Editor: **Ciera Jaspan**
Google
ciera@google.com

Editor: **Collin Green**
Google
colling@google.com

What Makes a Good Productivity Metric?

Collin Green, Ciera Jaspan, and Majed Samad

// In this article, we describe some criteria we consider when building a productivity metric, illustrating with an example of a novel metric for work schedule consistency that we built in the aftermath of the COVID pandemic //

IN A PRIOR column, we wrote about how measuring productivity can be viewed as a form of modeling and that all models are wrong, but some are useful. That discussion centered on the idea of ensuring that a productivity model was inclusive of multiple metrics and that those metrics covered the various facets of productivity and covered each facet reasonably well. In that article, we set aside the question of what makes a good *individual* productivity metric that can be combined with others into a (hopefully) useful model of productivity. In this article, we'll share some things we consider when building an individual metric, including an example of a novel metric we built in the aftermath of the COVID pandemic.

A Metric for Work Location Consistency

Back when studying hybrid work was cool, our team sought to understand how office usage patterns related to engineering productivity metrics. Previously, we relied on a scheme borrowed from product engagement: we defined "office engagement" as the regularity of working from an office. But we were unhappy with that approach: it conflates the frequency of working from the office with the consistency of one's working schedule. Discussions with stakeholders revealed that many people had hypotheses about whether and how frequency and consistency each mattered and—importantly—that leaders setting policy around hybrid work wanted to understand whether their policy should explicitly address each. Quantifying the frequency of office use was easy for us to make and easy for stakeholders to understand. Quantifying the *consistency* of office use was not. In this article, we're going to explore what makes for a useful metric, using this "consistency of office schedule" metric as an example. We explored 13 candidate metrics to capture "office consistency," and we landed on "Weekly Markov Consistency"—the metric described in the sidebar. As the reader can see, it's a rather complex metric. So why this metric?

What Makes a Good Metric?

Each case of metric design is different, but there are a few criteria we pay attention to as we proceed with metric design and selection:

1. The metric is intuitive to a broad audience.
2. The metric reflects patterns that we understand to exist in the world.
3. The metric doesn't reflect patterns that we know don't exist in the world.
4. The metric is useful in decision making.

We'll unpack these criteria in the context of our effort to quantify the consistency of engineers' work schedules to understand whether and how the consistency of working location affects productivity.

A Good Metric Is Intuitive

A good metric has to help stakeholders understand the world: its design should not be difficult to understand in the context of the business. That doesn't mean it has to be simple, but it does mean that there should be a reasonably accurate nontechnical explanation of the metric that stakeholders can understand and reason about. Things that may make a metric less comprehensible include arbitrary or uninterpretable parameter values or an ill-defined or seemingly arbitrary range of possible values (e.g., a metric that ranges from 0.33

Digital Object Identifier 10.1109/MS.2026.3662031
Date of current version: 14 May 2026

WEEKLY MARKOV CONSISTENCY

After exploring 13 candidate metrics for "consistency of office schedule," we settled on a metric we call "weekly Markov consistency." This metric was inspired by Vasilescu et al. (2016)[1]: we adapted the quantity known as Markov entropy to measure week-to-week consistency. The intuition around a Markov process is that it describes a time-varying random variable for which the probability of its value at the next time point only depends on its value at the current time point, $p(x_{t+1} \mid x_1, x_2 \ldots x_t) = p(x_{t+1} \mid x_t)$ where $p(x_t)$ is the probability of being in state x at time t.

To address week-to-week variation, we define the process across a moving window that is 7 days wide. As there are three possible states on each day (work from office, work from home, not working), we have a 3x3 matrix that determines the transition probabilities for all weekly transitions (i.e., transitions from day t to day t + 7). The formula for the Markov entropy of that 3x3 matrix is:

weekly Markov entropy

$$= -\sum_{i=1}^{3}\left[p_i \sum_{j \in \pi_i} p(j \mid i) \log_2 p(j \mid i)\right] \quad (1)$$

where p_i is the fraction of days the person spent in state i, $p(j \mid i)$ is the conditional probability of a weekly transition from state i to state j (defined as the i, j element in the 3x3 matrix divided by the sum of the ith row), and π_i s the set of outgoing states for state i (i.e., the ith row of the 3x3 matrix).

The quantity in Equation 1 produces values ranging from 0 to $-\log_2(1/3)$, which can be seen by considering that the quantity is maximized when the Markov process has the greatest degree of uncertainty, which occurs when the 3x3 matrix is fully uniform since that means that all transitions are equally likely. In that case, the $p(j \mid i)$ terms will all equal 1/3. The inner summation in Equation 1, therefore, produces $\log_2(1/3)$. Assuming further that the distribution of states across the time window is also uniform, we have that $p_i = 1/3$ for all i. The outer summation therefore yields $-\log_2(1/3)$ as the maximum that Equation 1 can take.

