Volume 7 | Autumn 2019

DEFENCE STRATEGIC COMMUNICATIONS

The official journal of the
NATO Strategic Communications Centre of Excellence

ISSN: 2500-9486
DOI: 10.30966/2018.RIGA.7

THE ELEPHANT IN THE ROOM: MEASUREMENT OF EFFECT

A Review Essay by Gary Buck

The Book of Why
Judea Pearl, Dana Mackenzie. Basic Books, 2018

Measuring Social Change: Performance and Accountability in a Complex World
Alnoor Ebrahim. Stanford Business Books, 2019

Keywords—*Measurement of Effect, causality, strategic communications, information operations, assessment strategy*

About the Author

Dr Gary Buck is an Applied Psychologist and heads the Research and Insight team at M&C Saatchi World Services. He has served in military IO roles in Afghanistan, Bosnia and Iraq.

DOI 10.30966/2018.RIGA.7.6.

Introduction

Measuring effect is a little like obeying the speed limit when driving; everyone agrees in public that it's a good idea, but no one seems to do it. I am not convinced it is as hard as it seems. As a practitioner of PSYOPs, Information Operations, and Strategic Communications for the last twenty years, both in military and civilian capacities, I have thought long and hard about the issue. Reading two excellent books on the subject has prompted me to re-engage with it. *The Book of Why,* by Judea Pearl and Dana Mackenzie, seeks to debunk the idea that it is too difficult to attribute causality when measuring effect. Alnoor Ebrahim's *Measuring Social Change: Performance and Accountability in a Complex World* also examines this issue, along with various other questions, such as whether to measure goals in the short or long term. In this, I hope to broaden the discussion around the state of play of Measurement of Effect (MoE). I refer to these authors when considering which objectives to measure and when to measure them. I also address the issue of attributing causality to communication campaigns and conclude by describing an approach to MoE, currently under development, that I believe will allow us to both qualitatively and quantitatively assess the impact of context on campaign outcomes.

To structure this exploration, I shall employ Rudyard Kipling's 'six honest serving men'—*What and Why and When and How and Where and Who*—to ask the following questions:

Why should we measure effect? *What* should we measure? *When* should we undertake measurement? *Whom* should we measure? *Where* should we measure? *How* do we know we've had an effect?

In each instance I shall address points of received wisdom, so often cited with regard to Measures of Effect, which I am not convinced withstand detailed scrutiny. Specifically, I shall address the following claims:

- MoE is only or primarily about evaluation
- Campaign activity is reported by activity type
- Evaluation is about big studies that have a midline and an endline assessment

- Longitudinal studies are the only or best study design
- It is too difficult to measure online activity
- Statistical significance is both a necessary and sufficient condition
- It is impossible to attribute causality

These are big issues and I have attempted to be provocative to stimulate what I hope is a useful and informative debate.

Why measure effect?

The first and perhaps most fundamental question to be addressed is why should we be measuring effect? This answer also speaks to the first point of received wisdom: 'MoE is about evaluation.' We obviously conduct measures of effect to assess the outcomes from our campaigns. However, the thinking often stops there—too often MoE is seen as a process that occurs at the end of (and possibly midway through) a campaign to justify actions by showing some effect being achieved on the ground.

There are two issues on this point. First, we should be trying to extend our thinking in terms of justifying and assessing the Return on Investment (ROI) from campaigns. This is more than merely justifying actions by pointing to some effect being achieved on the ground. Those involved in measuring effect should also factor in the cost of conducting the activity as well as accord some value to the benefit the effect delivers.

The second reason to measure effect—and one often overlooked—is to test and adjust. To determine the extent to which we are making progress towards achieving the desired outcome, we should measure effect at different points during the campaign. By doing so we can both demonstrate incremental progress towards the end state and identify early what isn't working and take corrective action. When we look under the bonnet of a campaign to see what is happening, it is helpful to have a model against which to structure our thinking. As a starting point, we can turn to a model that should be familiar to most practitioners and customers of Information Operations—the OODA Loop (Figure 1 below).

DOI 10.30966/2018.RIGA.7.6.

