"When a Measure Becomes a Target, It Ceases to be a Good Measure"

Christopher Mattson, MD
Reamer L. Bushardt, PharmD, PA-C, DFAAPA
Anthony R. Artino Jr, PhD

Imagine you are leading a residency program at a large academic medical center, and the program is preparing for the annual Accreditation Council for Graduate Medical Education (ACGME) Resident/Fellow Survey. You are concerned that 80-hour workweek violations have recently occurred and will be reported to the ACGME. You email the residents one month before the survey to announce forthcoming schedule changes to decrease residents' current workload. You also mention that an ACGME citation for work hour violations could have major negative consequences for the program and recruitment efforts. On the day of the survey, most residents respond by answering "never" or "almost never" when asked about the frequency of work hour violations.

In the 1970s, British economist Charles Goodhart described the pitfalls of measuring the effectiveness of fiscal policy based on monetary growth targets. What is now known as Goodhart's law is most often generalized in a quote from anthropologist Marilyn Strathern, "When a measure becomes a target, it ceases to be a good measure."[1] In its original form, Goodhart's law stated, "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."[2,3] What was initially a jocular aside has become a widely disseminated and universally applicable idea.[4] For learners, teachers, clinicians, and scholars, Goodhart's law speaks to a fundamental truth in health professions education. In particular, the practice of targeting measures and then using them to assess learners and evaluate programs, even when the measures are no longer credible, is quite pervasive in graduate medical education (GME).

Our goal in this editorial is to revisit Goodhart's law and related ideas from other fields and to provide strategies that can be used to mitigate the undesirable effects of this law. Our hope is that those involved in GME will thoughtfully discuss the unintended consequences of measures used as targets and seek to continuously improve their programs' assessment and evaluation practices.

DOI: http://dx.doi.org/10.4300/JGME-D-20-01492.1

Related Ideas and GME Examples

The principle underlying Goodhart's law is not limited to economics. Numerous scholars have published similar ideas in other fields, including social scientist Donald T. Campbell. A pioneer of experimental and quasi-experimental study design methods, Campbell noted, "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."[5] Campbell's and Goodhart's ideas challenge GME faculty tasked with developing and maintaining assessment models and operating programs under existing, often mandated, evaluation systems.

In GME, examples of the potential (negative) consequences exist widely. In the opening example, we highlight a known situation in which responses on the ACGME Resident/Fellow Survey are targeted by the ACGME to serve as a measure of work hour compliance or noncompliance. Program directors are aware of how their residents' responses are used, which creates pressure to coach residents on how best to respond. As a result, noncompliance with work hour regulations may go undetected. By targeting this measure, the ACGME is influencing program director and resident behavior in a way that may distort the measure itself, which renders the measure less useful for its intended purpose. As a result, the validity of the decisions being made based on the measure may be negatively affected by "corruption pressures."

United States Medical Licensing Examination (USMLE) Step 1 scores are often used by residency program directors when screening resident applications and ranking residents. Step 1 scores assess medical knowledge and are used as a surrogate for overall applicant quality. This practice is well known to medical students, who focus a significant amount of time and effort on preparing for the USMLE Step 1. The scores then begin to represent this increased focus, including the amount of dedicated study time and access to test preparation resources, rather than learned medical knowledge and future potential. This

focus also comes at the expense of other learning activities, such as studying for local course examinations, actively participating in small group and peer-learning activities, or developing clinical skills.[6,7] Ultimately, the targeting of USMLE Step 1 scores by GME faculty influences medical student behaviors in ways that may negatively affect their preparation for residency and practice.

Finally, the fixation in academia on "number of publications" and journal impact factor is also felt in GME research environments.[8] Department chairs and promotion committees use these numbers to help make appointment and promotion decisions. As such, faculty are incentivized to focus on the *quantity* of papers published, and the reported quality of journals, erroneously measured by the flawed journal impact factor, over the *quality* of the research itself. Focusing on these targets is widely known to encourage suboptimal research methods.[9] It also adds pressure to engage in other questionable research practices such as "salami slicing"[10] and honorary authorship, both of which are common in health professions education research.[11] In the TABLE, we provide additional examples of Goodhart's and Campbell's laws in action.

Mitigating Unintended Consequences

GME faculty should anticipate negative consequences when specific measures become targets. Recognizing the unintended consequences is the most important step; this can stimulate important discussions when developing assessment and program evaluation plans. Likewise, it is vital to consider how these negative effects might be mitigated. Said another way, we should consider what behaviors will be rewarded given the system that currently exists.[12] A logic model is a common planning tool that is useful in identifying rewarded behaviors.[13] Logic models depict the relationships between program activities and intended effects. Such a tool graphically depicts the shared interactions between the resources, activities, outputs, outcomes, and impact of a program. Through detailed analysis of a logic model, GME faculty can identify unintended consequences and corruption pressures that might distort the processes and outcomes they intend to monitor and improve.