We use this fact to normalize the metric by dividing it by $-\log_2(1/3)$, and further subtract it from 1 to invert its meaning from a measure of disorder to a measure of order, so that it becomes a viable metric for schedule consistency. The final formula is therefore:

weekly Markov consistency

$$= 1 - \frac{\sum_{i=1}^{3}\left[p_i \sum_{j \in \pi_i} p(j \mid i) \log_2 p(j \mid i)\right]}{\log_2\left(\frac{1}{3}\right)}. \quad (2)$$

Не самая простая, но и не очень сложная

to 1.7 is worse than one that ranges from 0 to 1). Composite metrics or metrics with complexities like weighting or nonlinearity are also difficult to reason about: if a composite metric's value changes, one has to immediately ask about which underlying components changed.

While none of our 13 candidate metrics are mathematically simple (including the one we chose!), there is variety in how intuitive they are. For example, we can provide a relatively simple and accurate description of weekly Markov consistency as "The metric shows how easy it is to predict an engineer's work status on any day based on their status the same day in the prior week." Even more simply "If an engineer does the same thing every Wednesday, that's a consistent schedule." Contrast this with another candidate metric we examined, one based on a fully connected multilayer neural network (with one hidden layer and three output units) that predicts working state from the prior 28 days. There is no plain language description of what such a metric is doing aside from "It uses a statistical model to predict work status from the past 28 days of status." This explanation isn't intuitive and doesn't aid reasoning.

Чем-то напоминает сложность по Колмогорову

Weekly Markov consistency is strong in other aspects of intuitiveness, too. The metric's two parameters reflect the length of week and the number of possible working states. It includes no other weights or thresholds. In contrast, the neural network candidate metric has many parameters corresponding to network connection weights that are difficult to interpret practically. Weekly Markov Consistency also has the desirable property that it ranges from 0 to 1.0 for minimum

and maximum consistency, respectively. The neural network metric shares this property, but other candidate metrics did not (e.g., Gini Index autocorrelation ranges from 0.8 to 1.0; Cramer's V autocorrelation ranges from 0.3 to 1.0).

An intuitive metric, even if it is mathematically complex, supports stakeholder confidence and adoption.

> The metric shouldn't be merely a proxy for some other factor; if it is, there's no substantial value in retaining it.

A Good Metric is Right in Important Ways

It's important that a metric can make expected "table stakes" predictions. For example, outlier cases should map to extreme values of the metric and typical cases should fall in the middle of the metric's range of values. Ideally, the metric should change smoothly and monotonically between extremes. Metrics should show known relevant seasonal patterns. The metric should also not be too volatile; no one wants to spend excessive time discriminating signal from noise.

For our office consistency metric, we came up with a series of validations to ensure that it captures expected patterns. For example, we explored how candidate metrics responded to "mutations." We took a uniform schedule (all in the office) and introduced increasing amounts of noise, randomly mutating one day at each increment. A metric that is "right in important ways" would show a smooth monotonic decrease as schedules mutated from uniform to completely random. Weekly Markov consistency passed this test while several other candidate metrics did not.

We observed sensible seasonal movements in the metric: Months where engineer show lower weekly Markov consistency are those where schedules are known to be perturbed (i.e., in June, November, and January we observe lower median consistency: U.S. employees show more variation in how they take time off in these months and that's reflected in the metric).

We also evaluated whether weekly Markov consistency reflected human judgment. To do this, we sampled 50 schedules (five per decile of the distribution) and asked eight team members to rate them as either *consistent* or *inconsistent*. This exercise revealed two important things: there was a significant positive correlation between human ratings of consistency and the metric's value; and we were able to estimate the "point of subjective equality" for the metric. That is, we found that schedules with a weekly Markov consistency around 0.42 are equally likely to be labeled as consistent or inconsistent. This also corresponds to the metric's value in our mutation testing exercise when about half the mutations had been introduced, giving us an extra layer of confidence that this metric is right in important ways.

A Good Metric Isn't Wrong in Important Ways

It's important that the metric not have undesirable properties (no "deal breakers"). A metric can be right in a whole bunch of important ways and still fail to land with a stakeholder if it shows troubling anomalies. For example, if a metric shows values that make no sense in the real world (e.g., if you're measuring the duration of a process that must take time, a duration of 0 is impossible) then it will be hard for stakeholders to trust the metric. The metric should yield sensible values for all interesting cases. Additionally, the metric shouldn't be merely a proxy for some other factor; if it is, there's no substantial value in retaining it.

We evaluated our candidate metrics against some negative cases: we looked for patterns that would cast doubt on the proposed metric. For example, we checked whether candidate metrics suggested unrealistic behavior from engineers. While humans can be unpredictable in some ways, most engineers don't have wildly unpredictable schedules. If most engineers were labeled "inconsistent" by our metric, that would give us pause about its ability to describe the world and be trusted by stakeholders. Therefore, we checked the value of the metric for the actual engineering population at Google. While there is variation across individuals and over time (see previous regarding seasonal patterns), the median engineer is labeled as consistent by weekly Markov consistency.

