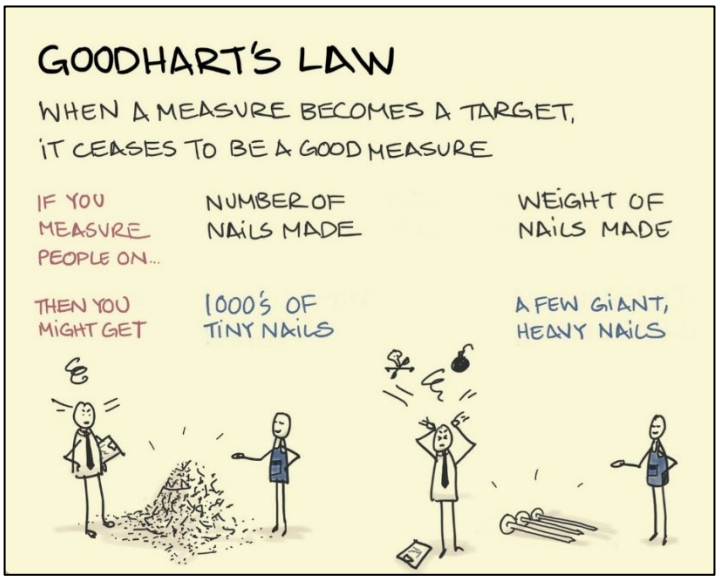
CNA

September 2022



Goodhart's Law: Recognizing and Mitigating the Manipulation of Measures in Analysis

Michael F. Stumborg, Timothy D. Blasius, Steven J. Full, Christine A. Hughes

With contributions from Jennifer M. Scherer

COP-2022-U-033385-Final

Abstract

Goodhart's Law states that "when a measure becomes a target, it ceases to be a good measure." In other words, when we use a measurement to reward performance, we have provided an incentive to manipulate the measurement in order to receive the reward. The result can sometimes be actions that actually damage the effectiveness of the measured system while paradoxically improving the measurement of system performance. This report provides examples of Goodhart's Law in defense analysis and offers analysts and the organizations that employ them techniques to identify and mitigate the pervasive and pernicious effects of measurement manipulation.

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.
Administrative or Operational Use 9/26/2022

This work was created in the performance of Federal Government Contract Number N00014-22-D-7001.

Cover image: https://sketchplanations.com/goodharts-law

Approved by: **September 2022**

Dr. Timothy A. Roberts
Director, Strike & Air Warfare Program
System, Tactics & Force Development

This document may contain materials protected by the Fair Use guidelines of Section 107 of the Copyright Act, for research purposes only. Any such content is copyrighted and not owned by CNA. All rights and credits go directly to content's rightful owner.

Request additional copies of this document through inquiries@cna.org.

Executive Summary

Goodhart's Law states that "when a measure becomes a target, it ceases to be a good measure." In other words, when we use a measure to reward performance, we provide an incentive to manipulate the measure in order to receive the reward. This can sometimes result in actions that actually reduce the effectiveness of the measured system while paradoxically improving the measurement of system performance.

Illustrative examples of Goodhart's Law include the experience of British officials in colonial India. They offered a bounty on cobra skins to reduce the cobra population, only to have entrepreneurial citizens actually breed cobras for their skins and then turn them loose when their fraud was discovered, thus increasing, rather than decreasing, the cobra population.

Examples of Goodhart's Law abound in defense analysis. They include manipulated measures of performance as diverse as body counts in Vietnam, modern day fighter aircraft readiness rates, and the use-it-or-lose-it "rule" in Department of Defense program management.

The manipulation of measures resulting from Goodhart's Law is pervasive because direct measures of effectiveness (MOEs), which are more difficult to manipulate, are also more difficult to measure, and sometimes simply impossible to define and quantify. As a result, analysts must often settle for measures of performance (MOPs) that correlate to the desired effect of the MOE. MOEs are difficult to measure and difficult to manipulate. MOPs are easier to measure, but also easier to manipulate. Thus, the negative impacts of Goodhart's Law are commonplace.

These negative effects can sometimes be avoided. When they cannot, they can be identified, mitigated, and even reversed. Analysts, and the organizations that employ them, often in partnership with the customer organizations that receive their analytical products, can take concrete actions to avoid or mitigate the negative effects of Goodhart's Law. This report recommends that analysts do the following:

- Use MOEs instead of MOPs whenever practicable and possible
- Use the scientific method to generate new measurement data, rather than harvesting existing and possibly compromised data
- Help customers establish authoritative and difficult-to-manipulate definitions for measures
- Identify and avoid the use of manipulated data and data prone to manipulation

- Use measurement data not generated by the organization being measured
- Collect data secretly or after a measurable activity has already occurred
- Measure all relevant system characteristics rather than just a representative few
- Randomize the measures used over time
- Wargame or red team potential measures

This report recommends that the organizations that employ analysts should do the following:

- Return to the roots of operational research to focus more on direct measurements in the field
- Answer the questions that *should be answered*, rather than the questions that *can be answered* simply because the required data are already available
- Train analysts on MOEs, MOPs, and Goodhart's Law and how they are interrelated
- Make recognition of Goodhart's Law part of the internal peer review process and part of all delivered analytical products
- Identify and share mitigation best practices

Organizations that use data and develop measures that have consequences—both positive and negative—for the persons, organizations, and processes they are charged with measuring and improving should act on these recommendations to identify, understand, avoid, mitigate, or reverse the effects of Goodhart's Law.

Implementing these recommendations will benefit individual analysts, the organizations that employ them, and the organizations that they support. Admittedly, these recommendations constitute additional burdens on project managers already stressed by limited budgets and tight schedules. Because the negative outcomes of not assuming these burdens will not occur until a future date, they may go unrecognized, or it may be tempting to dismiss them if recognized. Therefore, institutionalizing these additional actions by including them in an already required process (such as peer review) is essential to avoid the pervasive and pernicious effects of Goodhart's Law.

Contents

Introduction

Objective

Although this report seeks to advance the state-of-the-art in defense analysis—analysts from other disciplines can certainly use it to improve their craft. The target audience of this report is the analysts and analytical organizations that seek to improve either the effectiveness of military operations or the efficiency of the business processes used by these militaries to design, develop, field, and sustain their forces. The improvement sought is to inculcate within the defense analytical establishment an ability to avoid (or to identify and mitigate) instances where the accuracy of analytical results is diminished because the measures used to characterize the operation or process under study were manipulated by the effects of Goodhart's Law.

What is Goodhart's Law?

It is axiomatic that we cannot, without great difficulty, improve the performance of systems unless we can measure them [1]. The persons and organizations responsible for system performance often employ operations researchers and analysts of all types to carry out these measurements and to make recommendations on how to improve system performance based on what the measurements tell them about the system under study.

To improve performance, some measures must be maximized, some minimized, and others optimized to a specific value. For example:

- To improve readiness, a military service will ***maximize*** the percentage of military equipment that is in working order and available for combat.
- To improve profits, a corporation will ***minimize*** production costs.
- To improve crop yield, a farmer will ***optimize*** fertilizer use, adding enough to promote plant growth, but not so much that the fertilizer burns the root and kills the plant.

In these instances, the value of the measurement becomes *a proxy representation* for the performance of the measured system. The persons and organizations responsible for system performance therefore have a powerful incentive to move the measurement in the direction indicative of improved performance. They can do so by legitimate means—by actually improving the attributes of the system being measured—but they can also manipulate the

measure in a way that paradoxically improves the value of the measured attribute while failing to improve the performance of the underlying system. Such manipulation demonstrates Goodhart's Law, which states:

> *"When a measure becomes a target, it ceases to be a good measure."*

This quote is actually a popular rephrasing of a statement made by Charles Goodhart [3]. The original statement appeared in a 1975 paper examining the relationship between money supply and inflation in the United Kingdom [2]: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." Social scientist Donald Campbell similarly (and more clearly) stated: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" [3].

To achieve the objective of this report, we first present occurrences of Goodhart's Law from outside the defense analysis sector to demonstrate its widespread prevalence. We then present examples of Goodhart's Law within defense analysis to demonstrate the significant costs that militaries may incur if they are not aware of its negative effects.

We then discuss measures of effectiveness (MOEs) and measures of performance (MOPs) because choosing these measures incorrectly—or being forced by circumstance to use MOPs over MOEs—significantly affects the analyst's ability to avoid, or identify and mitigate, the negative effects of Goodhart's Law.

We conclude this report with recommendations—some directed at individual analysts (and teams of analysts) and others directed at the analytical organizations that employ them. Some recommendations from each of these categories will best be implemented in partnership with the military organizations that leverage defense analysis to improve their military operations and their supporting business processes.

