To appear in Theory and Society

Interrogating the "cargo cult science" metaphor[1]

Andrew Gelman[2] and Megan Higgs[3]

19 Feb 2025

Abstract. Over the past fifty years, the term "cargo cult" has been used to describe the actions of scientists who appear to follow forms of scientific inquiry but without the understanding and self-criticism that are essential to real scientific progress. The term has served a useful role by providing a short and catchy label to something that is otherwise difficult to explain. However, the term is also fraught with historical and cultural baggage and, in our opinion, encourages crossing of a subtle line between criticizing discipline methodological norms and criticizing the individuals currently carrying out those norms as part of a complex and context dependent social process. We find that carefully interrogating the term itself holds some important lessons for improvement in the science reform movement.

Keywords: ritual science, science reform, sociology of science, statistics

Background

In 1974, physicist Richard Feynman used the term "Cargo Cult Science" to describe practices that look like scientific investigation in form, but are missing something essential, which he describes as a type of "utter scientific integrity" that leads scientists to work diligently to not fool themselves, or others (at least in the domain of scientific inquiry). The analogy is to groups of people living in isolated Pacific islands after the end of the Second World War who apparently followed ritualistic practices in an unsuccessful attempt to replicate the cargo deliveries that had been delivered by the Allies when using these islands as airstrips during the war. Feynman was not the first to compare these behaviors to ritualistic scientific practices, but his use of the term likely helped secure Cargo Cult a place in popular culture and motivated related extensions by swapping out Science for Programming, Software Engineering, and more recently, Statistics (McConnell, 2000, Stark and Sartelli, 2018). The term "cargo cult science" points to over 2500 papers on Google Scholar in fields including biology, political science, computer science, agriculture, psychology, philosophy, policy analysis, education research, dentistry, management, religious studies, medicine, artificial intelligence, along with physics and statistics. The term is also used frequently in news outlets, blogs, and social media.

"Cargo cult" has become a catchy phrase that elicits a resonant image of islanders in the South Seas creating non-working replicas of runways, traffic control huts, and headphones in hopes of obtaining goods via airplane, based on memories of goods delivered during World War II. But the history, usage, and opinions on cultural appropriateness of the term are complicated and still

[1] We thank Pamela Reinagel, Bruce Mannheim, Susan Gelman, and Samuel Bowles, the editor, and two reviewers for helpful discussion and feedback.
[2] Department of Statistics and Department of Political Science, Columbia University, New York.
[3] Critical Inference LLC, Bozeman, Montana.

unsettled (e.g., Jarvis, 2019). The science and statistics references are now commonly used in discussions of science reform, and our impression is that most users of this phrase, including ourselves (Gelman, 2017), have been largely unaware of the analogy's problems. There is irony in the use of the term "cargo cult" in this context, as scientists continue to employ it without interrogating its meaning enough to avoid fooling themselves into using an analogy that may be inadequate or misleading. As statisticians who are involved in the science reform movement, we see an opportunity for self-criticism, to learn about the analogy itself and about the practices it has been assumed to capture. It has long been recognized that much can be learned by interrogating our metaphors to reveal aspects of our underlying framing (Lakoff and Johnson, 1980).

What is cargo cult science meant to capture?

Before going further, we take a closer look at what Feynman actually said in his now famous speech (Feynman, 1974). He appealed to the analogy to help express concerns over "science that isn't science," particularly referring to some research in education and psychology. He presented a simplified view of a cargo cult in the "South Seas" where the people saw airplanes deliver materials during the war and wanted the same thing to happen again. "So they've arranged to make things like runways . . . to make a wooden hut for a man to sit in, with two wooden pieces on his head like headphones and bars of bamboo sticking out like antennas—he's the controller—and they wait for the airplanes to land. . . .The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things Cargo Cult Science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land."

It's easy to end here and fill in what might have been missing in a variety of settings, which gets us to a common interpretation of the phrase as capturing the idea of practitioners going through motions that appear important but do not ultimately deliver the end they purport to be a means for. It may be that the people carrying out the methods believe they will work and have not sufficiently interrogated them to learn about why they don't work, or it may be more of a ritual carried out as part of cultural or social practice without any real expectation of success. This interpretation focuses on the practices and not the reasons for the practices or the individuals carrying them out.