We also worried that the metric might be biased toward specific

kinds of schedules. We didn't want to see different values for uniform schedules of different types (purely work-from-office, purely-hybrid, etc.). We directly examined whether each metric assigned its maximum value to a perfectly consistent schedule, irrespective of the specific nature of that schedule. Weekly Markov consistency assigns a value of 1.0 to a uniform schedule that reflects a 5-day in-office work week with weekends off and it assigns that same 1.0 value to a uniform schedule that reflects a 3-day in-office hybrid work week. Other metrics favored specific types of consistent schedule and were therefore rejected. For example, a sample entropy metric assigned higher values to uniform schedules that were entirely work-from-home or entirely work-from-office relative to uniform hybrid schedules.

Finally, we started this project in an attempt to disentangle the frequency of office use from the consistency of office use. If our metric is highly correlated with the frequency of office use or with the frequency of working at all, that would make it less valuable for both research projects and policy-making stakeholders. Luckily, we found a low correlation between weekly Markov consistency and the rate of working-from-an-office; the correlation with overall working days was even lower. This indicates we are not measuring a related concept that might be a poor proxy for our intended use.

While there may be other anomalies in how weekly Markov consistency behaves that we've not yet discovered, evaluating against these patterns helped us feel confident that the metric was useful and that stakeholders would trust it.

ABOUT THE AUTHORS

COLLIN GREEN is the user experience research lead for the Engineering Productivity Research team at Google, Mountain View, CA 94043 USA. Contact him at colling@google.com.

CIERA JASPAN is the software engineering lead for the Engineering Productivity Research team at Google, Mountain View, CA 94043 USA. Contact her at ciera@google.com.

MAJED SAMAD is a quantitative user experience researcher on the Engineering Productivity Research team, Google, Seattle, WA 98109 USA. Contact him at majs@google.com.

Главный вывод статьи и он действительно интересный

A Good Metric is Useful In Decision Making

Of course, we never set out to construct a metric just to have one. We make metrics to serve a purpose, whether that be to identify opportunities for improvement, evaluate the impact of interventions, or more deeply understand a phenomenon so that we can inform our next actions. In this case, we wanted to understand the productivity impacts of consistent and inconsistent work schedules, how consistency interacts with other factors in driving those impacts, and what such findings imply about the policy for hybrid work. There's no room in this article to unpack all our findings, but we'll give you the punchline: consistency in work schedule is at least as important (probably more important) as the frequency of working in any particular location. For example, we observed that engineers who worked consistent hybrid schedules (1–3 days in office per week) had CL throughput indistinguishable from those that worked consistent schedules entirely in the office. Importantly, those engineers that worked in the office one day a week or less and scored as consistent had CL throughput indistinguishable from those that worked most days in the office but held inconsistent schedules. There are many factors that affect how an individual behaves, but these results sum up what we see in a population of tens of thousands of engineers in our data set: consistency matters to productivity

(and so this metric is useful for our stakeholders).

Good Metrics and Validity

In sum, we think weekly Markov consistency is a good and useful metric because it's intuitive, it reflects several patterns that we know to exist in the world, it doesn't show anomalous behavior that would undermine confidence in it, and it's useful for decision making.

It's worth noting that the aforementioned considerations are related to formal concepts of validity that are commonly discussed in behavioral and social science research. For example, intuitiveness is related to the notion of *face validity*.[a] The extent to which a metric does and does not reflect the world as we know it is related to *convergent validity*[b] and *discriminant validity*.[c] Evaluating whether a metric makes useful predictions is related to *criterion validity*.[d] The principles underlying these formal concepts are helpful in metric design or selection even when specific statistical concepts don't apply.

Hopefully, our discussion of the consistency metric makes clear some considerations for designing and evaluating a productivity metric. Every case is a little different and domain expertise, human judgment, and practicality must play a role in the decision making around metrics. As we've written previously,[2] measuring productivity is an exercise in model building and all models are wrong. Thinking carefully about principles like those described here will hopefully enable readers to build models that are useful.

[a] https://en.wikipedia.org/wiki/Face_validity

[b] https://en.wikipedia.org/wiki/Convergent_validity

[c] https://en.wikipedia.org/wiki/Discriminant_validity

[d] https://en.wikipedia.org/wiki/Criterion_validity

References

1. B. Vasilescu et al., "The sky is not the limit: Multitasking across GitHub projects," in *Proc. IEEE/ACM 38th Int. Conf. Softw. Eng. (ICSE)*, Austin, TX, USA, 2016, pp. 994–1005, doi: 10.1145/2884781.2884875.
2. C. Jaspan and C. Green, "Measuring productivity: All models are wrong, but some are useful," *IEEE Softw.*, vol. 42, no. 2, pp. 13–17, Mar./Apr.2025, doi: 10.1109/MS.2024.3511735.

Digital Object Identifier 10.1109/MS.2026.3685763