation, where managers forget that measures are imperfect proxies, and

improperly reify the measures as identical to their goals. (Choi, Hecht, & Tayler, 2012)

Causal Forethought. Sometimes the metric measures something related to the intended goal with an unclear or non-causal relationship. If this is the case, a reward system using that metric can creates incentives that make the relationship between the metric and the goal disappear. For example, measuring attendance in class may increase attendance, but if the otherwise-absent attendees spend their time in class sleeping, or being disruptive, it is possible that nothing will be gained. A theory of change is helpful for clarifying these relationship and avoiding this class of error. (See Taplin and Clark's book[1] (Taplin & Clark, 2012), for a clear introduction to theory of change.)

Structured Discussions and Compromise. In situations of deep uncertainty and conflicting goals there is often a need for discussion and compromise. While no compromise can achieve conflicting goals, deep exploration of problems can often lead to agreements that are better for all participants than the alternatives. (Rosenhead & Mingers, 2001) While useful, these methods require extensive and costly analysis and discussion, and are therefore ill-suited to many smaller-scale problems.

Pre-Gaming. If a metric is proposed, the exercise of imagining how it could be gamed, and building incentives aimed at forestalling gaming, can be useful. This idea is closely related to research about the effectiveness of such planning by Mitchell, Russo, and Pennington(Mitchell, Edward Russo, & Pennington, 1989), which Gary Klein later popularized as a "pre-mortem"(Klein, 2007). If done well, these can be very helpful - but they are oten done poorly(Klein, Sonkin, & Johnson, 2019). After identifying likely failure modes, it may be possible to improve the metric, or add explicit conditions to the rewards to thwart the failure modes that were discovered. Despite the desire to restrain gaming, however, care should be taken to ensure that the metric does not dictate exact methods, which can stifle innovative for accomplishing the overall goal. For example, measuring hours of classroom time spent by a teacher may discourage time spent on lesson planning, peer consultation, and other activities that improve effectiveness of the time spent in class. Explicitly requiring each of those specific activities to account for the potential failure, however, removes discretion that allows teachers to pick the activities that are most beneficial in their case.

Monitoring Behaviors. Even when well designed and initially effective, metrics have a tendency to go awry over time as systems and behaviors change. Explicitly setting checkpoints and reviews for metrics may be useful for ensuring that these systemic drifts are limited in scope. This is especially useful when it is easy it is easy to detect behaviors which effectively cheat[2]. For example, metrics often promote a short term intermediate goal, like sales of a certain product, or short term ad-revenue. Incentives may start encouraging overzealous sales activities, or placement of ads that interfere with user happiness or engagement, in each case potentially preventing longer-term growth. Overzealous sales activities would be visible in lower repeat sales or reduced customer satisfaction, maknig detecting this failure relatively easy. Designing perfectly coherent

[1] Available online here: `http://www.theoryofchange.org/wp-content/uploads/toco_library/pdf/ToCBasics.pdf`

[2] I am grateful to Davide Balzarotti for this insight.

metrics aligned with goals for the system overall may be infeasible, but monitoring behaviors that metrics incentify can detect or prevent larger distortions and later systemic failures.

To conclude the discussion of design processes, it is critical to again note the trade-off between ease of measurement, cost of measurement, and the better solutions that can result from the above processes. This means that the time invested in metric design should be commensurate with the importance of the metric, the potential impacts, and the likelihood of manipulation or perverse effects. Sometimes these issues are minimal, and ease of measurement is paramount. Still, the choice of easy or convenient metrics should be intentional rather than a default caused by ignorance of the potential issues.

Metric Features

The desiderata are difficult to balance, and the processes suggested can clarify goals and weaknesses. While considering design strategies, there are features of metrics that can allow for different and sometimes better trade-offs. The below list is not a full review of metric properties, but includes several general points about what can be done, and includes some critical suggestions relevant to avoiding perverse outcomes discussed earlier.

Data Sources. There are many places that can be used for understanding a system, and not all of them are immediately obvious to metric designers. For example, in medicine, administrative data can sometimes be as useful as clinical data for understanding risk, but does not need to be gathered separately .(Flacker & Kiely, 2003) Similarly, for web sites, user behavior can be gathered from site logs and used to infer issues, rather than fielding user surveys to ask about the experience.

Diversification. Often, no single metric can be found that both aligns well with the goal, and isn't manipulable. Introducing additional metrics, even if they are individually less well correlated to the goal, can sometimes improve the system overall. In a similar way, it is often the case that multiple different metrics are better aligned with the true goal than any single metric.

Aggregation Diverse and compound metrics can also be used to mitigate problems with incoherence, such as disagreement or lack of causal understanding. This is because a scattershot approach will tend to limit the degree to which any one measure influences the system. Designers with conflicting goals can choose measures that assist with each, and the combination may be an acceptable compromise. Similarly, if the causal relationships are unclear, targeting multiple different parts of the system may constrain the amount by which the system is changed due to the new incentives.

Secret Metrics. If the metric is not known to participants, they cannot game it. The existence of an un-revealed metric can still incentivise participants to achieve the goals they think most likely to be measured or rewarded, and to the extent that they understand the goal but not the metric, this will align incentives while preventing or at least hindering manipulation.