Feynman admits it is hard to tell us what's missing: "it would be just about as difficult to explain to the South Sea Islanders how they have to arrange things so that they get some wealth in their system. It is not something simple like telling them how to improve the shapes of their earphones." But, he does identify one feature: "scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards." He goes further by saying "Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can—if you know anything at all wrong, or possibly wrong—to explain it. . . . In summary, the idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another."

Feynman conveys something we agree is vital to science, but it's quite a jump to get from the image of the South Seas cargo cult to the substance of his speech. The implication seems to be that the members of the cult did not possess the type of integrity scientists are expected to operate under, they didn't work hard enough to bend over backwards to show how they might be wrong, and perhaps they ignored information given to them about the shortcomings of their methods. Setting aside questions of historical accuracy, the members of Feynman's cargo cult lacked access to various aspects of modern science and technology, but there is no reason to believe they were being dishonest or lacking integrity within their own cultures. To put it another way, all the honesty, openness, and scientific integrity in the world wouldn't have allowed them to leap the technological gap, any more than one could have expected Benjamin Franklin's electrical experiments to allow him to force a quick end to the Revolutionary War by building an atomic bomb.

Problems with the metaphor

We see problems with the "cargo cult" metaphor—not only in its historical and cultural insensitivity, but also in that it is in many ways misleading. There is an important difference between commenting on practices that people carry out as part of a social group (e.g., building fake runways) and commenting on the character of the individuals involved (e.g., honesty and integrity), and we expect the distinction affects interpretation and usefulness of the analogy in terms pointing to problems or suggesting reforms. A focus on personal character can quickly lead to feelings of disrespect and defensiveness which may take focus away from understanding the underlying reasons for problematic practices. The members of the cargo cults may have lacked access to various aspects of science and technology available to those in industrialized nations, but there is no reason to believe they were dishonest or lacking integrity within their own cultures. To put it another way, all the honesty, openness, and scientific integrity in the world wouldn't have allowed them to leap the technological gap.

Setting aside the historical inaccuracies, the analogy also implies that the practitioners of what is called cargo-cult science or engineering are somehow technologically behind or lacking access to technology that would otherwise help them achieve the goals of doing science. But the people and practices we are talking about are not associated with less or inferior technology; the problems identified are more about how technologies are used, rather than lack of access. Again, the cargo cult analogy falls short.

We assume that the historical reasons for cargo cults are complicated (political, social, and cultural). Likewise, we assume reasons for researchers engaging in practices labeled as "cargo cult" are complex; our colleagues across the sciences are struggling to balance the goal of learning from data while not being fooled, as well as working to survive, or thrive, in current scientific culture. Jumping to labeling them or their practices as "cargo cult" does not seem a productive path forward. To the extent that the "cargo cult" expression reflects a dismissive attitude of science reformers toward working scientists, perhaps exploding the metaphor will reduce this unproductive mode of thought on the part of ourselves and others in the science reform movement. This work thus fits into the reflective attitude toward scientific reform as expressed by Devezer et al. (2020). It requires something industrialized societies don't have the

best track record for—listening, respecting cultural differences, and learning from each other through gaining awareness.

Application to statistical methods

We and others have used the "cargo cult" epithet to disparage unfortunately common statistical practices such as the proclamation of discoveries based on statistical significance tests, the use of statistical models that are not justified by data collection or the underlying science (Gelman, 2017), and more generally the "mechanical, ritualistic application" of methods, where practitioners "invoke statistical terms and procedures as incantations, with scant understanding of the assumptions or relevance of the calculations" (Stark and Saltelli, 2018).