Goodhart's Law in Action

The desire to measure human and organizational behaviors in order to improve performance is so prevalent that examples of Goodhart's Law can be found in most, if not all, fields of human endeavor.[1] The following is therefore just a representative sampling of Goodhart's Law in action, assembled to illustrate its pervasive nature.

Pest eradication

Seeking to reduce the population of cobras in colonial India, British authorities put a bounty on cobra skins. Entrepreneurial citizens began to breed cobras for the money. When the authorities discovered this and ended the program, cobra breeders released their now-worthless cobras, and their population in the wild actually increased [3]. French colonial authorities in Indochina had a similar experience with rats in Hanoi. Needing only the tails for proof of a kill, enterprising bounty hunters would not actually kill the rats, but would amputate the live rats' tails, leaving them to breed additional rats with valuable tails. Rats were also bred in the countryside and smuggled into Hanoi [4]. The US Army attempted to eradicate feral hogs from the grounds of Fort Benning, Georgia, asking only for the tails as proof of a kill, with equally disastrous results [5].

Academic testing

The No Child Left Behind Act sought to improve the academic achievement of disadvantaged students [6]. To achieve this goal, it created a dependence on standardized testing to measure the "adequate yearly progress" of students. Test scores were in turn used to assess the performance of teachers and schools, with implications for student promotion, teacher retention, and school closures. This led school teachers and administrators to cheat by providing test answers or by changing poor test scores [7]. Additionally, although it cannot technically be considered cheating, students were also "taught to pass the test" [8] instead of being taught to master all of the material required to have a well-rounded education. Evidence suggests that this act actually retarded the yearly progress of gifted students [9] simply because they required no additional instruction to pass the now-paramount annual

[1] Entering the terms "Goodhart's Law," "Campbell's Law," "the cobra effect," "gaming the system," or "the law of unintended consequences" into any internet search engine will return multiple additional examples of measure manipulation.

standardized test. They were left under-instructed and unchallenged by teachers and administrators who shifted their attention away from these students who they already knew would pass the test. Gifted students from all backgrounds were left behind by the No Child Left Behind Act.

Search engine optimization

Internet search engines use algorithms to rank the "relevance" of web pages and display the highest-ranked pages first. Because a high page rank creates more page views, web masters seek to get their pages ranked as highly as possible. The entire search engine optimization industry (and occupation) was created by this desire to manipulate the measurement of web page relevance. This manipulation sometimes takes the form of reverse engineering the page-ranking algorithms [10]. Manipulation tactics also include keyword spamming, generating massive numbers of low-quality pages, creating artificial link networks, and creating deceptive web pages that appear differently to users and search engines [11]. As a result, page-ranking algorithms are by necessity constantly updated and closely guarded proprietary secrets [12].

Package and mail deliveries

The US Postal Service has a contract with Amazon to deliver packages. Contract renewal is contingent on their on-time delivery rates. The delivery time is measured by scanning the packages at the time of delivery. However, mail carriers in Atlanta state that they were told to "pull your truck over to the side of the road and scan every single one of your [still-undelivered] Amazon packages" at 7:15 p.m. so they are not counted as late deliveries [13]. The US Postal Service also contracts airlines for on-time delivery of overseas mail. Three airlines recently paid over $80 million in fines for falsifying on-time mail delivery documents [14].

Coronavirus deaths

Determining whether a patient "with" a coronavirus infection died "of" that infection depends on the clinical discretion of the physician completing the death certificate [15]. Political incentives sway that determination, with accusations of both fraudulent overcounting [16] and undercounting [17]. In at least one instance, a financial incentive exists to falsify these records: a Louisiana county coroner noted that family members of deceased patients who argued against citing coronavirus as a cause of death sometimes reversed course when told that the federal government pays burial expenses for coronavirus victims [17].

Health care providers

Heart and organ transplant surgeons who are evaluated based on the survivability rate of their patients may not operate on seriously ill patients who are more likely to die [8, 18]. Similarly, fertility clinics that are evaluated on the rate of successful pregnancies may decline to take the hardest cases [19] and, in extreme examples, may even take patients with no fertility problems just to boost their pregnancy success rates.

Airline schedules

Airlines began being rated based on their on-time arrival rates in the 1980s. As a result, they simply increased the estimate of their flight times, artificially driving up their "on-time" rates [8].

Agile software development

All of the previous examples contain an element of corruption, but a person need not be corrupt to be corrupted by Goodhart's Law [20]. In agile software development, teams estimate point values for proposed work modules and are (sometimes) assessed based on points completed. A team that consistently "underperforms" compared to others based on this system may *honestly* ask themselves if they are undervaluing the degree of difficulty of their own work when they assign point values. They may then assign more points to proposed future work to bring their (measured) performance into line with expectations and norms.

We say "sometimes assessed" because agile software development teams are run by scrum masters, who use these point assignments with varying levels of formality. At one extreme, points completed can serve as an input to formal compensation and promotion decisions. At the other extreme, points are used by the scrum master only to schedule work assignments and track progress. Between these two extremes, public disclosure of points completed during team meetings can serve as a team motivational tool, or as a subtle form of peer pressure. The potential for recourse to Goodhart's Law depends on the consequences associated with scrum points.

Social credit scores

Though most measurements of human behavior seek to incentivize and reinforce desirable behavior, some measurements may seek to punish behaviors deemed undesirable by the measuring entity. Social credit systems such as those used in China fall into this category [21]. These systems are not limited to authoritarian governments [21], or even to governments.

The US company AirBnB patented what can be considered a social credit scoring system [22] that calculates a trustworthiness score of a person based on an analysis of online and offline documents about, or authored by, that person. The AirBnB "dictionary of behavior or personality traits" includes "badness, anti-social tendencies, goodness, conscientiousness, openness, extraversion, agreeableness, neuroticism, narcissism, Machiavellianism, and psychopathy." Though the correlation with trustworthiness is intuitively obvious for some of these terms (psychopathy), the correlation with others (extraversion) is less so.

"Untrustworthy persons" are presumably banned from the AirBnB platform. Some individuals deliberately and openly employ Goodhart's Law to counter this sort of real or perceived discrimination, noting that "these systems are flawed, and we should do everything we can to game them. After all, hacking is just lying to the machine, and in this world, it is often our moral duty to do so" [23].

Goodhart's Law in Defense Analysis

Goodhart and Campbell studied measure manipulation in economics and social programs, respectively, and the previous section provides examples from other disciplines. We should not be surprised, then, to find examples of measure manipulation in the defense sector as well.

Body counts in Vietnam

Military leaders fighting the insurgency in Vietnam could not use "territory liberated" as a measure of military progress as they had done in previous wars. Then-defense secretary Robert McNamara's emphasis on quantitative measures led to the counting of enemy soldiers killed in action—body counts—as one measure of progress. Higher body counts equated to greater progress in this paradigm, and military leaders were graded harshly on achieving progress in this unpopular war. According to retired LTG Robert Gard Jr., who served in Vietnam: "If body count is your measure of success, then there's a tendency to count every body as an enemy soldier" [24]. Evidence of body count inflation exists in mismatches between "enemy soldiers" killed or captured and enemy weapons recovered. For example, the US Army 9th Infantry Division recorded 10,899 enemy killed, 2,579 enemy captured, but only 748 recovered weapons [24].

Fighter aircraft readiness

In September 2018, then-defense secretary James Mattis directed the Navy and Air Force to increase the mission-capable rates for fighter aircraft to 80 percent [25-26]. Although the Navy declared success before the September 2019 deadline [27], the Air Force never did [28].

An analysis by the Government Accountability Office (GAO) questioned the validity of the Navy analysis, noting that the Navy tracks aircraft readiness using two separate databases—one that records daily aircraft status at a certain time each day—the Aviation Maintenance Supply Readiness Report (AMSRR), and another that records the percentage of the total time an aircraft is available—DECision Knowledge Programming for Logistics Analysis and Technical Evaluation (DECKPLATE) [29]. Even though Navy officials acknowledged that DECKPLATE data provide a "more comprehensive measure" of aircraft health, they used AMSRR data, which showed that three aircraft types met the 80 percent goal, while only one did using DECKPLATE data [29].

Individual Marine Corps aviation squadrons also counted mission-capable aircraft differently. Some used a prior 30-day average, others used a specific point in time, and still others used a 7-day forecast [30]. In these examples of Goodhart's Law, behavior is not modified to manipulate the measure, but rather the definition of the measure itself is manipulated.

Collapse of the Afghan national defense and security forces

The growth and development of the Afghan national defense and security forces became a precondition for reducing the US presence in Afghanistan [31] and, thus, had to be measured. In a manner similar to the previous description of how unscrupulous entrepreneurs leveraged the bounty on cobras [3] and rats [4] by breeding them, corrupt Afghan officials created "ghost soldiers" [32], who existed only on paper, to collect their paychecks. Though the inflated and fictitious number of actual soldiers is not likely the sole cause of the Afghan Army's rapid collapse in August of 2021, it is very probably a contributing factor. This focus on the quantity rather than the quality of solders did result in their having significant deficiencies as a fighting force [31]. This occurred despite warnings from the Special Inspector General for Afghanistan Reconstruction dating back to at least 2016 [33].