Post-Hoc Specification. If the metric is chosen after all actions are taken, participants view the metric as secret, but because the order of choices is reversed, attempted

gaming of the metric can be punished, or at least detected and ignored. Unfortunately, this can be perceived as allowing unfair discretion, and may lead to new forms of corruption in both the technical sense of invalidating the metric, and the typical sense of dishonest and fraundulent conduct by those choosing the metric.

Randomization. Even if a metric is known beforehand, if the specific components or the relative weights and rewards are uncertain, gaming the metric is harder and in expectation less rewarding. In addition, many forms of randomization can allow later evaluation of success via econometric methods, which is especially useful for monitoring the usefulness of the metric or reward system. Again, however, this reduces perceived fairness

Soft Metrics. Human judgment, peer evaluation, and other techniques may be able to reduce gaming specific to metrics. Metrics are often seen as a way to avoid subjectivity, but a combination of metrics and human judgment may be able to capture the best of both worlds.

Limiting Maximization. Failures are often the result of too much pressure on the optimization. By using metrics to set a standard or provide a limited incentive instead of a presenting value to maximize, the overoptimization pressure can sometimes be mitigated.

Abandoning Measurement. Sometimes, the value of better incentivising participants and the potential for perverse incentives issues make it worthwhile to be wary of what Muller refers to as metric fixation.(Muller, 2018) As he suggest, sometimes the best solution is to do nothing - or at least nothing involving measurement.

CONSIDERING APPLICATIONS AND FEATURES IN PRACTICE

Not all strategies are appropriate in all domains, and implementation is critically dependent on factors specific to a given system and the relevant actors. Still, systems chosen by public authorities face a higher burden for fairness and non-corruptibility, while those implemented in private business often require more immediacy. Incentives intended to motivate non-experts benefit more if they are simpler and easily understood, and those that impact people or organizations which must participate in a system, such as employees, or those that involve high reward, may need to be more game-proof.

The different issues that are implicated necessitate a broader discussion of some of the complex trade-offs. The variety of concerns that exist, however, make it worthwhile to illustrate the relationship between the metric desiderata and the process of designing better metrics, and how the different specific metric strategies will affect the desiderata. The table below attempts to do this briefly, followed by a discussion of how desiderata can differ based on more specific context.

	Availability	Cost	Immediacy	Simplicity	Fairness	Non-corruptibility
Considering Coherence		+		#	+	+
Causal Analysis		-		-		
Structured Compromise				-	+	
Pre-Gaming		-			+	+
Monitoring Behavior	-	-	-	#	+	+
Diversification		-		-	+	
Aggregation				-	+	+
Secret Metrics			-		#	-
Post-Hoc Specification	+	+	-			-
Randomization			#	-	-	+
Soft Metrics	+	-	#	#		
Limiting Maximization				-	+	
Abandoning Measurement			+	+	-	-

The table indicates which desiderata (top) are likely affected by first, each strategy and second, each metric characteristic. Positive effects on each desideratum are indicated with a plus, while negative ones are indicated with a minus. Complex interactions are complex are noted with a hash, as these are sometimes positive and sometimes negative. These are discussed in more detail below.

There are many examples of considering these and related desiderata. Reviewing a few recent, exemplary examples allows us to highlight how context-specific features lead to desiderata and approaches that are unique to that context.

Fraade-Blanar et al's "Measuring Automated Vehicle Safety," which usefully distinguishes between "measures (concepts)," which can be thought of as soft metrics, and "metrics (a defined calculation)". In this framing, they note that measures can be leading or lagging, so that the leading measures are indications, typically without a clear causal relationship with the goal, which "serve as proxies or surrogates for lagging measures," which may come too late, but can be more precise and causally connected to the goal. They suggest that the measures should be valid, reliable, feasible (low-cost,) and non-manipulatable (non-corruptible.) Fraade-Blanar2018 Because they focus on leading indicators, the discussion of validity drops their earlier and critical discussion of how measures should have causal, in this case physics-based, relationships with the phenomenon of interest. Reliability is important in their context because the metrics is used across all vehicles and vehicle types, and measures may differ in their validity ore usefulness between vehicles.

O'Keefe et al's Windfall Clause discusses designing a quantifiable future trigger for companies that capture large windfall profits from being the first to invent general artificial intelligence, and consider desiderata that include transparency, elasticity and adequacy (fairness,) and a number of less generally applicable desiderata. (O'Keefe et al., 2019) The less applicable desiderata here are interesting because of the speculative

nature of the metrics - there is no way to validate them before the potential one-time event they are supposed to influence.

Development Impact Bonds are an application of metrics that faces many challenges due to being directly financially incentified. In addition, they need a metrics specified in advance which resolves quickly at the time of the bond maturity, so that groups designing such bonds must be very careful. Sturla, Shah, and McManus present a very useful summary of the lessons learned by IDInsight in this domain. First, they need to measure carefully, using "outcomes that: 1) capture real improvements in people's lives, 2) can be measured, and 3) hold up under pressure." Second, the impact must be accurately and convincingly attributed, implicating both trust and transparency. In this case, attribution also requires careful understanding of the causal basis of the measurement. Third, the goals need to allow for discretion in implementation, and allow adaptation during the process so that innovation is possible. Fourth and finally, design needs to carefully consider trade-offs, especially because these bonds are ideally designed so that the measurement can be done at low cost. (Sturla, Shah, & McManus, 2018)

Given these concrete examples, it is now worth considering what things should be considered for building and calculating metrics, and see how they can help.