An example of such a ritualistic application is researchers copying the superficial trappings of a statistical procedure (such as computing a p-value or a Bayesian posterior probability calculation) without implementing, or even understanding, the fundamental assumptions on which the validity of the procedure is predicated—or worse, without even asking what is needed for the method to be valid and reliable enough to use in practice. The practices typically being criticized are uses of scientific or statistical procedures that do not actually put one's scientific hypotheses to the test but rather are used to confirm a pre-existing belief (consider the all-too common formulation, "We designed an experiment to prove . . .") and perhaps to reject a straw-man null hypothesis. Social or cultural pressures, such the need for strong preliminary results for a grant proposal, are part of the story, whether we like it or not.

In many natural and social sciences, formal training in statistical inference is absent or, at best, limited to teaching numerical or computational procedures for pushing data through tests, typically long after experiments have been done and maybe even after conclusions have been reached, and even more quantitative fields such as economics often rely on cookbook methods. This is inevitable and, to a large extent, desirable, in that it would be a great loss if serious research could only be conducted by teams with statistical expertise. It would be a mistake to propose a "science reform" that would require innovative statistical design, data collection, or analysis for each new study.

Indeed, users of default statistical procedures have a sincere and serious goal: making valid discoveries and avoiding being fooled by spurious coincidences. And although there may be statistical reasoning and methods that could assist them toward this goal, the options that readily appear in initial training or previous publications may not. This common situation stems from an honest lack of training and knowledge coupled with misguided social pressure, not from blatant lack of scientific integrity. However, there may be some point at which enough questioning of current practices from others outside the field should sound an alarm and send a scientist to interrogate the methods they have used, or intend to use, to the point they understand the foundations and assumptions relative to their own work. We, as statisticians or proponents of science reform, should be doing our part to instigate the alarms, assuming integrity and honesty of those employing the questionable practices. Labeling the people or practices as "Cargo Cult" is unlikely to be productive.

This then raises the question of how to sound the alarm—how to respect the need for researchers to use routine methods while also giving them the tools to interrogate and criticize their own work in the way that Feynman associates with good science. The Cargo Cult analogy has been a convenient way to get attention paid to problems—but it's not clear that it does more good or harm.

At this point we could just laugh (or cringe) at our cultural myopia—and the obvious ways in which cargo cult metaphor does not apply to science—and move on. But in simply abandoning the epithet entirely, we might miss an opportunity to learn more about why it has been so popular and accepted. Exploration of the metaphor can be an entry point into thinking more clearly about scientific practice and, from our perspective as statisticians, why it has seemed so fitting when used to describe statistical misconceptions and misuse of particular methods. There is much agreement that an alarm is needed, but we would like to more accurately diagnose and address the aspects of statistics, and the scientific method more generally, that get in the way of openness and self-criticism. Some problems are obvious—for example, the traditional publication of summary statistics but not raw data—while others are more subtle, such as the practice of dichotomizing inferences based on statistical significance or other thresholds, which both discards information and moves researchers toward an inappropriate sense of certainty (Greenland, 2017).

Moving from "cargo cult" to more specific statistical concerns

Thoughtless use of statistical methods is a serious problem across the sciences and engineering. However, for the reasons given above, we do not think "cargo cult" is a good label here. The term "social rituals," as used by Gigerenzer (2004) seems more appropriate, as this does not imply a technological gap between the misusers of methods and their developers, but rather a mindlessness: "repetition of the same action . . . wishful thinking and delusions that virtually eliminate critical thinking." The problem is not that credentialed scientists are behaving like ignorant South Sea islanders but that they follow rules that could make sense in social settings ranging from religion to politics but that have no place in the laboratory.

Gigerenzer's article is focused on the "null ritual" of "the magical 5% number" and "wishful thinking about . . . the p-value," but one could similarly talk about rituals in survey research, biostatistics, Bayesian inference, machine learning, software developers, and any area where applied workers rely on templates to guide their processes.

Indeed, one could ask where to draw the line between "rituals" (bad) and "defaults" (good). Statistics is said to be the science of defaults, with the properties of a statistical method defined by averaging over hypothetical replications (frequentist inference) or a hypothetical prior distribution (Bayesian inference). More generally, templates and recipes are an essential part of science: we would not expect a single researcher, or even a single research team, to come up with a new theory, design and build their own instruments, and collect and analyze their data using methods devised just for this new problem. The problem with "ritual science" or "ritual statistics" is not the use of standard methods but rather a lack of understanding of the conditions under which the methods would be appropriate—or a lack of willingness to consider or interrogate the limitations of the methods being used.