Ship maintenance delays

Not all "beneficial adjustments" to measurements are corrupt manipulations. In some instances, adjustments are warranted. For example, the Naval Sea Systems Command, which is responsible for scheduling shipyard maintenance and ensuring that ships finish their maintenance availabilities on time, recently redefined how they calculate the amount of time a ship will remain in the yard [34]. These new (longer) estimates reduced the number of "days lost" due to maintenance overruns from 7,000 to just 1,100.

The improved measurement resulting from this redefinition of maintenance availability times might appear to be an instance of Goodhart's Law in action. In reality, though, the Naval Sea Systems Command commissioned an analysis that showed they had previously been packing an unrealistic amount of maintenance work into the availability timeframes, making schedule overruns inevitable [34]. Unlike the artificial extension by the airlines of estimated flight times described earlier, this schedule extension is legitimate and defensible by analysis.

Technology transition

The issue of technology transition—ensuring that investments in basic and applied research eventually turn into viable products—is not limited to government-funded military research and development. Private companies also fund product development research. We include it here under defense analysis examples because technology transition is a major concern within the Department of Defense (DOD) [35-36].

Governments around the world spend heavily on defense research. The US alone spends over $100 billion in annual military research expenditures [37]. Each country seeks to measure the return on its research investments; China provides an example[38]. The Chinese government sought to measure both their research output and their technology transition rate. However, they noted a significant disconnect between their chosen measures for each. They measure research production by counting research publications and citations—a number that is increasing rapidly. They measure technology transition by counting commercialized patents, and this number languishes at about 10 percent of awarded patents.

Further investigation into these quantitative measures showed that Chinese academics were purposely publishing "garbage papers" to boost their Science Citation Index (SCI) scores [38]. This practice became so prevalent that the Chinese military published an article titled "Military Research Should Guard Against 'SCI Worship.'" In this way, a measure designed to measure research productivity was manipulated by researchers seeking career advancement.

The use-it-or-lose-it "rule"

One method that federal comptrollers use to manage accounts is reprogramming. They identify and recoup funds allocated to programs that have not yet been spent and provide those funds to other (related) programs that need them. Reprogramming happens continuously, but it is the most pronounced at the end of the federal fiscal year [39]. The use-it-or-lose-it "rule" incentivizes program managers to spend money for fear that they will lose it, and that failure to spend it will justify decreasing their budget for the next year. It is technically illegal to not spend appropriated funds [40], but about 1.6 percent are returned to the treasury each year [41]. This fact (and the paperwork required to return unspent funding) provides additional incentive to quickly spend any soon-to-expire funds.

Spending in the last few weeks of the year is almost five times higher than average and is often wasteful or inefficient [42]. Thus, the allocated budget becomes a target and is no longer a good measure of what the program manager actually needs to execute the program. With this "rule" in effect, analysts can measure how much was spent but not how much money was needed [43]. Ironically, the only time the data are accurate under these circumstances is when a budget

is overrun. Federal government program managers are put in the untenable position of not being permitted to go over or under budget—they may only break even.

The 355-ship fleet

Despite the fact that different naval vessels provide different forms of combat power in different ways, the number of ships in the fleet is now an accepted measure of the "combat power" of the US Navy. Perhaps part of the attractiveness of this measure is that it is easily grasped by the public and by the political leaders who fund the construction, operation, and maintenance of ships. In 2017, attainment of a 355-ship fleet became national policy [44].

An accurate ship count is so important that the Secretary of the Navy publishes an instruction [45] on what ships can and cannot be counted. This instruction—dating to before the advent of plausible unmanned US Navy ships—does not require US Navy ships to be crewed. Coupled with the policy requirement for a 355-ship fleet, the downward budget pressures, the upward spiral of new ship construction costs, and the (perceived) relative inexpensiveness of unmanned ships, this omission tempted the invocation of Goodhart's Law by adding unmanned ships to the official US Navy ship count.

Indeed, in 2020, then-secretary of defense Mark Esper stated that "unmanned will enable us to grow the US Navy well beyond 355 ships" [46]. Though this approach would help satisfy the policy requirement to get to a 355 ship fleet, it did not necessarily lead to a fleet with increased combat power (which is the real goal) because these ships remain untested.

Chief of Naval Operations (CNO) Admiral Michael Gilday stated as much: "There are a lot of assumptions that go along with unmanned because they're pretty much conceptual [capabilities right now]. And so the final [official ship count] numbers that will come out in a couple of weeks … will not include unmanned" [47]. Unlike many of the previous examples in this report, this is a refreshing example of principled resistance to the corrupting influence of Goodhart's Law.

This extensive and wide-ranging list of instances of Goodhart's Law leads us to the question, What are the forces that drive analysts to choose measures that are so easy to manipulate? We turn to that question in the next section.

Measures of Effectiveness and Measures of Performance

The previous examples of Goodhart's Law all have three things in common: there is a desired effect to be achieved, there is a need to measure the degree to which that effect is being achieved, and direct measurement of the desired effect is either difficult or impossible. The analyst must then identify some other attribute of the system under study that is not difficult or impossible to measure, and whose value correlates in some known way with achievement of the desired effect. The analyst is stuck measuring this proxy variable.

In these examples, measurement of the proxy variable is also a measure of the performance of some person or organization. Sometimes this person or organization is actually responsible for achieving the underlying desired effect, and the measurement is made for accountability purposes. Other times they are measured and rewarded in order to incentivize a behavior that achieves the desired effect.

Thus, two related but very different concepts are to be measured here: ***effectiveness*** and ***performance.*** In our examples, the former cannot be measured, but the latter can be. We will now show that understanding each and (more importantly) the relationship between them can help the analyst recognize and mitigate the negative consequences of Goodhart's Law.

Definitions

Unfortunately, multiple accepted definitions for *measures of effectiveness (MOE)* and *measures of performance (MOP)* exist across the DOD. The Defense Acquisition University provides definitions applicable to the development of military equipment and systems (quoted):

- MOE: The data used to measure the military effect (mission accomplishment) that comes from using the system in its expected environment. That environment includes the system under test and all interrelated systems, that is, the planned or expected environment in terms of weapons, sensors, command and control, and platforms, as appropriate, needed to accomplish an end-to-end mission in combat [48].
- MOP: System-particular performance parameters such as speed, payload, range, time-on-station, frequency, or other distinctly quantifiable performance features. Several MOPs may be related to achieving a particular Measure of Effectiveness (MOE) [49].

Note that the latter definition requires the MOP (but not the MOE) to be "distinctly quantifiable," and that the MOP is "related to" (i.e., correlated with) achieving a particular effect.

Joint Publication 1-02, the Department of Defense Dictionary of Military and Associated Terms [50], provides definitions for *MOEs* and *MOPs* that are instead more applicable to military operations and military forces (quoted):

- MOE: A criterion used to assess changes in system behavior, capability, or operational environment that is tied to measuring the attainment of an end state, achievement of an objective, or creation of an effect.
- MOP: A criterion used to assess friendly actions that is tied to measuring task accomplishment.

Here, the link from the MOP to the MOE is not explicitly stated, nor is the requirement that the MOP be quantifiable. This publication in turn references Joint Publication 3-0 Joint Operations [51], which does not define these terms but does state the intention of MOEs and MOPs succinctly by posing two questions (quoted):

- MOEs help answer the question, "Are we creating the effect(s) or conditions in the operational environment that we desire?"
- MOPs help answer the question, "Are we accomplishing tasks to standard?"

We direct the reader to consider these two questions as we demonstrate that a thorough understanding of MOEs and MOPs can help analysts identify and mitigate the negative effects of Goodhart's Law. We do so by using a specific example and then by extrapolating these specific observations to the more general case.

A specific example

Consider the previously mentioned case of cobra eradication in British India. The desired effect was a reduction in the number of cobras. A suitable MOE would be to simply count the cobra population *directly* before and after some eradication program to see whether that program was indeed effective. Unfortunately, employment of this MOE has two very clear disadvantages. First, periodically counting all the cobras would be labor intensive (and dangerous). Second, if there is more than one eradication program, or more than one person executing an eradication program, then it becomes difficult to attribute (and reward) reduced cobra counts to any one program or person, making it difficult to incentivize whatever behavior may have resulted in fewer cobras.