Data Sources

New or unexploited sources of data can be very valuable. Often, new sources are marginally the most valuable sources for metrics because they provide novel insights(Hubbard, 2007). At the very least, the novelty is itself can temporarily forestall cheating and gaming the metrics. At the same time, new instruments and data sources will have new and unforeseen challenges, and the ways in which they fail can be far less obvious.

Diversification

When goals are complex but cannot be directly measured, measures of various components or correlated outcomes can be used. This may make the goal easier to achieve, since it replaces an unclear target with clear sub-targets, but it may also make it harder for participants to decide what they should focus on. This means that gaming of metrics will be harder, but each additional metric creates the need to identify how it can be gamed, and how to prevent that gaming.

When a metric includes only some parts of a goal, it implicitly pushes emphasis away from the others, and diversified metrics can mitigate this issue. If reading and arithmetic are each 50% of the measured outcomes from school, it means that science, art, and physical education are all 0%. Because the easy to measure parts of a system are quickly accounted for and optimized for, even rudimentary or obviously biased measures of the remaining outcomes can offer significant marginal value. (Hubbard, 2007). Measuring additional features therefore removes the implicit pressure to minimize the previously unmeasured parts of the goal. For example, adding measures of time spent in arts classes will at least mitigate the pressure to remove those classes completely - and by doing so, lose important longer term benefits that are more difficult to measure for short-term

evaluation(Hess, 2018). Note, however, that simply adding metrics may not be wise, especially if they are all capable of being exploited in the same way. For example, testing students on various subjects to diversify metrics for learning instead of focusing on just mathematics and language does nothing to prevent metric failure if students cheat on tests. In addition, it may aggravate issues with losing class time due to testing, and moving focus from learning to teaching to the test.

Aggregation

Metrics which amalgamate multiple simple measures are often useful when individual measures are insufficient, as noted in the discussion on diversification. Recalling an example above, the choice of the best basketball players is better predicted by a combination of metrics than any single one. As noted earlier, aggregation can be used to side-step issues with finding a consensus for a single metric, and are also useful when the causal relationships are unclear. In either of these cases, however, the metrics are unlikely to be coherent. Still, because the different metrics typically require different behaviors, and they will be to some extent in tension with one another, they can make gaming harder. The complexity of aggregate metrics can sometimes reduce the degree to which participants can game metrics, but simultaneously make it harder for the designers to identify ways that participants may find to game the system.

Note that diversification and aggregation can be complementary, but diversification does not require a single aggregate metric. In fact, disaggregated metrics can identify and prevent problems caused by Simpson's paradox. Comparing subgroup outcomes directly can reduce the incoherence of comparing implicitly aggregated overall outcomes, which is sometimes important. For example, Leibowitz and Kelley show examples where different sub-population sizes can make ranked education outcomes reverse direction markedly. Once the success of subgroups is considered, diverse areas which perform worse in aggregate are found to better serve every sub-population, making the aggregate metric for success not only incomplete, but incoherent.(Liebowitz & Kelly, 2018)

Unfortunately, keeping metrics disaggregated can make it hard to compare or incentivise results, and any method of combining conflicting or varied measures will make the overall system more complex, or incoherent. Such complex and incoherent metrics may also be less effective at motivating desired behavior, since the complexity that makes gaming less likely makes it harder for participants to identify how to target the compound metric at all.

Secret Metrics

When qualitative goals are understood, keeping participants from knowing the details of the measurement system will limit the degree to which they can exploit the system. This requires some conception of the goal independent of the metric. In the worst case, the awardees don't understand the goal at all, and they will not be motivated by the seemingly-arbitrary rewards.

This is an effective strategy for preventing gaming, especially when pre-gaming methods discover important vulnerabilities of the various metrics that are hard to avoid. This works well if the metrics can be gathered without informing participants, and where the metrics that would be used are not obvious. The strategy will be less effective at preventing gaming if they can guess or infer the metric which will be used. Similarly, if the data collection to support evaulation of the metric is visible to participants, such as requiring them to take a test or gather specific data, it will be harder to hide.

Unfortunately, secrecy is prone to degrade over time as rewards are received and people can infer what is being evaluated. If a metric must be used repeatedly or in real time, it will be difficult to keep participants unaware of the details of the system. Similarly, if managers or regulators who implement the system are themselves being judged on the basis of the measured results, or they can be induced by participants to divulge information, they may intentionally degrade the secrecy needed. For this reason, secret metrics are more helpful if used one time then changed, as occurs when new tests are written for students each year - and as that case illustrates, knowledge of the types of questions commonly asked can still confer unfair advantages.