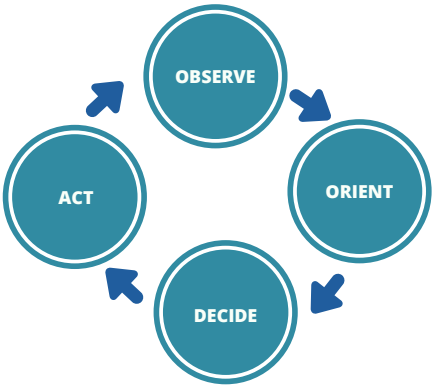

Figure 1. The OODA Loop

The OODA Loop is a decision-making model developed by USAF Colonel John Boyd, who saw the process as a cycle involving four steps: Observe, Orient, Decide, and Act. Boyd's model provides a starting point for assessing a target audience's (TA) progress towards a desired effect.

At each stage of the loop we use a specific type of measurement to assess progress regarding a key objective in the process.

Once we make a baseline assessment of the target audience and deliver the campaign message, the three key objectives around which we can build our MoE framework are:

- *knowledge transfer*—the extent to which the TA is aware of the desired behaviour
- *attitudinal shift*—the extent to which the TA believes it is correct to engage in the desired behaviour
- *behavioural change*—the extent to which the TA actually engages in the desired behaviour

Considering Measurement of Effect through the lens of the OODA Loop means we can structure our assessment around the changes in *knowledge*, *attitudes*, and *behaviours* (KAB) the TA will need to make if we are to achieve the desired effect. Breaking down the desired outcome of the campaign into specific objectives at each stage allows us to define the content of a measurement framework.

OODA Loop stage	Key objective	Measure of Effect
Observe	Message Exposure	Performance (MoP)
Orient	Knowledge transfer	Understanding (MoU)
Decide	Attitudinal shift	Outtake (MoO)
Act	Behavioural change	Impact (MoI)

What should we measure?

By thinking about the behavioural change process in terms of the OODA Loop, we can answer the question: what should we be measuring? Undertaking an assessment at each stage of the loop gives us four different Measurements of Effect:

- Observe—Measure of Performance (MoP)
- Orient—Measure of Understanding (MoU)
- Decide—Measure of Outtake (MoO)
- Act—Measure of Impact (MoI)

Measures of Performance (MoP) show the extent to which the TA has been exposed to the message or campaign and has actually perceived it. These measures are relatively simple to collect (and, unfortunately, are often mistaken for Measures of Effect in their own right). Measuring performance allows us to demonstrate progress in the early stages of a campaign—we can show that we have reached the TA. If we are not seeing an effect, we must determine the cause—has the TA not been reached and, if so, why?

MoP's should be familiar to everyone. Typically, MoP's are reported and included in the MoE analysis on the basis of the nature of the output, i.e. x number of TV commercials aired, y number of impressions on social media. The problem is that it makes the analysis more difficult in that it is hard to equate campaign activity with outcomes. I would argue that rather than reporting MoP's at the

overall level, campaign activity should be reported thematically against objectives. It then becomes possible to tack the impact of greater activity against different objectives. In this way, the evaluative comparison can be made in a more direct fashion and it becomes easier to assess what objectives the Target Audience has been exposed to.

Assuming we have reached the TA sufficiently at the Observe stage, and following the OODA Loop round, we can assess the extent to which the exposed members of the TA have understood the message at the Orient stage. Measures of Understanding (MoU) are essentially about the transfer of knowledge from sender to recipient. If the target hasn't received the message and understood it, it will not be processed correctly. Again, measuring at this point can demonstrate further progress and determine if there is a problem with comprehension of the message.

Continuing the cycle, Measures of Outtake (MoO) assess how far the TA has decided to act on the information contained in the message. The decision that is reached involves, to a greater or lesser extent, a shift in stance regarding the issue in question. Thus, MoO measures shifts in attitudes. Measuring at this point provides further demonstration of progress and can identify problems—namely, has the mindset of TA shifted in the right direction? Assuming it has, the final step is for the TA to change its behaviour and to act in line with the desired outcome. This is the impact that we should be seeking to achieve and so the actual desired behavioural change can best be measured with Measures of Impact. This is our final, evaluative assessment.