The beauty of directly measuring the effect (counting remaining cobras in this case) is that it cannot be manipulated. The effect is the effect. Either the cobra population is declining as desired, or it is not. Directly measuring the effect itself is a measure of effectiveness—an MOE—of the program or person eradicating cobras.[2]

Unfortunately, the two disadvantages noted above drove British officials away from the use of this direct MOE and toward the use of a MOP that served as a *convenient and reasonable proxy measurement* of the desired effect. They reasoned that one cobra skin was equivalent to one less live cobra, and that the person presenting that skin had killed the cobra and was thus entitled to the agreed upon bounty for killing a cobra. We know now that though these assumptions were perfectly reasonable, they turned out to be false, thanks to Goodhart's Law.

Generalizing this specific example

The logic applied to the "cobra effect" example can be applied to all of our examples, showing that a common theme runs through them (indeed, it runs through much of the field of operations research as currently practiced). The desired effect is known, but it is difficult or impossible to make a direct measurement of the desired effect; therefore, we identify some more-easily measurable item or activity that is directly correlated with the desired effect, and we measure that. When that measurement also measures the performance of some person or organization (as is often the case), we have then moved from a measure that is difficult to make and difficult to manipulate (the MOE) to a measure that is easier to make but also easier to manipulate (the MOP).

We must note here that MOEs may not be available for reasons beyond impracticality. Sometimes effectiveness simply cannot be measured. How, for example, does one measure a child's educational attainment, or the combat power of a naval fleet? No underlying theory exists to describe how these (dependent) output variables relate to their (independent and modifiable) input variables. This larger issue is more suitable to a report examining MOEs and MOPs specifically. We note it here, though, because shifting from MOPs to MOEs to mitigate the effects of Goodhart's Law is not an available option if no theory exists to articulate and quantify the MOE for a particular case.

We also note another consideration briefly alluded to in the section on the use-it-or-lose-it "rule" when we noted that federal program managers are not supposed to make (or lose) money. Private-sector business examples are notably absent from our list, which makes sense upon closer examination. Businesses literally have a "bottom line" that serves as a perfect MOE.

[2] Assuming, of course, that all other variables remain constant. For example, there is no disease, increase in the mongoose population (a cobra predator), or habitat destruction that could reduce the cobra population.

Businesses are designed to make money. Any proposed changes to business operations can be immediately and ruthlessly evaluated based on how they affect profits. The simplicity of this MOE is a luxury that business analysts enjoy but that other operations analysts may not.

The need to resort to reasonable and convenient proxy MOPs instead of direct MOEs invites the emergence of Goodhart's Law. The typical analyst who seeks to mitigate the negative effects of Goodhart's Law would therefore be wise to understand the differences between MOEs and MOPs, their availability, and their relationship to each other. The analyst can use MOEs whenever possible *and* practicable to mitigate Goodhart's Law, and MOPs may be used only if and when they must be. The informed analyst is now better equipped to identify and mitigate the negative consequences of Goodhart's Law.

Recommendations

Having demonstrated the pervasive and pernicious effects of Goodhart's Law in the preceding examples, and having discussed the differences and interrelationships between MOEs and MOPs, we now turn to recommendations for how to avoid or mitigate the effects of Goodhart's Law. We provide recommendations for individual analysts (and for their analytical teams) and for the organizations that employ them. Some recommendations will be best implemented in partnership with the customer organizations supported by the analysts. Many are applicable to all analysts and all analytical organizations. Others are applicable only to public-sector analysis, while still others are applicable only to defense analysis.

We note here that the utility of these recommendations depends on the individual analyst's circumstances. It is entirely possible that implementing a recommendation to avoid or mitigate the effects of Goodhart's Law may introduce some other problem or impracticality. These recommendations should be read with this caution in mind.

Recommendations for analysts

Paraphrasing an observation by 19th-century French economist Frédéric Bastiat:[3]

> ***There is only one difference between bad analysts and good analysts: bad analysts confine themselves to the visible effect; good analysts take into account both the effect that can be seen and those effects that must be foreseen [52].***

To be a good analyst, one must have (among *many* other skills) the ability to foresee the corrupting influences of Goodhart's Law and to recognize that they can exist in the past, the present, and the future:

- **Past.** Analysts must recognize that the pre-existing data they are preparing to use may have already been corrupted by Goodhart's Law.
- **Present.** When developing measures that are to become part of their delivered analysis, analysts should endeavor to identify and avoid the use of measures that may result in the emergence of Goodhart's Law after delivery.

[3] We replace "economist" with "analyst," and we use the gender-neutral plural rather than the masculine singular.

- **Future.** When the use of potentially corruptible measures cannot be avoided, the analyst must communicate this possibility to customers and also provide advice on how the customer can identify and mitigate the corrupting effects of Goodhart's Law.

What, then, are the potential mitigation techniques available to the analyst (the "concrete steps we can—and must—take" [53]) to avoid the corrupting effects of Goodhart's Law—past, present, and future? We recommend the following.

Use MOEs whenever practicable

A problem avoided is a problem solved. Given that Goodhart's Law often emerges when we resort to the use of MOPs over MOEs, choosing to directly measure effects rather than indirectly measuring performance solves the problem of Goodhart's Law. Unfortunately, as we noted in the previous section, Goodhart's Law is pervasive partly because MOEs are inherently difficult to measure and sometimes impossible to even quantify. Still, the analyst should consider the feasibility of using a directly measured MOE before resorting to the proxy MOP and its accompanying difficulties.

Use the scientific method

In the scientific method, a testable hypothesis leads to a question, the scientist designs an experiment that will generate the data needed to answer the question, the experiment is conducted, the data are collected and analyzed, the question is answered, and the hypothesis is verified or refuted. *The data are tailored to the question, rather than the question being tailored to the available data.* The deliberate choice and collection of data provides an additional opportunity to ensure that the data measure an effect rather than performance, and are thus not subject to the corrupting influences of Goodhart's Law.

Help customers establish defensible authoritative definitions for measures

Perhaps the easiest way to manipulate a measure is to manipulate its definition. When then-secretary of defense Mattis directed the US Navy and Air Force to achieve 80 percent [25] readiness rates without providing an authoritative definition of readiness, he invited manipulation of the measure by manipulation of its definition [29-30].

The US Navy's decision to use AMSRR instead of DECKPLATE data to measure readiness [29] provides an example. Though it might appear that this was a judicious choice of a *database*, it was actually a judicious choice of a *definition*, as each database aggregates the data within it to arrive at different definitions of aircraft availability.

Despite acknowledging to the GAO in 2020 that DECKPLATE data provide a "more comprehensive measure" [29] of fighter aircraft readiness, the Navy declared in response to a follow-up report in 2022 [54] that AMSRR is its authoritative data source for fighter aircraft readiness calculations. The response noted that readiness calculations using DECKPLATE are based on "out of date targets" [54] and that they now measure "number of mission capable aircraft required, vice a percentage goal" [54]. The Air Force rejected the 80 percent readiness target as an improper metric in 2019, but the Navy waited until 2022 to do so.

Because of the definitional disagreement between the Navy and the GAO, knowing the true state of fighter aircraft readiness remains elusive. A solution to this conundrum would be to establish a transparent, justifiable, and defensible definition and to make it the authoritative definition. The recalculated shipyard maintenance availability time discussed earlier is a good example of how this could be accomplished.

Technically, defense analysts do not have the authority to establish an authoritative definition—their customers hold this power. However, defense analysts (at least the ones aware of Goodhart's Law) do have the skills and abilities needed to make the authoritative definition defensible. Analysts who are unaware of Goodhart's Law may blindly accept the problematic measure and measure it.

If Navy and GAO analysts had been more aware of Goodhart's Law in 2019, then they might have joined the Air Force in pushing back against Mattis' insistence on using this incongruous MOP as a proxy for the true desired effect of fighter aircraft. The Air Force did so then, and CNO Admiral Gilday pushed back against then-secretary of defense Esper's similar attempts to count experimental unmanned vessels as part of the fighting fleet in 2020 [47].

Identify decoupling opportunities

If the person or organization being measured has no power to manipulate the measure being used to evaluate their performance, then Goodhart's Law is mitigated. A recent study on operation and support costs for ships and aircraft [43] provides an example. The military services collect data on these costs using authoritative definitions and databases. But using these data to predict future costs is inadvisable because the data can be manipulated by the forces of the use-it-or-lose-it "rule" [43].