Selecting criterion-referenced over norm-referenced assessments is another strategy to mitigate Goodhart's and Campbell's laws in action. For example, mastery learning techniques have been described as "an instructional approach in which educational progress is based on demonstrated performance, not curricular time. Learners practice and retest repeatedly until they reach a designated mastery level."[14] Instructors and curriculum designers focus on determining the knowledge, skills, and attitudes that are needed for individual success, rather than focusing on ranking individuals relative to one another. Competency-based frameworks are an example of applied mastery learning, and competency-based assessment systems have shown promise in identifying individuals who are struggling.[15] The focus on learning and finding struggling learners rather than identifying the highest performers should be a primary goal in GME. Criterion-referenced assessments also help to eliminate some of the competition incentives that may exist among peers who are accustomed to functioning within more traditional assessment systems.

An additional, albeit controversial, strategy that focuses on criteria over norm-referenced outcomes is the use of a lottery for medical school admissions.[16] By defining specific criteria necessary for success in medical school and using them as entrance criteria to the lottery, there may be less pressure on applicants to attempt to inflate their metrics beyond these thresholds.

GME faculty can also fortify their assessment and evaluation systems with a focus on the processes of learner and program growth versus specific time-point outcomes. This approach has been described in medical education in the context of "thinking longitudinally and developmentally."[17] It challenges faculty to move beyond *how* an individual or program performs (eg, "the first-year resident performs at the level of a senior resident") and towards *why* an individual or program performs the way they do (eg, "the first-year resident shows an ability to independently review personal practice data and improve practice, and also leads health care team discussions of complex patients").

Finally, avoiding overreliance on "the numbers" in assessment and evaluation can mitigate some of the effects of Goodhart's and Campbell's laws. This idea has been previously discussed through the lens of *avoiding the quantitative fallacy* in GME.[18] Numbers are quite limited in the range of competencies that they can completely capture. Further, as noted by Cook, et al, "Numeric scores are inherently limited to capturing attributes and actions prospectively identified as important."[19] In contrast, narrative assessments allow faculty to uncover information that might not have been intentionally sought or otherwise discovered. Because narrative approaches do not reduce complex behaviors or activities into a numerical surrogate, they provide a means to identify and explore nuance and context.

Along with the movement away from numeric assessments and evaluations comes the need to acknowledge and embrace subjectivity.[20,21] This approach encourages faculty to welcome the

TABLE
Examples of Goodhart's and Campbell's Laws in GME

Measure	Used By	Intended Use	Distortions/Corruptions
First-time board pass rate	▪ US specialty boards for Accreditation Council for Graduate Medical Education (ACGME) reviews	▪ Knowledge and skills delivered to trainees ▪ Quality of didactic and workplace-based teaching provided by residency programs ▪ Preparation of residency graduates for independent practice	▪ Selecting applicants with good standardized test skills ▪ Failing to recognize the growth of individual residents relative to their prior standardized test performance
ACGME Resident/Fellow Survey	▪ ACGME	▪ Clinical learning environment in residency/ fellowship programs ▪ Compliance with rules/ regulations ▪ Monitoring of programmatic improvement over time	▪ Coaching on "appropriate" survey responses ▪ Losing the capability to detect areas for improvement
Number of publications	▪ Department chairs ▪ Appointment, promotion, and tenure committees	▪ Scholarship quality ▪ Broad impact	▪ Promoting so-called "salami slicing" ▪ Awarding honorary authorship ▪ Valuing quantity over quality
United States Medical Licensing Examination Step 1 scores	▪ Residency/ fellowship programs	▪ Medical knowledge ▪ Applicant readiness for residency	▪ Studying/teaching to the test ▪ Focusing on medical knowledge at the expense of other clinical competencies
Medical student performance evaluation (MSPE) letters	▪ Residency programs	▪ Residency applicants' clinical skills and personal attributes	▪ Avoiding negative comments by MSPE authors ▪ "Sugarcoating" evaluations ▪ Using "code words"

complexity and messiness of narrative assessments. Qualitative research approaches and narrative assessments are inherently rich, are harder to manipulate, and can produce credible decisions.[19,22] Narrative assessment often requires multiple observations to ensure complete construct sampling. When multiple observations are used for a quantitative measure, one marker of the measure's quality is the lack of variability between iterative measurements. Individuals or programs can change their behavior such that the same outcome is achieved every time. The existence of a single "right answer" to be achieved every time explains why Goodhart's and Campbell's laws are particularly relevant in the context of quantitative measures. However, when multiple observations are used for a narrative-based measure, the measure's quality is determined by differences that are elucidated through different perspectives. The lack of a single expected outcome renders narrative comments much more difficult to manipulate.