Using this structure, we can both conduct evaluative assessments and demonstrate progress towards the desired objectives. And, crucially, this structure allows us to diagnose problems along the way. Understanding why we should measure effect helps us understand how often we should measure, and that takes us to the next question: when should we be measuring?

When should we measure?

The implication of measuring effect to demonstrate the progress of a campaign or test and adjust its direction is that it must be assessed early and regularly.

This is where the dictum 'fail fast' is often applied. Use this phrase with senior officers in the military or account managers in a civilian agency and they turn pale, spluttering about not wanting to fail. I think the term misses the point. A much better phrase would be 'learn quickly'; you don't necessarily need to fail to learn valuable lessons. Learn quickly better encapsulates the notion that we need to measure early and often throughout the lifetime of a campaign.

This point addresses one of the themes Ebrahim discusses in *Measuring Social Change*. He presents a contingency framework for assessment comprised of four different measurement strategies (niche, emergent, integrated, ecosystem) that he argues should be adopted under different circumstances (low uncertainty/low control, high uncertainty/low control, low uncertainty/high control, and high uncertainty/high control, respectively). The choice of strategy is driven by the degree of certainty one has about the causal relationship between campaign activities and intended outcomes, and the ability to control those outcomes. A key point he makes is that under circumstances of high uncertainty and low control, which is the case for most strategic communications campaigns, it is preferable to measure in the short-term as high uncertainty about cause-effect makes it difficult to understand the best course of action for achieving desired outcomes.

Ebrahim argues for flexibility. We need agility in our approach to assessment. Measuring early and often can provide crucial new information and make it possible to readjust the course of the campaign. This refutes the next point of received wisdom, that 'evaluation is all about big studies that have a midline and an endline assessment'. Comprehensive assessment once or twice in a big campaign does not provide the agility needed to make best use of the assessment process. Although these large-scale assessments are important, there is also a need for more frequent assessments of smaller samples that provide quicker feedback. I agree with a colleague's assertion that an: 'N of 40 gives you 80% of the answer'. It is at this figure that an unbiased sample begins to approximate a normal population (and thus begins to lend itself to the assumptions about distribution of error upon which statistical testing is based). Using smaller sample sizes is, in a way, more rigorous in that larger (more meaningful) effects must be demonstrated for them to be statistically significant. I am not suggesting that all sample sizes should contain only 40 subjects. Should be used smaller samples to produce more frequent assessments

DOI 10.30966/2018.RIGA.7.6.

to provide the agility necessary to learn quickly and respond flexibly in low control contexts. I will return to the issue of statistical significance later.

Another reason why measuring more frequently makes sense is the high reliance on social media within populations and the increasing use of online platforms to deliver digital content within influence campaigns. The 'always-on' media environment, within which we are measuring messaging, is a rapidly changing context. Therefore, we need quicker assessments to keep pace and get inside the TA's OODA Loop.

We do not have the luxury of being able to conduct large-scale baseline, midline, and endline assessments spread months apart. The environment is changing too quickly. Any time gap between assessments leaves room for circumstances and events to change that significantly impact a target audience and thus confuse the picture. Greater access to information and a rapidly changing media environment pose challenges that can be answered by more frequent assessment and, possibly, by changing whom we assess.

Whom should we measure?

I would like to challenge the default setting for measurement approaches—the longitudinal study. This is the approach recommended in the US Department of Defense (DoD) handbook on measurement, but I suspect it is too restrictive and does not correspond to the reality on the ground.

I propose a flexible approach to study design and making use of Quasi-Experimental Designs where we measure a cross-section of samples exposed and not exposed to our message. Because measurements take place at the same time, we can form judgements about the effectiveness of a campaign with greater speed, allowing us to 'learn quickly'. Because there is no intervening time period, making more direct comparisons reduces the potential impact of confounding events or variables. The difficulty arises with finding a suitable comparison group. But provided samples are matched across relevant demographic factors this should not pose a problem. Going one stage further, adopting a 'mixed method' strategy—combining a multiple sample (exposed/non-exposed) approach with a

longitudinal design—means we could track changes in knowledge, attitudes, and behaviour of the exposed TA against the control group across time as well.

The dynamic nature of the media environment, especially the pace of audience reaction on social media, demands greater agility around how often and at which stages to measure. Target audiences' use of social media requires greater consideration when measuring the effectiveness of campaigns.