Using a machine learning technique, the study identified a strong correlation between the numbers of personnel assigned to ships and aircraft and their operation and support costs [43].[4] In the US Navy, personnel accounts are funded by the Deputy Chief of Naval Operations for Manpower, Personnel, Training, and Education (OPNAV N1), and platform operation and

[4] Personnel costs do make up the majority of operations and support costs, but this correlation existed even for platforms with no crew, and for costs not associated with personnel.

support accounts are funded by the Deputy Chief of Naval Operations for Warfighting Requirements and Capabilities (OPNAV N9). Any attempt by OPNAV N9 to manipulate the measurement of operation and support costs[5] by changing the numbers of personnel assigned to ships and aircraft would provoke an immediate reaction and a demand for justification from OPNAV N1, because it would upset their carefully crafted personnel management plans and budgets. In fact, one could postulate that any wild deviations from the manpower-to-operation and support costs relationship could serve as an indicator that the budgets for operation and support costs are being manipulated by the use-it-or-lose-it "rule."[6]

Avoid the use of manipulated data

Good analysts already verify the quality of the data they need to use. The data must be, among other things, accurate, auditable, complete, consistent, credible, current, and timely [56]. Pre-existing data that may have been manipulated by the effects of Goodhart's Law may not be accurate or credible. Identifying and discarding manipulated data is a worthwhile investment of an analyst's time as failure to do so results in a "garbage-in-garbage-out" analytical product.

Avoid the use of data prone to manipulation

One way to "pre-validate" a data source against prior manipulation is to identify who collected the data and determine whether their performance was assessed based on that data. Data collected by unaffected and thus disinterested persons are preferable to data collected by affected persons. With exceptions, data collected by machines are to be preferred over data collected by humans.

Collect or analyze data secretly

If the persons or organizations being measured are not aware of the data being used to evaluate their performance, then they cannot manipulate that data. The primary example of this technique is the constant back-and-forth actions of internet search engines and the search engine optimization industry [57]. It is no secret that the search engines collect data on web page relevance; but how they calculate the resulting relevance rankings is kept secret [3].

[5] This is not to imply that OPNAV N9 would do such a thing. This is merely an illustrative example of a decoupling opportunity familiar to the authors.

[6] We must note here that because personnel costs are such a large portion of total operations and support costs, the Navy has sought to reduce crew sizes for new ship classes for the past three decades. As a result, the interests of the personnel resource sponsors and the platform resource sponsors are not always cleanly decoupled—at least not during the acquisition phase. The limited success in reducing operations and support costs by reducing crew sizes is the subject of multiple reports by the Naval Research Advisory Council [55].

A variation on this technique might be to openly collect multiple types of data but not divulge which data are used to measure performance. Alternatively, a combination of public and private measures—public measures to drive behavior in the desired direction and "secret" private measures held back as a corruption check may present an acceptable balance.

Secrecy presents other problems that might limit its utility as a mitigation technique. Sunshine laws exist to prevent secrecy [58], and in some cases, secretly collecting data about an individual or their activities might break privacy laws [59]. Analysts would be wise to seek legal counsel before using secrecy as a mitigation technique.

Avoid suboptimization

Choosing measures (or collections of measures) that represent *all* aspects of the system under study mitigates the effects of Goodhart's Law simply because it is more difficult to manipulate all measures than it is to manipulate just one or two [3, 53].

By not measuring overall performance, analysts might actually do more harm than good; for example, a focus on one measure that incentivizes resources to flow toward fixing that measure may negatively affect other measures in a resource-constrained environment.

Choosing math and reading as the subjects covered on standardized tests is one example. The effects of Goodhart's Law mean that these tests will be poor measures of overall educational attainment if the arts and sciences are subsequently slighted in the curriculum: "That schools have focused an ever-increasing number of hours on these subjects, to the near-exclusion of all else, is a reasonable, if undesirable, response" [53].

Fighter aircraft readiness measures provide another example. The flight-worthiness of the aircraft is not the only requirement for military readiness [60]. A flyable military aircraft with a poorly trained pilot may be a ready aircraft, but it is not a ready combat capability. The same is true of an aircraft that can fly but has only one operable sensor or weapons system.[7]

Although suboptimization in educational measurements might be merely detrimental—producing students who could thrive as mathematicians but would fail as artists or scientists—suboptimization in fighter aircraft readiness measurements can be catastrophic. This is because the fighter aircraft, its subsystems, and its crew make up a kill chain, and all links in

[7] Fighter aircraft mission capable rates are defined as the percentage of time an aircraft can fly and perform *at least one* mission [29].

the chain are required to achieve the desired effect. Strengthening one link in the chain at the expense of the others may result[8] in a more-easily broken chain—and mission failure.

We must also note that sometimes suboptimization may be doing the right thing—just for the wrong reasons. Returning to our two examples, if math happens to be a student's weakest subject, then focusing on math in order to pass a math test has the fortuitous—if unintended—side effect of focusing additional instructional resources where they will do the most good for overall educational attainment. Similarly, if airframe readiness is the weakest link in the military kill chain, then the demonstrated maintenance improvements that came out of the attempt to meet the 80 percent readiness rate [62] focused scarce resources precisely where they would do the most good—if only by happenstance instead of deliberate design.

Focusing on just one or a few measures is attractive for many reasons: available analytical resources may be limited, some measures are easier to quantify and understand than others, and the data for some measures may be more difficult to obtain. We recognize the natural desire to suboptimize because of these considerations, but the analyst must also recognize that this choice makes manipulated measures more probable.

Randomize measures over time

As noted above, it is preferable, but not always possible, to use all relevant measures. When this is not possible, one mitigation technique might be to randomize the use of the suboptimal measures. Returning to the example of standardized testing, if the objective is to improve the quality of education overall, and not just in math and reading, then the other subjects of interest should also be measured. Randomly (and secretly) picking which subjects will be measured on the next standardized test would force educators to cover all potentially testable subjects. Math and art proficiency could be measured one year, reading and science the next.

In the military readiness example, aircraft availability could be measured one year, pilot proficiency the next, and weapons or sensor systems the next (though not in a predictable order or combination). We acknowledge that randomized measures have the obvious drawback that it then becomes more difficult to track adequate yearly progress in every educational or readiness category, as not all categories are tested each year.

Use post hoc measures

When a measure is chosen after all actions are taken (post hoc), then it is effectively secret [63]. The measure cannot be manipulated because it does not yet exist as a measure. This technique

[8] Investments in individual kill chain links are sometimes synergistic rather than mutually exclusive. For example, better aircraft maintenance leads to more available aircraft, which in turn leads to more flying hours for pilots, resulting in higher pilot readiness [61].

may be of limited utility to most analysts because they are engaged in taking measures in order to improve future performance, rather than just to assess past performance. Another potential issue is that historical data can lose context as circumstances in the measured environment change over time. Post hoc measures may not be applicable to current operations.

Wargame or red team potential measures

Although a *wargame* has many definitions [64], one definition from the Marine Corps Warfighting Laboratory, Wargaming Division makes it clear that wargames need not be limited to analyzing the conduct of war: "A method wherein the human intellect uses a synthetic construct that replicates a conflict and requires decisions for resolution in order to consider a real problem" [64].

A person or organization whose performance is being measured is presented with a potential conflict when they are made aware of the measures used to evaluate them. This requires a decision on how to respond. This response may include manipulation of that measure, so wargaming and red-teaming activities are a good way to identify the incentives and opportunities that may lead to manipulated measures [63].

Such activities could be conducted with the customer of the analysis before the final analytical product is delivered. In this way, the potential for manipulation can be discovered before it occurs, and in an environment free of negative recriminations. Once identified, measures subject to potential corruption can be replaced or tracked to halt attempted manipulation.

Recommendations for analytical organizations

We contend that arming analysts with the knowledge of Goodhart's Law and the mitigation techniques above to combat it is not sufficient. The effects of Goodhart's Law are so pernicious and pervasive that analytical organizations should *institutionalize* these mitigation techniques [53]. We provide six institutional-level recommendations to that effect.

Return to the roots of operations research

This first of our organizational recommendations supports the first two recommendations above for analysts: to use MOEs whenever practicable, and to use the scientific method. The typical analyst faces significant barriers in attempting to implement these two recommendations that require some effort by their employing organizations to remove. Consider the following barriers:

- Analytical questions often arrive with short deadlines, and annual government contracting practices limit the time available to answer them—let alone collect the

required data. As a result, defense analysts (and perhaps others) become accustomed to answering the question at hand with the data at hand.

- Although not strictly limited to military culture, the officers who are the recipients of defense analysis are trained to, are accustomed to, and are comfortable with making quick decisions based on imperfect information. Matching this culturally ingrained decision speed forces the analysts to answer the question at hand with the data at hand. Critical questions for which no data exist beforehand go unanswered.
- The recent pervasiveness of data analytics approaches often operates on an implicit assumption and expectation that an organization is awash in exploitable data. This is not always so [65]. This limits the analyst to questions that can be answered (based on data availability) rather than the questions that should be answered.