Post-Hoc Specification

When results are seen and analyzed before the metric is chosen, there are a variety of ways to prevent gaming while preserving the transparency of the rewards.

Designing measures completely post-hoc often involves justifying intuition or decisions already made. To avoid this, post-hoc specification should be limited to only include some parts of the metric. For example, the weights on various measures may be chosen after all activities have finished, or certain measures may be discarded based on analysis of the outcomes. If this process is known to participants beforehand, the potential for metrics to be discarded or given low weights can serve as an incentive not to game them.

The first, and most significant disadvantage for such post-hoc decisions is unfairness, both actual and perceived. Transparency in the process for the post-hoc selection can mitigate this problem, as can ensuring that the decision is made by a party that is not directly involved. The second significant disadvantage is that the feedback and reinforcement is delayed, which can significantly reduce the effectiveness of a reward system. A key advantage is that the post-hoc specification can keep the measures simple and easy to understand.

Randomization

Randomization can be used to choose between different proposed metrics when there is disagreement, or can be used within the metric itself. Choosing metrics via chance may avoid difficult compromise that leads to incoherent results. Allowing part of a metric or incentive to be determined by chance can be useful for preventing exploitation. Like secrecy and post-hoc specification, randomization reduces the direct connection between behaviors and metrics, which has some of the same positive and negative impacts.

To the extent that the weights and rewards are randomized instead of chosen intentionally, the incentives will be less well aligned with the actual goal. The uncertainty may also be perceived as adding significant and hard to understand complexity, and reduce motivation to achieve goals. On the other hand, exploitation is similarly less rewarding. Randomization can also be perceived as unfair, either because it rewards individuals differently, or because it rewards factors in a way not proportionate to importance.

Randomization works particularly well in combination with other methods. For instance, the randomization of the outcomes of a metric based on diverse inputs can assign random weights to already-known components. Similarly, it can be used to remove concerns about corruption for post-hoc specification, by pre-specifying the randomization to be performed at the end of a time period. If used beforehand to assign different metrics or different weights on metrics to different groups, it can also be valuable for analyzing the outcomes from using various metrics and incentive systems.

Soft Metrics

Metrics can include quantitative evaluations of more subjective factors that require data collection. These soft metrics are often able to avoid certain pitfalls of focusing on quantifying extant data. For example, peer ratings by programmers will not reward behaviors that help achieve measurable results like rapid but sloppy development at a high cost to the overall goals and maintainability of a system. Such measures have their own potential for exploitation, where participants game the system via currying favor, "sucking up," or taking measures to appear more productive than the reality.

Such data gathering can be done routinely, which has the advantage of providing feedback rapidly, but if participants need to routinely spend otherwise productive time doing evaluations, the cost of such measurement can be very high. They can also be perceived as unfair, and this can also lead to fighting or backstabbing - especially if the rewards are zero-sum.

Limiting Maximization

Metrics do not need to be maximized to be effective. If the metrics is used as a minimum for some incentive, the overoptimization may dissapear. By replacing optimization with what Simon terms satisficing(Simon, 1947), many of these problems can be avoided. For example, bonuses for salespeople who hit sales number targets is less likely to lead to overly competitive employee dynamics, where employees try to "steal" credit, or aleinate customers with overly aggressive tactics.

This strategy is not always appropriate, and using metrics in this way will not completely avoid the issues of participants gaming metrics, nor will it necessarily eliminate the pressure to perform well, or the stigma of performing poorly. Further, metrics are also often abused for control and direct feedback.

Steven Shorrock noted that "when you put a limit on a measure, if that measure relates to efficiency, the limit will be used as a target."(Shorrock, 2019) His original example was of flight duty times, where a regulation limiting the maximum number

of duty-hours that airlines crews can work led to use of that minimum as a target for airlines. Now that they must measure crew-duty times, airlines try to ensure their employees are as close to the limit as is possible. By introducing this new measure, it is possible crews are now more overworked than they were before any measurement of duty-hours was required.

Once metrics exist, they will often be abused for control even when inappropriate. The UK has a "Year 1 Phonics Check" in schools, which was developed and has been found useful for diagnosing "at-risk-readers" (Duff, Mengoni, Bailey, & Snowling, 2015). From the proposal's idea of diagnostic testing, it quickly turned into an "accountability agenda" almost completely useless to supplement extant assessment procedures, but very valuable for grading teachers and schools success at teaching reading. (Bradbury, 2014)

Satisficing can also allow complacency once targets are reached. Climate legislation limiting total emissions have failed because they were not ambitious enough, and "the shortcomings identified... are inherent to crediting mechanisms in general" (Cames et al., 2016). That report found, as one important shortcoming, that transferrable emissions credits were worthless in part because there were too many credits that were being generated effectively for free. This was made worse because of the ability to transfer the credits from countries that exceeded the goal to places where the goal was not met. Because no further incentive was in place once targets were met, there was no need to embark on more ambitious projects. In such a case, structuring the incentive differently might have been more effective. For instance, a moderately-sized tax on emissions could provide incentive to do some amount of mitigation without providing a potentially unlimited incentive to artificially game the system the way refundable tax credits might.