Where should we measure?

A conventional wisdom is the perception that it is too difficult to measure effect online. However, given that a great deal of campaign activity is delivered online and audiences spend much of their time engaging with this space, it is important to assess effect in this domain. Problems centre on using automated sentiment analysis. Social media analytical tools are often criticised for not being able to identify and factor in sarcasm and irony. This adds up to a lack of assessment in the online space when we should also be measuring effect here.

One approach to measuring effect online is to measure shifts in narratives used by a target audience in relation to the campaign theme. The TA can be members of the public in the Area of Operations or, indeed, the Adversary. This approach means viewing a shift in narrative as an instance of behavioural change—the TA has changed its behaviour and shifted the narrative it uses in the online domain. To achieve this, we need to break down the narrative into more measurable units; namely, its component parts:

- *Narrative*—the overarching position or explanation stated by the target audience
- *Themes*—key points or topic areas that, when combined, make up the overall narrative
- *Stories*—specific events or accounts used by the TA to illustrate the themes being discussed

DOI 10.30966/2018.RIGA.7.6.

Breaking down the narrative in this manner, then measuring changes at the story level, would make an assessment more robust and easier to conduct. Shifts at the story level could then be assessed along two dimensions: change in topic and change in sentiment. For example, a Strategic Communications campaign might be focused on highlighting the desirability of its target audience engaging in community activities as a way of fostering a return to normalcy after a period of conflict. If members of the TA who had been exposed to the campaign message began discussing security concerns less frequently and instead started to relate stories about getting involved in community activities, they could be assessed as having shifted their narrative in line with the campaign objective. This would be a shift in topic. Shift in sentiment can also be used as a measure of effect. For example, a shift from being against the unification of a nation to being open to the concept could be assessed as a measure of effect.

It is possible to measure effect online. Indeed, any measurement framework should include both offline and online assessment. Regardless of where measurement takes place, it is vital to determine if there has been an effect and the extent to which a campaign was responsible.

How do we know we've had an effect?

I've left the biggest and perhaps trickiest issue until last. This subject is discussed in Ebrahim's *Measuring Social Change* and forms the core theme of Pearl and Mackenzie's *The Book of Why*. There are two issues of interest: the focus on statistical significance and the assumption that it is impossible to attribute causality.

Assessment studies typically identify and report on results that are statistically significant. This is right and proper. If a result is not statistically significant there is an unacceptable chance—greater than 5% probability—that it might have occurred through an error in the measurement process. More specifically that the sample is not representative of the population as a whole and, if another sample were taken, it might yield a different result. Problems arise when a statistically significant result is taken to be a meaningful result; this is not necessarily the case. A result may be significant statistically but not tell us anything meaningful. This is because the calculations used to estimate statistical

significance are largely dependent on sample size. The greater the sample size the smaller the effect (the difference between two groups on an opinion survey question, for example) required for the result to be statistically significant.

It is important to look not only at the significance level, but at the meaningfulness of the result. We should not confuse significant with meaningful. Statistical significance is a necessary but insufficient condition for reporting a result; what does the result tell us about the impact of our campaign? Have we made clear progress towards the desired end state? In other words, have we seen a real effect?

Even if we do see an effect, perhaps the most difficult question to answer is what caused it. Was it our campaign? Was it outside factors? Both? How do we attribute causality? The received wisdom is that we can't attribute causality to a campaign. I believe we can.

Pearl and Mackenzie introduce a number of useful concepts when considering this issue, such as the Ladder of Causation and the use of causal diagrams to map the flow of causality. Their ideas are useful and stimulating. However, I would like to come at this from a different perspective. To answer the problem of causality, we must determine the extent to which an observed effect is driven by campaign activity and/or by possible confounding variables. To address this issue, we need to monitor the operating environment during a campaign to measure changes in the context that might have made an impact on the target audience. For example, we might be running a campaign to increase a target audience's sense of agency and optimism for the future in a particular country. One impact indicator might be a survey question asking how optimistic respondents feel. After a comparison of baseline and endline assessments, it may turn out that the TA has become more optimistic, which might lead us to think that our campaign has been effective. However, an upturn in the economy and concurrent rise in the employment rate may also have had a positive impact on the TA and their survey responses.