What does the metaphor capture well?

In a perhaps too optimistic look at the metaphor, we might take away a message of what can happen when we (in a very general sense) simply don't have the information and motivation necessary to step back and see why something isn't working—instead, it can be easy to just keep making more non-working headphones and runways in response to pressure from people in positions of power and incentive systems. Such momentum, both for individual researchers and for scientific fields as a whole, makes it hard to stop long enough to ask if and how we might be fooling ourselves and where the assumptions we take for granted might fall apart. In science, it's often not obvious that the valuable goods are not arriving, particularly if some goods do appear to be arriving to benefit the careers of individual scientists (e.g., increasing chances of publication by reporting small p-values or using the most popular complex model even if not appropriate or needed for the context). What are the signs that practices aren't working? This isn't an us vs. them interpretation; it applies to all.

There are many statistical methods that are widely used, widely misunderstood, and deserve criticism, with the common thread being that these methods are used ritualistically or mechanically and often based on assumptions that are not acknowledged or assessed in a meaningful way. For a familiar example, the observed errors of election polls are about twice as large as would be predicted from the theory of random sampling (Shirani-Mehr et al., 2008), but the usual analyses of surveys ignore this nonsampling error.

Given the problems, both cultural and descriptive, with the cargo cult analogy, one might ask why it has been so often used to label problematic practices in some disciplines. Why speak of "cargo cult science" rather than, say, "cookbook science" or "black-box science" or some other metaphor capturing mechanistic application of procedures that do not meet standards of quality and lack adequate transparency or introspection? We argue that the metaphor has had value beyond the perhaps-amusing conjuring of images of teams of lab-coated Ph.D.'s as members of technologically unsophisticated societies. "Cargo-cult" goes beyond "cookbook" etc., in conveying that the processes being employed are not just automatic and poorly understood, but also that they don't work at all. To put it another way, if the problem is "cookbook science," this can be solved by awareness of the limitations of existing recipes and understanding the principles underlying them, so they can be creatively altered to develop new or improved dishes. Similarly, one can go beyond the "black box" by understanding its internal workings. Neither of these solutions are easy, but they assume the "cookbook" and "black box" methods do align with the larger goals, or at the very least are not a detriment to those goals. The cargo cult epithet, despite its problems, may better capture the complexity of the inherently social parts of the problems and the challenges to finding solutions. Ironically, the term may not yet be well recognized within many social sciences.

Another appeal of the cargo-cult metaphor is the sense of ritual it conveys. Ritual has been described as having predefined sequences characterized by rigidity, formality, and repetition that are embedded in a larger system of symbolism and meaning but contain elements that lack direct instrumental purpose (Hobson et al., 2018). These features are present for many of the practices that have been labeled as cargo cult, such as publication decisions based on p-values or default

use of a method regardless of clear violations of assumptions. A ritualistic attitude does seem to describe much of the work in the natural and social sciences that we find troubling.

One complicating factor is that there can be a pressure for even routine science to be presented as revolutionary. This is the scientific surprise two-step: a finding is presented as a newsworthy surprise, causing us to reassess our understanding of the world, and at the same time as consistent with a well-established scientific theory with a rich literature. Similarly when applying for grants, researchers describe their work as uncharted territory that is also eminently reachable. This tension is implicit in statistical terms such as hypothesis testing and false discovery rate, which if taken literally would imply that scientists are regularly overturning hypotheses and making discoveries. We bring this up here to emphasize that there is controversy not just in the processes of science but in its glamorization of discovery.

What does the metaphor say about how to fix the problem?