Abandoning Metrics, or Using them Diagnostically

Despite their general usefullness, metrics are sometime bad. For instance, situations where measuring outcomes is too expensive to be justified by the potential improvement that it could create. This often occurs when the complexity needed to correctly represent the system can require a business structure that is unreasonably or inefficiently complex(Poulis & Poulis, 2016). In other cases, them metric is likely to lead to distorted incentives rather than the initial goal. Many note that what isn't measured isn't managed, and the aphorism is correct. Still, when choosing between not managing part of a system by not measuring it, or measuring it in a way that makes it worse, the choice should be clear.

The negative impacts of poorly designed metrics are felt by multiple parties, not only those who the metrics are intended to help. Obviously, the people who promote the metrics would prefer if their actual goal were pursued, rather than chasing the metric. Anyone who attempts to target the ultimate goals of the system and ignore the perverse incentives are implicitly punished for not playing these games. They would prefer better metrics that reward their efforts, or no metrics, so they are not punished. The people who do adopt strategies to exploit the perverse incentives may benefit directly, but even they would often be happier not to be forced to play the game of understanding and

exploiting complex, changing, and often harmful systems. Their exploitation of metrics also has impacts well beyond the management and the players of these games, since the economic waste and negative externalities created by exploiting poorly designed metrics can be significant.

The Gravity of Abandoning Metrics, and the Alternative

Choosing not to manage a system is a decision that should not be made lightly - especially not before seriously considering whether an alternative measurement might be useful. On the other hand, putting in place a mediocre measurement system prematurely is often far worse. Until serious consideration has been given to the processes and alternatives identified above, it may be better to wait, or to abandon incentives based on measurement, rather than deploy a system that will be ineffective or worse. As Muller puts it, "sometimes, recognizing the limits of the possible is the beginning of wisdom. Not all problems are soluble, and even fewer are soluble by metrics."(Muller, 2018)

These limits Muller notes are particularly relevant if participants will be drawn to the explicit rewards that are less well suited to accomplishing the goal than those who would participate regardless. The limitations are also critical if participants feel discouraged by the extrinsic motivation and measurement, especially in domains where intrinsic motivation is primary. This is supported by the empirical work by Rasul et al. showing that autonomy, which is incompatible with extensive measurement and accountability systems, is more effective for civil service. (Rasul & Rogger, 2017; Rasul, Rogger, & Williams, 2017, 2018)

However, it is critical not to throw out the measurement baby with the perverse incentives bathwater. In most cases, metrics can be used as a feedback mechanism, rather than using them for any direct reward system, or abandoning them completely. This approach is particularly useful when qualitative feedback and supervision are useful. For example, instead of using metrics to determine who gets a year-end bonus, the same measures of performance might be used to identify which people are excelling and which are falling behind so that the former can mentor the latter.

Transitioning to monitoring via measurement is also very useful if the diagnostic measures cannot identify what is failing, or are known to be causally unconnected to the goal. Identification of an issue can be useful without diagnosis, much like noise from a car engine is (usually) of limited value in diagnosing a problem, but of immense value in noticing that some such problem exists. In systems that are poorly understood quantitatively, diagnosing issues might require intensive investigation and intervention, but some numeric measures can provide early warning of a problem, and often they are valuable in doing so.

Still, as discussed above when considering limiting maximization, there are a number of problems that occur simply because measurements or concrete criteria are introduced. In addition to the above concerns, the use of quantifiable guideposts adds new failure modes. For example, these can be used to make claims unrelated to the purpose of the measurement, as in the earlier example of Phonics testing, or can be used as a way to accomplish other goals, sometimes undermining the purpose of the diagnostic measure.

As an example of how diagnostic measures intended for evaluation can be misused, consider diagnostic criteria in mental health. Used properly, criteria are interpreted with a careful view to contextual factors, the presence or absence of extrinsic causes, the existence or non-existence of a functional impairment, and so on. These diagnostic criteria are intended to be flexible, and provide insight and assistance for clinical work. "A too-rigid categorical system does not capture clinical experience," but it is all-too-easy for non-experts (or experts) to use the diagnostic criteria far outside of what the criteria writers intended. (APA (American Psychiatric Association), 2013) In extrema, this leads to "amateur, at-a-distance diagnosticians," applying such criteria as a political statement, rather than for diagnosis or treatment. (Frances, 2017)

The same domain also illustrates the abuse of diagnostic criteria to accomplish other goals. Mental health diagnoses are used by American insurance companies to determine whether to reimburse treatments, or how much to pay for a given service. In doing so, assessments can be turned into games played by clinicians (or their billing departments) to enable individuals to get needed care. This turns diagnostic measures back into metrics, with all of the accompanying failure modes. For example, diagnostic accuracy may be replaced with practical concerns. An insurance provider may not pay for counseling or medication in the case of a generalized anxiety disorder, but the service or medication is covered if the patient is instead diagnosed with panic disorder. If a patient cannot otherwise afford care, the temptation for providers to slightly modify patient diagnoses may be overwhelming.