Many analysts have a scientific background. Rather ironically (or perhaps not), the field of operations research as pioneered before and during World War II by the British and US militaries is predicated on sending scientists to the field to practice the scientific method in the field by collecting and analyzing the data required to solve operational problems [66]. Some organizations still do this—at least temporarily—with many or all of their analysts [67]. This field presence gives analysts the opportunity to identify and collect data on *direct* and less manipulatable MOEs, rather than having to rely on indirect or proxy MOPs. More importantly though, they need not limit themselves to the data at hand. Limiting analysis to MOPs and the data at hand can increase the chances that the data have been, or could be, manipulated by the perverting incentives present in Goodhart's Law.

If analysts are to use MOEs, and the scientific method more extensively, their employing organizations must reduce, and preferably remove, the three barriers listed above. Specifically:

- Analytic organizations under contract to the US government should communicate the data collection limitations created by rigid annual contracting cycles to their government contracting officers. They should further communicate the potential benefits of increased use of MOEs (versus MOPs) and the scientific method and seek to either create, or expand, multi-year contracting relationships.
- Analytic organizations with customers for whom quick decision-making based on imperfect information is the cultural norm should work with these decision-makers to strike an appropriate balance between rapid, data-sparse analysis and deliberate, data-rich analysis, rather than reflexively acquiescing to the cultural norm that limits analytically supported decisions to those with the (often inadequate) data at hand.
- Analytical organizations that provide data analytics services should do two things:
    - Help their supported organizations expand and mature their in-house data infrastructure so that they actually are "awash in exploitable data," thereby greatly expanding the opportunities to leverage data analytics [56, 68].

- Help their supported organizations recognize that quick-turn data analytics is limited to the data at hand, which may cause decision-makers and the analysts who support them to limit themselves to the questions that *can* be answered, rather than seeking out the questions that *should* be answered.

Train analysts on MOEs and MOPs

Some newly hired analysts may have a degree in operations research that would provide them with previous exposure to the concepts of MOEs and MOPs, but most newly hired analysts, even if they are being hired into a position with the title of operations researcher, come from the hard sciences or social sciences where they have had no such previous exposure. Given the strong connection between MOEs, MOPs, and Goodhart's Law, some instruction in this area would seem appropriate. Even when MOPs do not result in measure manipulation via Goodhart's Law, an understanding of MOEs and MOPs is still clearly of value to the operations researcher.

Train analysts on Goodhart's Law

Simply making analysts aware of Goodhart's Law is a necessary, but not sufficient, first step to mitigate its effects. A good next step might be to expose new analysts to Goodhart's Law use cases—the examples in this report would suffice—and then conduct a training exercise in which they must identify an instance of Goodhart's Law from their own personal experience or observations.

Alternatively, analysts could be presented with a use case, instructed to assume as a null hypothesis that Goodhart's Law is affecting the measure used, and then prove that it is not. The training cohort could then discuss all identified instances and postulated mitigation techniques—a procedure that we should expect them to conduct as members of future analytical teams.

Organizations that hire analysts trained in the hard sciences (e.g., physics, chemistry) may have to pay particular attention to Goodhart's Law as a training requirement. Unlike analysts from the social sciences (e.g., economics, sociology), their professional training is with elements of the natural world that are incapable of manipulating measures. Social scientists, in contrast, are more likely to already have professional analytical experiences that dealt with measuring potentially manipulative human beings. Organizations that hire analysts from both the hard sciences and the social sciences can leverage the latter group to help the former group acclimate to the potential instances of Goodhart's Law.

Make MOEs, MOPS, and Goodhart's Law part of the peer review process

Presumably, any good analytical organization has some form of a peer review process. These processes often have "required elements"—issues that must be addressed before an analytical product is approved for release to the customer. Examples include verifying that conclusions are traceable to findings and verifying that previous work is properly cited.

To truly institutionalize the techniques that mitigate manipulation of measures, they must be part of an institutionalized process. Asking questions during the peer review process regarding the three past, present, and future implications of Goodhart's Law would accomplish this:

- Did the analyst examine the data used in the study to determine whether it might be subject to (prior) manipulation by the original data collector?
- Did the analyst use MOEs that cannot be manipulated, or MOPs?
- Did the analyst identify any MOPs used that are vulnerable to future manipulation?

This recommendation has the added advantage of being relatively easy and inexpensive to implement.

Make the potential consequences of Goodhart's Law a required part of delivered analysis

Naturally, if Goodhart's Law is part of the peer review process, and that process identifies potential consequences for the customer's organization, then the analyst is obliged to communicate this danger to the customer. Additionally, actions that the customer organization can take to identify when their measures are being manipulated, or actions they can take to prevent this manipulation, should be communicated in the final analytical product.

Given that various time and resource constraints often affect analysis, it may not always be possible to follow this and the other recommendations we have made. In such situations, analysts should still explicitly communicate any mitigation steps that could not be taken and the resultant effects that Goodhart's Law may have on their delivered analysis.

Identify and share best practices

The mitigation techniques described in this report come from a brief literature review [2-3, 5, 8, 23, 53, 63, 69-70] and from the collective analytical experience of the authors. If the preceding recommendations are taken by an analytical organization, then their analysts are likely to discover and develop additional mitigation techniques. When new techniques appear in the peer review process, they can be captured and injected into the analyst training process.

Conclusion

Mitigating the negative effects of something that rarely occurs might not be worth the effort of institutionalizing the use of mitigation techniques. The same is true for negative effects that are minor. Goodhart's Law, however, has been shown to be both pervasive and pernicious—so much so that the corrupting influence of Goodhart's Law should be treated as the null hypothesis (i.e., only dismissed when it can be explicitly demonstrated to not be present). It can exist in any system in which humans have a vested interest in measurements and have the means to manipulate these measurements. It can have significant negative effects if not mitigated. In defense analysis, these negative effects may include (and have included) unintentional death.

The observations made in this report might seem rather dire. The pervasive and pernicious effects of Goodhart's Law might lead one to conclude that analysis in pursuit of improved operational effectiveness is futile. Our analysis (of analysis) indicates that this need not be the case. Analysts, analytical organizations, and even analytics customers can take steps (recommended here) to avoid the circumstances that lead to measure manipulation. When unavoidable, other steps (also recommended here) can be taken to recognize, mitigate, and even reverse its emergence. These steps should be coupled with penalties for knowingly engaging in measure manipulation and an organization-wide understanding (by analysis providers *and their customers*) that such activities are being monitored and will be called out. Doing so can greatly reduce the occurrence of measure manipulation, thereby greatly increasing the value of the customer organization's investments in analytical support. The negative effects of Goodhart's Law are only dire for those who remain ignorant of it, and who ignore the recommendations in this report.

Organizations that use data and develop measures that have consequences—both positive and negative—for the persons, organizations, and processes they are charged with measuring and improving should therefore act on these recommendations to identify, understand, avoid, mitigate, or reverse the effects of Goodhart's Law.

The recommendations we have offered for individual analysts and for the organizations that employ them can prevent or otherwise mitigate the effects of Goodhart's Law, benefiting the analysts and the organizations that they support. Admittedly, these recommendations constitute additional burdens on project managers already stressed by limited budgets and tight schedules. Because the negative outcomes of not assuming these burdens will not occur until a future date, they may go unrecognized, or it may be tempting to dismiss them if recognized. Therefore, institutionalizing these additional actions by including them in a

required process (like a peer review) is essential to avoid the pervasive and pernicious effects of Goodhart's Law.

References

[1] Stumborg, Michael F. 2007. "The Elements of Successful Military Transformation: Applying Lessons Learned from Science, History, and Corporate America." *High Frontier* 3 (4): 46-8. https://www.afspc.af.mil/Portals/3/documents/HF/AFD-070814-023.pdf.

[2] Goodhart, C.A.E. 1975. "Problems of Monetary Management: The UK Experience." *Papers in Monetary Economics* I (Reserve Bank of Australia). https://cyberlibris.typepad.com/blog/files/Goodharts_Law.pdf.

[3] Anzil, Federico. "Goodhart's Law." Economic Point. https://economicpoint.com/goodharts-law.

[4] Maunz, Shay. "The Great Hanoi Rat Massacre of 1902 Did Not Go as Planned." Atlas Obscura. Jun. 6, 2017. https://www.atlasobscura.com/articles/hanoi-rat-massacre-1902.

[5] Hartley, Dale. "The Cobra Effect: Good Intentions, Perverse Outcomes." *Psychology Today*. Oct. 8, 2016. https://www.psychologytoday.com/us/blog/machiavellians-gulling-the-rubes/201610/the-cobra-effect-good-intentions-perverse-outcomes.

[6] *The No Child Left Behind Act of 2001*. Jan. 8, 2002. The 107th Congress of the United States. https://www.congress.gov/bill/107th-congress/house-bill/1/text.

[7] "Teacher Cheating and Standardized Testing." FindLaw. Jun. 20, 2016. https://www.findlaw.com/education/curriculum-standards-school-funding/teacher-cheating-and-standardized-testing.html.