Feynman even points out the lack of a simple solution, even in the context of the "cargo cult" story: "it would be just about as difficult to explain to the South Sea Islanders how they have to arrange things so that they get some wealth in their system. It is not something simple like telling them how to improve the shapes of the earphones." Nor would the problems be solved by real runways and better earphones, as these would still not bring the deliveries. If members of industrialized societies took the time inquire and try to understand the underlying motivations and goals, they would likely find something much more complicated than simply wanting aviation technology: perhaps more productive agricultural methods or some other way to escape from colonialism. Or maybe the members of the cargo cult just want to be left alone. The members of industrialized society would also have to look honestly at their past actions that may have contributed to the existence of the cargo cult. We think there is something helpful in this deeper look when thinking about how to make progress on some of the problems with use of statistical methods.

For example, asking experimentalists or policy analysts about the goals of their research might reveal that the common approaches used, such as making a binary decision about "significance" of a result from a single study, are not aligned well with their scientific goals or even their actual research workflow. It may be that deviations from more appropriate statistical methods are serving a function in helping them avoid other more pervasive types of errors. Or they maybe have never been exposed to statistical approaches that would have real benefit to their research, as opposed to jumping through hoops of "doing statistics" such as carrying out null hypothesis significance tests that have little relevance beyond that of satisfying journal reviewers. Addressing the larger goals of working scientists and policy analysts should happen first, and a one-size-fits-all solution should not be expected. We conjecture that approaching a science reform effort from this collaborative perspective increases the chance that any changes made are actually improvements, are authentically understood by and thus reasonably employed by the practitioners, and are welcomed as a resource to be embraced, rather than resented as an impediment to be endured.

We are concerned that some people—including us—who have criticized science, or scientists, using the "cargo cult" analogy envision a solution in which scientists build better earphones or

runways, without solving underlying problems. Something that looks like an easy solution to an outsider (such as a statistician or scientist in another field) might not actually solve the underlying problem. In statistical terms, short-term solutions such as multiple-comparisons adjustments can represent potential improvements in carrying out the mechanics of an expected method without solving the problem of continued ritualistic use and without interrogation by the researcher of how the method might work (or not) to help accomplish the goal. We do not see easy-to-adopt improvements, or superficial replacements, in the methods themselves as a useful first step.

What does interrogating the metaphor offer to improve science reform?

Our goal in this paper is not to reform science so much as to achieve the more moderate goal of reforming science reform. Maybe it is not clear how "cargo cult" practices hinder the puzzle-solving processes that are central to science—especially social science. Many researchers believe that good work is being done and successes are happening, despite a layer of "cargo cult" labeled practices being embedded in the process, so it's not as simple as just a label for distinguishing between science and pseudoscience. For example, repeated and uninterrogated use of particular statistical methods, despite advice against them from statisticians, is an example of poor statistical practice showing up in science, but the use of such approaches labeled as "cargo cult" does not imply nothing of use is coming from the entire scientific process; it depends on how much the final conclusions depend on the practices. Questionable statistical practices may not be advancing science in the way we believe (or say) they are, but the extent to which the larger scientific endeavor is then called into question depends on how much it relies on the practices.

Our larger point here is that, when thinking about science reform, we should consider other scientists' worldly goals (publications, funding, esteem, and ability to do the interesting parts of science with minimal amounts of statistical "paperwork") but also their omnipresent (if sometimes implicit) goals of using data to suggest, refine, evaluate, and compare scientific hypotheses without being fooled by noise or biases.

We can consider three goals of empirical science: to make discoveries, to confirm/disconfirm scientific hypotheses, and to rule out alternative explanations. From this perspective, problematic statistical practices (such as the uncontrolled use of p-values or Bayesian inferences based on poorly calibrated models) are unsuccessful attempts to attain these goals—often without knowledge of how (un)successful they are. Criticisms, by ourselves and others, of questionable research practices tend to focus on ways in which these methods lead to overconfidence: strong claims which do not hold up over time and in retrospect are often based on shaky theory: hypotheses that are "more vampirical than empirical—unable to be killed by mere evidence" (Freese, 2007).

