A Watched Pot Seems Slow to Boil:

Why Frequent Monitoring Decreases Perceptions of Progress

Authors' note: Materials, data, and analysis code can be accessed online:

https://osf.io/kt8rj/?view_only=3722dd9d1edd442aab4fc2b24a20de27.

Word count: 10,084

Abstract

In evaluating endeavors, struggles, and accomplishments, perceivers care not only about targets' end products or states, but also the speed with which they have progressed. Two employees are likely to have different value in the eyes of their supervisor if they take different amounts of time to complete the same amount of work. Targets may be frequently monitored (e.g., subject to daily check-ins) or rarely monitored (left to proceed with only occasional inspection). The present work explores a *monitoring frequency effect* (MFE) whereby progress is seen to slow to the extent that it is monitored more frequently. Five studies in the main manuscript and two in the Supplemental Materials combine to document the MFE, identify causal antecedents that shape preferences for monitoring, demonstrate that the monitored can have counterproductive preferences for how frequently they are monitored, and distinguish between two mechanistic accounts to explain this effect. Although the present work identifies challenges in how perceivers track targets' progress across time, such biases did not explain the MFE. Instead, perceptions of progress seemed to disproportionately weight observed incremental progress, which will be smaller when the interval between monitoring check-ins decreases, and relatively neglect the amount of time that passes between such check-ins. Discussion focuses on how the MFE complements or superficially contradicts previous demonstrations of the ratio bias, unpacking, and the automatic tracking of features' frequency. The MFE identifies a qualitatively distinct way by which beliefs and expectations color evaluations of targets.

Keywords: monitoring frequency, temporal neglect, goal progress, productivity

A Watched Pot Seems Slow to Boil:

Why Frequent Monitoring Decreases Perceptions of Progress

As people make progress on a lengthy task or goal, such progress is observable over time. Furthermore, people typically have a choice over how often they check in to learn whether and how quickly progress is being made. A conscientious home financial manager will consult their credit card statement to see how quickly their balance is building. An overbearing factory supervisor may make the rounds more often to inspect the latest output. Anxious academics may keep a close eye on their Google Scholar citations, being attuned to the latest upticks in their own metrics.

On the one hand, keeping a watchful eye on change may seem tantamount to being fully in the know. The surest way of not knowing whether one has gained weight over the holidays is to stay off the scale. But in this paper, we instead consider some unforeseen consequences of the frequency with which one tracks progress. Drawing on social psychology, cognitive psychology, and judgment and decision-making literatures, we hypothesize that frequent (vs. infrequent) monitoring encourages the perception that progress is happening more slowly. In the process, we show not only that the watched pot *seems* slow to boil, but also explore why that is the case, identify several determinants of what leads (proverbial) cooks to keep a more watchful eye, and assess whether people readily intuit the broad applicability of this proverb or instead display counterproductive preferences for how frequently they would want to be monitored.

A brief example will both capture the intuition and provide an introduction to our basic paradigm. Consider a factory supervisor whose job is to monitor factoryworkers' output. Every so often, the supervisor stops by a worker's station to surreptitiously check the worker's output, counting how many parts have been completed and placed in an output tray before clearing the tray's contents until the next time they complete an inspection. Even if two workers produce output at the same rate, what the supervisor will see in the tray (a straightforward and salient index of output) will be larger or smaller depending on how frequently they check in. If the supervisor, perhaps more concerned about one of the workers than the other, finds themselves stopping by one worker's station twice as often as the other's, then they will always see half of the output in this frequently monitored worker's tray than they would have seen under the less frequent monitoring schedule. To the extent that the supervisor (at least partially) neglects how much time has passed between check-ins, then they may come to see the more frequently monitored worker as less productive. We call this the *monitoring frequency effect* (MFE).

More generally, we suggest that progress is often monitored by considering how much has changed since the previous check-in. Dieters who do weekly weigh-ins may focus more on how much weight they have lost more than their absolute weights. Public health departments may release the number of new COVID infections instead of simply updating a running tally of total infections. Investors may take note of percentage changes in the market instead of their preferred stock index's current value. PhD advisors may focus more on how many more pages of a draft their advisee sent over instead of on the current length of that draft. Given monitoring progress entails attention to, well, progress—or incremental change—the possibility that people may not draw normative conclusions based on how much progress they observe has potentially broad implications.