My team currently monitors the environments within which our campaigns are operating across five different factors (that are similar to the PMESII factors): Social, Military, Economic, Political, and Physical. They create daily information summaries, which are collated at the end of the week, and an assessment is

made of any changes in the environment that may have an impact on campaign objectives or sub-objectives. The table below provides examples of each of the five factors.

Contextual factor	Examples
Social	An increase in social unrest is likely to make the target audience less optimistic, whereas a period of stability or even rapprochement would most likely contribute to increasing feelings of optimism for the future.
Military	A reduction in levels of insurgency following the defeat of a terrorist organisation would most likely improve the target audience's view of the future.
Economic	A prolonged period of security and stability might lead to an upturn in the economy as jobs are created through foreign investment/support for local businesses. Increased employment would help increase feelings of optimism.
Political	Frequent changes in government or a series of scandals would create a sense of instability and thus reduce levels of optimism, while a period of good governance would have the reverse effect.
Physical	Changes in the environment, such as harvest failure or the adverse impacts of climate change, would have a negative impact on the target audience's feelings of optimism for the future.

This approach clearly raises questions about reliability and rigour in the assessment process. There is devil in the detail surrounding the robustness of assessments made. However, these are made against a set of clearly defined factors and based on a structured rating scale using a range of verified data sources. The assessments are then 'Red Teamed' through internal peer review and subjected to external confirmation.

This may seem like an overly complicated approach that requires a lot of effort. It certainly requires effort; however, the daily commitment produces a

secondary benefit in terms of situational awareness. Insights from the analysis are provided directly to the teams running the campaigns in order to help them identify 'hot topics' that they might wish to feature or avoid.

Weekly assessments are used to evaluate how changes in the environment might impact outcome measures. At the moment, whilst we are still collecting data, our assessments are conducted on a qualitative basis. For example, if the employment rate had fallen and security had declined during our campaign to increase the TA's sense of agency and optimism for the future, but the optimism rating for our target audience had still gone up, this would provide some justification that the effect was due to our campaign; or at least that a greater proportion of the effect was due to the campaign messaging. In this way, we can provide some context for our reporting.

A more interesting and useful possibility is to look at this quantitatively. Weekly assessments produce a numerical rating that evaluates the permissiveness of the environment in terms of our campaign messaging. In this way, we generate a fifth Measurement of Effect: Measure of Context (MoC)—currently a work in progress. If we were to generate a quantitative assessment for the confounding contextual variables (the MoC), we could then consider this along with our Measures of Performance (gathered as part of our ongoing monitoring of the campaign) and create a statistical analysis capable of apportioning causality to various factors within the MoE. We could then quantify the agency of the various factors contributing to the observed outcome (e.g. our campaign is responsible for 58% of the observed outcome, whilst other factors, such as economic factors, contributed 42%).

We are developing a statistical model on one of our campaigns to test this hypothesis; the results won't be available before this publication goes to press. Initial results are encouraging. I can of course be contacted directly. If we are able to measure the operating environment or the context in this way, we can attribute causality in a quantitative manner and help to answer one of the fundamental questions about the practice of measuring effect. This, along with the other ideas I have put forward, might help transform Measurement of Effect from the elephant in the room to the elegant swan gracing the lake of Strategic Communications.

DOI 10.30966/2018.RIGA.7.6.

Conclusion

I have employed Kipling's 'six honest serving men' (somewhat loosely) to ask some fundamental questions about Measurement of Effect and hope I have challenged some outdated received wisdom on the subject. My suggestions in response to these points can be summarised as:

1. Conduct MoE assessment for diagnostic reasons as well as evaluation.
2. Report and assess Measures of Performance by campaign themes.
3. Learn quickly—be more agile by conducting smaller studies more frequently.
4. Combine QED studies with longitudinal designs.
5. Measure shifts in narratives online.
6. Look for meaningful, not just statistically significant, differences.
7. Measure the context to control for confounding variables.

My hope is that this discussion has stimulated some thoughts, whether provoking agreement or disagreement. If I have, then I will have achieved my effect; now, how to measure it?