TOWARDS A COHERENT PROCESS FOR METRIC DESIGN

Given the various strategies and considerations discussed in the paper, as well as failure modes and limitations, it is useful to lay out a simple and coherent outline of a process for metric design. While this will by necessity be far from complete, and will include items that may not be relevant for a particular application, it should provide at least an outline that can be adapted to various metric design processes. Outside of the specific issues discussed earlier, there is a wide breadth of expertise and understanding that may be needed for metric design. Citations in this section will also provide a variety of resources for at least introductory further reading on those topics.

1. Understand the system being measured, including both technical (Blanchard & Fabrycky, 1990) and organizational(Berry & Houston, 1993) considerations.

    - Determine scope;
      What is included in the system?
      What will the metrics be used for?
    - Understand the causal structure of the system;
      What is the logic model or theory?(Rogers, Petrosino, Huebner, & Hacsi, 2000)
      Is there formal analysis(Gelman, 2010) or expert opinion(van Gelder, Vodicka, & Armstrong, 2016) that can inform this?

- Identify stakeholders(Kenny, 2014);
  Who will be affected?
  Who will use the metrics?
  Whose goals are relevant?

2. Identify the Goals

   - What immediate goals are being served by the metric(s)? How are individual impacts related to performance more broadly?(Ruch, 1994)
   - What longer term or broader goals are implicated?

3. Identify Relevant Desiderata

   - Availability
   - Cost
   - Immediacy
   - Simplicity
   - Transparency
   - Fairness
   - Corruptibility

4. Brainstorm potential metrics

   - What outcomes important to capture?
   - What data sources exist?
   - What methods can be used to capture additional data?
   - What measurements are easy to capture?
   - What is the relationship between the measurements and the outcomes?
   - What isn't captured by the metrics?

5. Consider and Plan

   - Understand why and how the metric is useful. (Manheim, 2018)
   - Consider how the metrics will be used to disagnose issues or incentify people. (Dai, Dietvorst, Tuckfield, Milkman, & Schweitzer, 2017)
   - Plan how to use the metrics to develop the system, avoiding the "reward / punish" dichotomy.(Wigert & Harter, 2017)
   - Perform a pre-mortem(Klein, 2007)

6. Plan to routinely revisit the metrics(Atkins, Wanick, & Wills, 2017)

CONCLUSION

Despite the intrinsic limitations of metrics, the frequent use of poorly thought-out and badly constructed metrics do not imply that metrics are doomed to eventually fail, or that they should not be used because they will be exploited. Instead, forethought and consideration of the problems with metrics is often worthwhile. This process starts by identifying and agreeing on coherent goals, then considering both what leads to the goals, and what parts of the system can be measured. After identifying measurable parts of the system, and considering how participant behavior might exploit the measurement methods or the measured outcomes, measures can be constructed. The construction of these metrics to avoid exploitation may involve multiple diverse measures, secret metrics, intentional reliance on post-hoc specification of details, and randomization. This may also include decisions about where subjective measurements are important, and consideration whether measurement will be beneficial. In building the metrics and deciding whether to implement them, attention should be paid to various important factors in the system, including immediacy of feedback, simplicity and understandability of the measurement system, fairness, and the potential for both actual and appearance of corruption in the metric and reward system.

Metric design is an engineering problem, and good solutions involve both science and art. Following these guidelines will not make metrics unexploitable, nor will it keep everyone happy with the results of a process. This is true of metrics used for employees, metrics used for monitoring systems, and even metrics used within machine learning algorithms - in each case, poorly designed metrics will be exploited. Occasionally, the suggested process will lead to investigation of potential improvements or strategies that are ultimately decided against. Despite this, it is a vast improvement on the too-common strategy of using whatever metric seems at first glance to be useful, or deploying metrics without considering what they in fact promote. Putting in the effort to build elegant and efficient solutions won't fix every problem, but it will lead to less flawed metrics and better results overall.

Acknowledgements

I am grateful to Abram Demski and Scott Garrabrant for prompting my initial interest in Goodhart's Law, Gregory Lewis for his work on the topic, and Venkatesh Rao for allowing me to explore the ideas on his "Ribbonfarm" blog in a series of guest posts. Conversations with other Ribbonfarm authors and readers, as well as twitter conversations about these ideas have been very helpful. I am especially grateful for the twitter-based insights provided by (in no particular order,) Steve Shorrock, Sam Gardner, James Pitt, Simon DeDeo, Thomas Dullien, Scot Hazeu, Alex Holcombe, Tiago Forte, Daniel Bilar, Nick Szabo, Peter Schryvers, and I am sure others who I am unfortunately omitting. I'd also like to thank Paul Davis for insight into complex policy systems generally, and useful feedback on the paper, as well as Osonde Osoba for early conversations and feedback about how randomization and secrecy would impact metrics. Lastly, I would like to thank the Machine Intelligence Research Institute for prompting, and the Berkeley Existential Risk Initiative for funding, my work on metric-alignment for artificial intelligence in

multi-agent systems, which has spurred much of my thinking on how the same dynamics affect human systems.

REFERENCES

APA (American Psychiatric Association). (2013). Diagnostic and statistical manual of mental disorders. BMC Med, 17, 133–137.