Summary

The implications of Goodhart's and Campbell's laws are now appreciated beyond their original contexts in economics and the social sciences. Risks exist in assessment and evaluation systems that rely on quantitative social indicators to inform social decision-making.[5] These concepts are relevant to GME, as demonstrated by the above examples. GME faculty are encouraged to recognize potential problems and take steps to prevent or minimize harms from Goodhart's and Campbell's laws in action. These approaches include: discuss the potential unintended consequences

of quantitative measures as you plan your assessment and evaluation system; apply a logic model or other structured approach in the design of your learner assessment and program evaluation efforts; consider criterion-referenced (over norm-referenced) assessments; and embrace subjective, narrative approaches to learner assessment and program evaluation.

References

1. Strathern M. 'Improving ratings': audit in the British University system. *Eur Rev*. 1997;5(3):305–321. doi:10.1017/s1062798700002660.
2. Goodhart C. Problems of Monetary Management: The UK Experience. In: *Papers in Monetary Economics*. Sydney: Reserve Bank of Australia; 1975.
3. Goodhart C. Monetary Relationships: A View from Threadneedle Street. In: *Papers in Monetary Economics*. Sydney: Reserve Bank of Australia; 1975.
4. Chrystal KA, Mizen PD. Goodhart's Law: its origins, meaning and implications for monetary policy. In: Mizen P, ed. *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart, Volume 1*. United Kingdom of Great Britain and Northern Ireland: Edward Elgar; 2003.
5. Campbell DT. Assessing the impact of planned social change. *Eval Program Plann*. 1979;2(1):67–90. doi:10.1016/0149-7189(79)90048-X.
6. Ronner L, Linkowski L. Online forums and the "Step 1 climate"—perspectives from a medical student reddit user. *Acad Med*. 2020;95(9):1329–1331. doi:10.1097/ACM.0000000000003220.
7. Chen DR, Priest KC, Batten JN, Fragoso LE, Reinfeld BI, Laitman BM. Student perspectives on the "Step 1 climate" in preclinical medical education. *Acad Med*. 2019;94(3):302–304. doi:10.1097/ACM.0000000000002565.
8. Fire M, Guestrin C. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *Gigascience*. 2019;8(6):giz053. doi:10.1093/gigascience/giz053.
9. Smaldino PE, McElreath R. The natural selection of bad science. *R Soc Open Sci*. 2016;3(9):160384. doi:10.1098/rsos.160384
10. Eva KW. How would you like your salami? A guide to slicing. *Med Educ*. 2017;51(5):456–457. doi:10.1111/medu.13285.
11. Artino AR, Driessen EW, Maggio LA. Ethical shades of gray: international frequency of scientific misconduct and questionable research practices in health professions education. *Acad Med*. 2019;94(1):76–84. doi:10.1097/ACM.0000000000002412.
12. Kerr S. On the folly of rewarding a, while hoping for b. *Acad Manag J*. 1975;18(4):769–783.
13. Frye AW, Hemmer PA. Program evaluation models and related theories: AMEE Guide No. 67. *Med Teach*. 2012;34:e288–e299. doi:10.3109/0142159X.2012.668637.
14. Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM. Setting mastery learning standards. *Acad Med*. 2015;90(11):1495–1500. doi:10.1097/ACM.0000000000000887.
15. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. *Med Educ*. 2014;48(6):614–622. doi:10.1111/medu.12408.
16. Mazer BL. Accepting randomness in medical school admissions—the case for a lottery [published online ahead of print October 17, 2020]. *Med Teach*. doi:10.1080/0142159X.2020.1832206.
17. Holmboe ES, Yamazaki K, Hamstra SJ. The evolution of assessment: thinking longitudinally and developmentally. *Acad Med*. 2020;95(11 Suppl):7–9. doi:10.1097/acm.0000000000003649.
18. Carmody JB. On residency selection and the quantitative fallacy. *J Grad Med Educ*. 2019;11(4):420–421. doi:10.4300/JGME-D-19-00453.1.
19. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words—validity evidence for qualitative educational assessments. *Acad Med*. 2016;91(10):1359–1369. doi:10.1097/ACM.0000000000001175.
20. Ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. *Acad Med*. 2019;94(3):333–337. doi:10.1097/ACM.0000000000002495.
21. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–568. doi:10.3109/0142159X.2013.789134.
22. Ginsburg S, Van Der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment—a quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617–1621. doi:10.1097/ACM.0000000000001669.

Christopher Mattson, MD, is a Pediatric Resident, Comer Children's Hospital, University of Chicago; **Reamer L. Bushardt, PharmD, PA-C, DFAAPA,** is Professor and Senior Associate Dean for Health Sciences, George Washington University School of Medicine and Health Sciences; and **Anthony R. Artino Jr, PhD,** is Professor and Interim Associate Dean for Evaluation and Educational Research, George Washington University School of Medicine and Health Sciences, and Deputy Editor, *Journal of Graduate Medical Education*.

Corresponding author: Anthony R. Artino Jr, PhD, George Washington University School of Medicine and Health Sciences, aartino@gwu.edu, Twitter @mededdoc