How does this relate to cargo cults? If the goal of cargo cult members was to obtain material goods, the associated practices have two key failures: first, a lack of good theory and, second, the practices do not actually work for the stated goal. That is, the analogy implies a disconnect between methods and goals, which is inappropriate in the setting of science and use of inferential statistics. Wooden headphones and simulated runways have no direct connection (in theory or

practice) to planes landing with cargo to distribute. However, they may serve other purposes, such as creating social connection or making a political statement, in the same way that poor statistical practices do not facilitate real discoveries but can allow their practitioners to get grants, publish, and climb the academic ladder: people are following the practice despite lack of understanding of its underlying principles or assumptions, either because they feel social pressure to or because they believe that it works, without being clear on what is meant by "it works." Just as anthropologists study social practices on their own terms and also by considering what they do for their practitioners, we can study statistical practices in the same way.

Perhaps it helps to consider that just about all of us are living much of our lives in a way that could be labeled as "cargo cult"—technologically (as when we drive a car or watch television with only the vaguest understanding of how these appliances work), logistically (in our reliance on an unseen supply chain to keep the power running and bring food to our supermarkets every day), and in the roles we play in society. Scientists rely on external authority when leaving their own fields of expertise.

Better practice in quantitative research is not just about improving our use of statistical tests (advice akin to saying "change the shape of the headphones" or "build bigger runways" in the cargo cult narrative) or even abandoning tests entirely ("fake headphones and runways serve no purpose at all"); it also involves studying real effects, controlling and adjusting for variation, and carefully integrating statistical models with scientific models. It involves a deeper understanding of available methods and when they are well aligned with the bigger goals of the science. In today's scientific culture, long-established principles of measurement and statistical design are often forgotten, or inadvertently ignored, because statistical significance and publication can be bought with the cheap coin of uncontrolled statistical analysis.

Improving statistical practice can be expected to reduce the rate of absurd claims that appear to be supported by data, but the real gains should arrive indirectly by incentivizing researchers to put in the effort to improve design, measurement, data collection, and the mapping of scientific theories to realistic models of data. Those pushing science reform efforts could benefit from respecting these challenges, rather than labeling other scientists, or their practices, as "cargo cult" and offering easy diagnoses that, while well-meaning, miss the mark.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Devezer, B., Nardin, L. G., Baumgaertner, B., and Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLoS One 14, e0216125.

Feynman, R. P. (1974). Cargo cult science. Commencement address, California Institute of Technology. http://calteches.library.caltech.edu/51/2/CargoCult.pdf

Freese, J. (2007). The problem of predictive promiscuity in deductive applications of evolutionary reasoning to intergenerational transfers: Three cautionary tales. In Caring and Exchange Within and Across Generations, ed. A. Booth et al. Washington, D.C.: Urban Institute Press.

Gelman, A. (2016). The problems with p-values are not just with p-values. American Statistician 70.

Gelman, A. (2017). It's not enough to be a good person and to be conscientious. You also need good measurement. Cargo-cult science done very conscientiously doesn't become good science, it just falls apart from its own contradictions. Statistical Modeling, Causal Inference, and Social Science blog, 21 Sept. https://statmodeling.stat.columbia.edu/2017/09/21/not-enough-good-person-conscientious-also-need-good-measurement-cargo-cult-science-done-conscientiously-doesnt-become-good-science-just-falls-apart-ow/

Gigerenzer, G. (2004). Mindless statistics. Journal of Socio-Economics 33, 587-606.

Greenland, S. (2017). The need for cognitive science in methodology. American Journal of Epidemiology 186, 639-645.

Hobson, N. M., Schroeder, J., Risen, J. L., Xygalatas, D., and Inzlicht, M. (2018). The psychology of rituals: An integrative review and process-based framework. Personality and Social Psychology Review 22, 260-284.

Jarvis, B. (2019). Who is John Frum? Topic Magazine (22), https://www.topic.com/who-is-john-frum

McConnell, S. (2000). Cargo cult software engineering. IEEE Software (March/April), 11-13.

Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018). Disentangling bias and variance in election polls. Journal of the American Statistical Association 113, 607-614.

Stark, P. B., and Saltelli, A. (2018). Cargo-cult statistics and scientific crisis. Significance 15 (4), 40-43.