Atkins, A., Wanick, V., & Wills, G. (2017). Metrics Feedback Cycle: measuring and improving user engagement in gamified eLearning systems. International Journal of Serious Games, 4(4), 3–19.

Berry, L. M., & Houston, J. P. (1993). Psychology at work: An introduction to industrial and organizational psychology. Brown & Benchmark/Wm. C. Brown Publ.

Blanchard, B. S., & Fabrycky, W. J. (1990). Systems engineering and analysis (4th ed.). Prentice Hall Englewood Cliffs, NJ.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological review, 111(4), 1061.

Bradbury, A. (2014, sep). 'Slimmed down' assessment or increased accountability? Teachers, elections and UK government assessment policy. Oxford Review of Education, 40(5), 610–627. Retrieved from https://doi.org/10.1080/03054985.2014.963038 doi: 10.1080/03054985.2014.963038

Cames, M., Harthan, R. O., Füssler, J., Lazarus, M., Lee, C., Erickson, P., & Spalding-Fecher, R. (2016). How additional is the clean development mechanism. Analysis of application of current tools and proposed alternatives. Oeko-Institut EV CLlMA. B, 3.

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and program planning, 2(1), 67–90.

Caplan, B. (2018). The case against education. Why the education system is a waste of time and money. Princeton University Press.

Caudill, H. L., & Porter, C. D. (2014, dec). An Historical Perspective of Reward Systems: Lessons Learned from the Scientific Management Era. International Journal of Human Resource Studies; Vol 4, No 4 (2014)DO - 10.5296/ijhrs.v4i4.6605. Retrieved from http://www.macrothink.org/journal/index.php/ijhrs/article/view/6605

Choi, J., Hecht, G. W., & Tayler, W. B. (2012). Lost in translation: The effects of incentive compensation on strategy surrogation. The Accounting Review, 87(4), 1135–1163.

Clifton, P. M., & Keogh, J. B. (2017). A systematic review of the effect of dietary saturated and polyunsaturated fat on heart disease. Nutrition, Metabolism and Cardiovascular Diseases, 27(12), 1060–1080.

Dai, H., Dietvorst, B. J., Tuckfield, B., Milkman, K. L., & Schweitzer, M. E. (2017, aug). Quitting When the Going Gets Tough: A Downside of High Performance Expectations. Academy of Management Journal, 61(5), 1667–1691. Retrieved from https://doi.org/10.5465/amj.2014.1045 doi: 10.5465/amj.2014.1045

Deresiewicz, W. (2015). Excellent sheep: The miseducation of the American elite and the way to a meaningful life. Free Press.

Duff, F. J., Mengoni, S. E., Bailey, A. M., & Snowling, M. J. (2015). Validity and sensitivity of the phonics screening check: implications for practice. Journal of Research in Reading, 38(2), 109–123.

Faeh, D., Paccaud, F., Cornuz, J., & Chiolero, A. (2008, apr). Consequences of smoking for body weight, body fat distribution, and insulin resistance. The American Journal of Clinical Nutrition, 87(4), 801–809. Retrieved from https://dx.doi.org/10.1093/ajcn/87.4.801 doi: 10.1093/ajcn/87.4.801

Flacker, J. M., & Kiely, D. K. (2003). Mortality-related factors and 1-year survival in nursing home residents. Journal of the American Geriatrics Society, 51(2), 213–221.

Fraade-Blanar, L., Blumenthal, M. S., Anderson, J. M., & Kalra, N. (2018). Measuring Automated Vehicle Safety.

Frances, A. (2017). Trump isn't crazy. Psychology Today. Retrieved from https://www.psychologytoday. com/blog/saving-normal/201701/trump-isnt-crazy.

Gelman, A. (2010). Causality and Statistical Learning. American Journal of Sociology, 117(3), 955–966. Retrieved from http://arxiv.org/abs/1003.2619 doi: 10.1086/662659

Goodhart, C. A. E. (1975). Problems of monetary management: the UK experience. In Papers in monetary economics. Reserve Bank of Australia.

Herzberg, F. (1968). One more time: How do you motivate employees. Harvard Business Review Boston, MA.

Hess, F. (2018, sep). Straight Up Conversation: Scholar Jay Greene on the Importance of Field Trips. Education Week. Retrieved from https://blogs.edweek.org/edweek/rick{_}hess{_}straight{_}up/2018/09/straight{_}up{_}conversation{_}scholar{_}jay{_}greene{_}on{_}the{_}importance{_}of{_}field{_}trips.html

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. JL Econ. & Org., 7, 24.

Hubbard, D. W. (2007). How to Measure Anything: Finding the Value of Intangibles in Business (Second ed.). doi: 10.1002/9781118983836

Kalra, N., Hallegatte, S., Lempert, R., Brown, C., Fozzard, A., Gill, S., & Shah, A. (2014). Agreeing on Robust Decisions New Processes for Decision Making Under Deep Uncertainty. World Bank Policy Research Working Paper, No. 6906(June). doi: doi:10.1596/1813-9450-6906

Kenny, G. (2014). Five questions to identify key stakeholders. HBR Harvard Business Review.

Klein, G. (2007). Performing a project premortem. Harvard Business Review, 85(9), 18–19.