[8] Waters, Tony. "Campbell's Law, Planned Social Change, Vietnam War Deaths, and Condom Distributions in Refugee Camps." ethnography.com. Mar. 14, 2010. http://www.ethnography.com/2010/03/campbell%e2%80%99s-law-planned-social-change-vietnam-war-deaths-and-condom-distributions-in-refugee-camps/.

[9] Spielhagen, Frances R. "Don't Leave Gifted Students Behind." Education Week. Feb. 12, 2012. https://www.edweek.org/teaching-learning/opinion-dont-leave-gifted-students-behind/2012/02.

[10] Bannister, Benjamin. "SEO Secrets: Reverse-Engineering Google's Algorithm." Free Code Camp. Apr. 26, 2017. https://www.freecodecamp.org/news/seo-secrets-reverse-engineering-googles-algorithm-92fad4f5a39/.

[11] Hochman, Jonathan. 2016. "What is Search Engine Optimization Manipulation?" *The National Law Review* XII (3). https://www.natlawreview.com/article/what-search-engine-optimization-manipulation#:~:text=SEO%20manipulation%20tactics%20include%20keyword,to%20users%20and%20search%20engines.&text=When%20a%20penalty%20occurs%2C%20the,claim%20against%20their%20SEO%20agency.

[12] Ofiwe, Michelle. "How Does the Google Search Algorithm Work in 2021?" Semrush. Oct. 7, 2021. https://www.semrush.com/blog/google-search-algorithm/.

[13] Renzulli, Dante. "Mail Carriers Accuse USPS of Faking Amazon Delivery Records so Customers Don't Get Free Stuff." *CBS News Channel 46 (Atlanta, Georgia)*. Nov. 15, 2017. https://www.cbs46.com/news/mail-carriers-accuse-usps-of-faking-amazon-delivery-records-so-customers-dont-get-free-stuff/article_5b4b48f0-cfc6-5b5a-8731-21cd4bc13a55.html.

[14] "Delta Pays $10.5 Million to Settle Post Office Allegations." Associated Press. 30 Jun. 2022. https://www.local10.com/news/politics/2022/06/30/delta-pays-105-million-to-settle-post-office-allegations/.

[15] Walensky, Rochelle, and Megan Henney. "CDC Director Criticized for Now Differentiating Between Dying 'From' vs. Dying 'With' COVID-19." *Fox News Channel*. May 16, 2021. https://www.foxnews.com/politics/cdc-director-walensky-criticism-updated-guidance-coronavirus-deaths.

[16] Neary, Graham. "Mainstream Media Notices the Fraudulent COVID Death Count." Apr. 18, 2021. https://grahamneary.wordpress.com/2021/04/18/mainstream-media-notices-the-fraudulent-covid-death-count/.

[17] Borresen, Jennifer, Janie Haseman, and Javier Zarracina. "Uncounted: Inaccurate Death Certificates Across the Country Hide the True Toll of COVID-19." *USA Today*. Dec. 22, 2021. https://www.usatoday.com/in-depth/news/nation/2021/12/22/covid-deaths-obscured-inaccurate-death-certificates/8899157002/.

[18] Agarwal, Nikhil, Charles Hodgson, and Paulo Somaini. "Choice and Outcomes in Deceased Donor Kidney Assignment." The Centre for Economic Policy Research. Jan. 17, 2021. https://voxeu.org/pages/about-vox.

[19] Palmer, Brian. "Fertility Clinic Data Means Bad Medicine: Success Rates are Misleading and Often Dishonest." Slate.com. Mar. 16, 2014. https://slate.com/technology/2014/03/fertility-clinic-success-rates-are-misleading-possibly-dishonest-and-promote-bad-medicine.html.

[20] Witek, Gregory. Nov. 3, 2020. "Agile Estimation - Why Story Points are Better Than Using Time?" Not Only Code. https://www.youtube.com/watch?v=4mrNtAHVEq8

[21] Donnelly, Drew. "An Introduction to the China Social Credit System." New Horizons. Oct. 26, 2021. https://nhglobalpartners.com/china-social-credit-system-explained/.

[22] Baveja, Sarabjit Singh, Anish Das Sarma, and Nilesh Dalvi. Mar. 2, 2021. *Determining Trustworthiness and Compatibility of a Person. US Patent Number 10,936,959.* AirBnB Inc. https://uspto.report/patent/grant/10,169,708.

[23] Hirsh, Jesse. "Social Credit: Gaming the System." Metaviews. Jan. 24, 2020. https://metaviews.substack.com/p/social-credit-gaming-the-system.

[24] Askonas, Jon. "A Vicious Entaglement, Part V: The Body Count Myth." War on the Rocks. Oct. 12, 2017. https://warontherocks.com/2017/10/a-vicious-entanglement-part-v-the-body-count-myth/.

[25] Mehta, Aaron. "Mattis Orders Fighter Jet Readiness to Jump to 80 Percent — In One Year." *Defense News*. Oct. 9, 2018. https://www.defensenews.com/air/2018/10/09/mattis-orders-fighter-jet-readiness-to-jump-to-80-percent-in-one-year/.

[26] Losey, Stephen. "Here's How Bad the Military's Aircraft Readiness Has Gotten." *Air Force Times*. Nov. 19, 2020. https://www.airforcetimes.com/news/your-air-force/2020/11/19/heres-how-bad-the-militarys-aircraft-readiness-has-gotten/.

[27] Eckstein, Megan. "Navy Surpasses 80% Aircraft Readiness Goal, Reaches Stretch Goal of 341 Up Fighters." *US Naval Institute News*. Sep. 25, 2019. https://news.usni.org/2019/09/25/navy-surpasses-80-aircraft-readiness-goal-reaches-stretch-goal-of-341-up-fighters?relatedposts_hit=1&relatedposts_origin=75671&relatedposts_position=2.

[28] Insinna, Valerie, and Stephen Losey. "US Air Force Bails on Mattis-era Fighter Jet Readiness Goal." *Defense News*. May 7, 2020. https://www.defensenews.com/air/2020/05/07/the-air-force-bails-on-mattis-era-fighter-jet-readiness-goal/.

[29] Maurer, Diana. Nov. 19, 2020. *Weapon System Sustainment: Aircraft Mission Capable Rates Generally Did Not Meet Goals and Cost of Sustaining Selected Weapon Systems Varied Widely.* US Government Accountability Office. GAO-21-101SP. https://www.gao.gov/products/gao-21-101sp.

[30] Reim, Garrett. Marine Corps Aircraft Readiness Numbers Inaccurate, IG Says." *Flight Global*. Aug. 10, 2018. https://www.flightglobal.com/fixed-wing/marine-corps-aircraft-readiness-numbers-inaccurate-ig-says/129209.article.

[31] Sep. 2017. *Reconstructing the Afgan National Defense and Security Forces: Lessons Learned from the U.S. Expereince in Afghanistan.* Special Inspector General for Afghanistan Reconstruction. https://www.sigar.mil/interactive-reports/reconstructing-the-andsf/index.html.

[32] Haltiwanger, John. "The Afghan Military Was Made Up of 'Ghost' Soldiers Who Didn't Actually Exist, and That's Why It Collapsed So Rapidly: Ex-Finance Minister." *Business Insider*. Nov. 10, 2021. https://www.businessinsider.com/afghan-military-was-made-up-of-fake-ghost-soldiers-ex-finance-minister-2021-11.

[33] John F. Sopko, Special Inspector General for Afghanistan Reconstruction. Aug. 5, 2016. Memorandum for the Hornorable Ashton B. Carter, Secretary of Defense. Subject: Efforts to Ensure Accuracy Across the Afghan National Defense and Security Forces Personnel Accountability Systems. https://www.sigar.mil/pdf/special%20projects/SIGAR-16-50-SP.pdf.

[34] Eckstein, Megan. "NAVSEA: Analysis of Ship Repair Processes Led to Better On-Time Rates, More Realistic Schedules." USNI News. Oct. 13, 2020. https://news.usni.org/2020/10/13/navsea-analysis-of-ship-repair-processes-led-to-better-on-time-rates-more-realistic-schedules.

[35] DOD Instruction 5535.03. May 21, 1999, Incorporating Change 1, Oct. 15, 2018. *DoD Domestic Technology Transfer (T2) Program*. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/553503p.pdf.

[36] DOD Instruction 5535.8. May 14, 1999, Incorporating Change 1, Sept. 1, 2018. *DoD Technology Transfer (T2) Program*. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/553508p.pdf?ver=2018-10-22-082514-847.

[37] Barnett, Jackson. "DOD asks for $112B in R&D funding in budget reques." FedScoop. May 28, 2021. https://www.fedscoop.com/dod-research-and-development-budget-request-fy-22/.