Klein, G., Sonkin, P. D., & Johnson, P. (2019). Rendering a Powerful Tool Flaccid: The Misuse of Premortems on Wall Street.

Lempert, R. J., Groves, D. G., Popper, S. W., & Bankes, S. C. (2006). A General, Ana-

lytic Method for Generating Robust Strategies and Narrative Scenarios. Management Science, 52(4), 514–528. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1050.0472 doi: 10.1287/mnsc.1050.0472

Liebowitz, S., & Kelly, M. L. (2018, nov). Everything You Know About State Education Rankings Is Wrong: Minds and dollars are a terrible thing to waste. Reason. Retrieved from https://reason.com/archives/2018/10/07/everything-you-know-about-stat

Liska, D. J., Cook, C. M., Wang, D. D., Gaine, P. C., & Baer, D. J. (2016). Trans fatty acids and cholesterol levels: An evidence map of the available science. Food and Chemical Toxicology, 98, 269–281.

Manheim, D. (2016). Overpowered Metrics Eat Underspecified Goals (Vol. 2016). Retrieved from https://www.ribbonfarm.com/2016/09/29/soft-bias-of-underspecified-goals/

Manheim, D. (2018). Value of Information for Policy Analysis (Doctoral dissertation, Pardee RAND). Retrieved from https://www.rand.org/pubs/rgs{_}dissertations/RGSD408.html

Manheim, D., & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. , 1–10.

Mika, E., & Lee, B. (2017). Who Goes Trump? Tyranny as a Triumph of Narcissism. St. Martin's Press.

Mitchell, D. J., Edward Russo, J., & Pennington, N. (1989). Back to the future: Temporal perspective in the explanation of events. Journal of Behavioral Decision Making, 2(1), 25–38. Retrieved from https://doi.org/10.1002/bdm.3960020103 doi: 10.1002/bdm.3960020103

Muller, J. Z. (2018). The tyranny of metrics. Princeton University Press.

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2019). The Windfall Clause: Distributing the Benefits of AI for the Common Good. arXiv preprint arXiv:1912.11595.

Poulis, K., & Poulis, E. (2016). Problematizing fit and survival: transforming the law of requisite variety through complexity misalignment. Academy of Management Review, 41(3), 503–527.

Rasul, I., & Rogger, D. (2017). Management of bureaucrats and public service delivery: Evidence from the nigerian civil service. The Economic Journal, 128(608), 413–446.

Rasul, I., Rogger, D., & Williams, M. (2017). Management and bureaucratic effectiveness: A scientific replication.

Rasul, I., Rogger, D., & Williams, M. J. (2018). Autonomy, incentives, and the effectiveness of bureaucrats. VoxDev. Retrieved from https://voxdev.org/topic/public-economics/autonomy-incentives-and-effectiveness-bureaucrats

Rodamar, J. (2017). There ought to be a law! Campbell v. Goodhart.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. New directions for evaluation, 2000(87), 5–13.

Rosenhead, J., & Mingers, J. (2001). Rational analysis for a problematic world revisited (No. 2nd). John Wiley and Sons.

Ruch, W. A. (1994). Measuring and managing individual productivity. Organizational linkages: Understanding the productivity paradox, 105–130.

Saltelli, A. (2020). Ethics of quantification or quantification of ethics? Futures, 116, 102509. Retrieved from http://www.sciencedirect.com/science/article/pii/S0016328719303714 doi: https://doi.org/10.1016/j.futures.2019.102509

Schoeller, D. A. (1990). How accurate is self-reported dietary energy intake? Nutrition reviews, 48(10), 373–379.

Shorrock, S. (2019, may). Shorrock's Law of Limits. Blog Post. Retrieved from https://humanisticsystems.com/2019/10/24/shorrocks-law-of-limits/

Simon, H. A. (1947). Administrative behavior; a study of decision-making processes in administrative organization. Macmillan.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological review, 63(2), 129.

Soares, N. (2015). Half-assing it with everything you've got. Retrieved 2019-07-22, from http://mindingourway.com/half-assing-it-with-everything-youve-got/

Strathern, M. (1997). 'Improving ratings': audit in the British University system. European Review. doi: 10.1002/(SICI)1234-981X(199707)5:33.0.CO;2-4

Sturla, K., Shah, B., & McManus, J. (2018). The Great DIB-ate: Measurement for Development Impact Bonds. Stanford Social Innovation Review. Available at: https://ssir. org/articles ....

Szajewska, H., & Szajewski, T. (2016). Saturated fat controversy: importance of systematic reviews and meta-analyses. Critical reviews in food science and nutrition, 56(12), 1947–1951.

Taplin, D. H., & Clark, H. (2012). Theory of change basics: A primer on theory of change.

van Gelder, T., Vodicka, R., & Armstrong, N. (2016, sep). Augmenting Expert Elicitation with Structured Visual Deliberation. Asia & the Pacific Policy Studies, 3(3), 378–388. Retrieved from https://doi.org/10.1002/app5.145 doi: 10.1002/app5.145

Wigert, B., & Harter, J. (2017). Re-engineering performance management. Gallup. com. Viewed: March, 6, 2019.