[38] Stone, Alex. *China's Model of Science: Rationale, Players, Issues.* China Aerospace Studies Institute. ISBN 9798407496274. https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Infrastructure/2022-02-07%20Model%20of%20Science.pdf.

[39] Fichtner, Jason J., and Joe Albanese. Jun. 2018. *Year-End Federal Spending and Government Waste: Reforming "Use It or Lose It" Rules.* Mercatus Center, George Mason University. https://www.mercatus.org/publications/year-end-spending-use-it-or-lose-it-rules.

[40] Yarmuth, John. "The Impoundment Control Act of 1974: What Is It? Why Does It Matter?". House Committee on the Budget. Oct. 23, 2019. https://budget.house.gov/publications/report/impoundment-control-act-1974-what-it-why-does-it-matter.

[41] Arkin, Jeff. May 25, 2021. *Federal Budget: A Few Agencies and Program-Specific Factors Explain Most Unused Funds.* US Government Accountability Office. GAO-21-432. https://www.gao.gov/products/gao-21-432.

[42] Liebman, Jeffrey B., and Neale Mahoney. Jan. 2018. *Do Expiring Budgets Lead to Wasteful Year-end Spending? Evidence from Federal Procurement.* National Bureau of Economic Research. Working Paper 19481. http://www.nber.org/papers/w19481.

[43] Stumborg, Michael F., Steven J. Full, Eric V. Heubel, Kyle F. Neering, and Shaye P. Storm. Jul. 2019. *Determining Cost Effectiveness of Unmanned Systems.* Distribution authorized to U.S. Government agencies and their contractors. Center for Naval Analyses. DRM-2019-U-020103-Final.

[44] Larter, David B. "Trump Just Made a 355-ship Navy National Policy." *Defense News*. Dec. 13, 2017. https://www.defensenews.com/congress/2017/12/14/trump-just-made-355-ships-national-policy/.

[45] SECNAV Instruction 5030.8C. Jun. 14, 2016. *General Guidance for the Classification of Naval Vessels and Battle Force Ship Counting Procedures*. https://www.nvr.navy.mil/5030.8C.pdf.

[46] Eckstein, Megan. "Esper: Unmanned Vessels Will Allow the Navy to Reach 355-Ship Fleet." *US Naval Institute News*. Sep. 18, 2020. https://news.usni.org/2020/09/18/esper-unmanned-vessels-will-allow-the-navy-to-reach-355-ship-fleet.

[47] Harper, John. "Navy Debates: Is a Robotic Ship a Ship?" *National Defense*. Jan. 22, 2020. https://www.nationaldefensemagazine.org/articles/2020/1/22/without-counting-unmanned-vessels-355-ship-navy-could-be-unobtainable.

[48] "Measure of Effectiveness." Defense Acquisition University Glossary. https://www.dau.edu/glossary/Pages/Glossary.aspx#!both|M|27925.

[49] "Measure of Performance." Defense Acquisition University Glossary. https://www.dau.edu/glossary/Pages/Glossary.aspx#!both|M|27926

[50] *Department of Defense Dictionary of Military and Associated Terms.* Nov. 8, 2010 (As Amended Through Nov. 15, 2014). Joint Publication 1-02. http://edocs.nps.edu/2014/December/jp1_02.pdf.

[51] *Joint Operations.* Jan. 17, 2017. Joint Publication 3-0. https://www.jcs.mil/Portals/36/Documents/Doctrine/docnet/jp30/story_content/external_files/jp3_0_20170117%20(1).pdf.

[52] Bastiat, Frederic. (reprinted) 1995. *Selected Essays on Political Economy*. Edited by George B. de Huszar, Translated by Seymour Cain. Irvington-on-Hudson, NY: Foundation for Economic Education. https://oll.libertyfund.org/title/bastiat-selected-essays-on-political-economy.

[53] Porter-Magee, Kathleen. 2013. *Trust but Verify: The Real Lessons of Campbell's Law.* The Thomas B. Fordham Institute. https://fordhaminstitute.org/ohio/commentary/trust-verify-real-lessons-campbells-law.

[54] Maurer, Diana. Jun. 15, 2022. *Air Force and Navy Aviation: Actions Needed to Address Persistent Sustainment Risks.* US Government Accountability Office. GAO-22-104533. https://www.gao.gov/products/gao-22-104533.

[55] "Naval Research Advisory Committee (NRAC) Published Reports." Office of Naval Research. https://www.onr.navy.mil/About-ONR/History/nrac/reports-and-executive-summaries/reports-alphabetical.

[56] Stumborg, Michael F. Mar. 2019. *Department of the Navy Data Readiness.* Distribution authorized to U.S. Government agencies and their contractors. Center for Naval Analyses. DRM-2018-U-018637-Final.

[57] David Bruce, Jr." All Saint's Media. https://allsaintsmedia.com/seo-cat-and-mouse-game-with-google/.

[58] Hayes, Adan. "Sunshine Laws." Investopedia. Feb. 24, 2021. https://www.investopedia.com/terms/s/sunshinelaws.asp.

[59] Hope, Alicia. "Google Involved in Yet Another Illegal App Tracking Privacy Lawsuit." *CPO Magazine*. July 24, 2020. https://www.cpomagazine.com/data-privacy/google-involved-in-yet-another-illegal-app-tracking-privacy-lawsuit/.

[60] Donovan, Matthew, Undersecretary of Defense, Personnel and Readiness. "The Wrong Way to Gauge Readiness." *Defense One*. Jan. 11, 2021. https://www.defenseone.com/ideas/2021/01/wrong-way-gauge-readiness/171301/.

[61] Eckstein, Megan. "Mission Capable: How More Ready Jets Is Helping the Navy Create Deadlier Pilots." USNI News. Apr. 23, 2020. https://news.usni.org/2020/04/23/mission-capable-how-more-ready-jets-is-helping-the-navy-create-deadlier-pilots#more-75722.

[62] Eckstein, Megan. "Mission Capable: How the Navy Harnessed Its Data to Achieve 80% Fighter Readiness." USNI News. Apr. 22, 2020. https://news.usni.org/2020/04/22/mission-capable-how-the-navy-harnessed-its-data-to-achieve-80-fighter-readiness#more-75671.

[63] Manheim, David. 2018. *Building Less Flawed Metrics: Dodging Goodhart and Campbell's Laws.* Ludwig-Maximilians-Universität München. Munich Personal RePEc Archive 98288. https://mpra.ub.uni-muenchen.de/98288/.

[64] Simpson, William L. Jul. 8, 2018. *A Compendium of Wargaming Terms.* Military Operations Research Society. https://www.mors.org/Communities/Communities-of-Practice/Wargaming.

[65] Stumborg, Michael. "See You in a Month: AI's Long Data Tail." War on the Rocks. Oct. 17, 2019. https://warontherocks.com/2019/10/see-you-in-a-month-ais-long-data-tail/.

[66] Stockfisch, Jack. 1987. *The Intellectual Foundations of Systems Analysis.* RAND Corporation. P-7401. https://www.rand.org/pubs/papers/P7401.html.

[67] "Field Program." Center for Naval Analyses. https://www.cna.org/centers-and-divisions/cna/special-programs/operations-evaluation/field-program.

[68] Stumborg, Michael F. 2018. *Manning, Training, and Equipping US Navy Big Data Analytics Teams.* Distribution authorized to U.S. Government agencies and their contractors. Center for Naval Analyses. DRM-2017-U-016661-Final.

[69] Campbell, Donald T. 1979. "Assessing the Impact of Planned Social Change." *Evaluation and Program Planning* 2 (1): 67-90. doi: 10.1016/0149-7189(79)90048-X. https://www.sciencedirect.com/science/article/abs/pii/014971897990048X?via%3Dihub.

[70] Rodamar, Jeffery. Nov. 28, 2018. *There Ought to be a Law! Campbell vs. Goodhart.* The Royal Statistical Society. https://doi.org/10.1111/j.1740-9713.2018.01205.x.

This report was written by CNA's Systems, Tactics, and Force Development Division (STF).

STF focuses on systems and platforms at the tactical level of warfare, providing classical warfare analyses to help the US Navy and Department of Defense win the great power competition while meeting other warfighting requirements to simultaneously deter and defeat lesser threats. The division's mission includes analyzing and assessing alternative combinations of networks, sensors, weapons, and platforms to provide maritime warfighting capabilities in all warfare areas under realistic employment conditions for current operations and future force architectures.

Dedicated to the Safety and Security of the Nation

CNA is a not-for-profit research organization that serves the public interest by providing in-depth analysis and result-oriented solutions to help government leaders choose the best course of action in setting policy and managing operations.

COP-2022-U-033385-Final

3003 Washington Boulevard, Arlington, VA 22201
www.cna.org 703-824-2000