Although we are aware of no previous exploration of the monitoring frequency effect, there is precedent for the idea that people neglect scope or duration when making judgments and evaluations. Kahneman (2000) has argued that, in affectively characterizing prolonged experiences, people exhibit temporal neglect and instead engage in evaluation by moments—

essentially how positive or negative an experience is at a single point (or moment) in time. In evaluating the experience of watching aversive film clips (Fredrickson & Kahneman, 1993), undergoing a painful colonoscopy (Redelmeier & Kahneman, 1996), or even prospectively considering a hypothetical experience from a description of its time course (Varey & Kahneman, 1992; Wang et al., in press), people seem relatively insensitive to the length of those experiences. Instead, they focus on a single moment or perhaps combine two moments—for example, leaning on a *peak-end rule*, the average of the most extreme as well as the final moments, to arrive at a summary evaluation (Kahneman et al., 1993). One consequence is that people will remember an interrupted painful experience as worse than an experience whose full duration was excruciatingly experienced, simply because the omitted (i.e., what would have been the additional, final) portion would have been less (but still) painful (Kahneman et al., 1993; Varey & Kahneman, 1992). The present work considers not how people summarize affective experiences, but how they synthesize feedback that is itself delivered across different moments (or check-ins). That people may apply a similar *evaluation by moments* to such contexts—those in which what feedback is delivered at each individual moment will be determined not merely by stable and relevant features of the target (e.g., their objective rate of progress) but by the monitoring frequency of the perceivers—is of course not a foregone conclusion. That said, previous empirical support for evaluation by moments lends plausibility to our proposal.

When previous research has examined how repeated sampling of information can distort judgment, it has typically considered how the *non-independence* of such information can throw things awry. Fiedler (2012) suggested that when people sample information, their judgments fail to take into account the degree of diagnosticity that each additional piece of information carries. Such a failure reflects *metacognitive myopia* (see also Fiedler et al., 2018). As Fiedler et al.

(2002) note, repeated sampling of information is beneficial when each draw is independent (A teacher gets a more accurate perception of their students' satisfaction by reading more students' course evaluations). But repeated sampling of information rarely produces independent draws, and thus it frequently yields redundant information.

Consider more generally how one might arrive at various conclusions. To know whether an opinion is popular, listen for it being expressed. To know whether a business is thriving, be on the lookout for relevant news reports. Problems arise when one hears a statement repeatedly because it is being continually espoused by one opinionated person, or when the positive news coverage one keeps hearing actually reflects many media outlets' reporting on the same event. Perceivers often act as if this actually redundant information is new and informative (Weaver et al., 2007), even when such redundancy is made salient to the perceiver (Begg et al., 1992; Fiedler et al., 2018; Hasher et al., 1977; Unkelbach et al., 2007).

Other research has examined how non-independence thwarts reasonable inferences from information that—like the focus of the current work—describes a target's progress. That work—much like that of Fiedler and colleagues described above—also considered how perceivers are not sensitive to (partial) redundancies that many descriptions of progress include (Alves & Mata, 2019; Alves et al., 2023; Grüning et al., 2023). Consider watching two basketball games in which one team ekes out a victory by 2 points. In one game, the winning team led the whole game. But in the other game, the two teams traded the leading position throughout the game. If you are like Alves and colleagues' participants, you may be more impressed by the former victor. But note that holding a steady lead is merely a commentary on the early-game performance, because only the opening-minutes' baskets factor into the score throughout the game. Points in the final quarter are reflected in the game score only for the final quarter. Absent a (non-intuitive) theory of why early performance is more diagnostic than late performance, attention to who led for the majority of the game reflects a counternormative weighting of redundant information.

Our own focus of study considers how tracking or monitoring progress can lead to counternormative observations, but not due to an inappropriate reliance on redundant information. Instead, we recognize that—much as did the factory supervisor who occasionally dropped by to assess a worker's incremental progress—progress is often tracked by noting changes in output or standing. In this way our own monitoring frequency effect—like Alves and Mata's (2019) cumulative redundancy bias—suggests that people focus and disproportionately weight a salient index of performance. But whereas Alves and Mata found that people fail to fully appreciate the significance of a global marker of progress (i.e., that repeatedly checking in on cumulative progress gives disproportionate weight to early performance), we instead consider how people fail to properly interpret a local marker of progress (i.e., that how much has been accomplished since a previous check-in depends not only on the amount of change that is observed but on how much time has passed).

When previous researchers have considered the frequency of monitoring, they have primarily focused on how often people track *goal* progress—i.e., whether a target (typically, the self) is progressing effectively toward a desired endstate. Furthermore, this research has focused more on determinants of monitoring frequency as opposed to consequences of it (Chang, Webb, & Benn, 2017; Chang, Webb, Benn, & Reynolds, 2017; Chang, Webb, Benn, & Stride, 2017; Webb et al., 2013). For example, people often avoid tracking goal progress, opting to bury their heads in the sand rather than keep close track on how they are faring. Webb et al. (2013) term this "the ostrich problem." When consequences of monitoring frequency have been studied, it has been in terms of actual performance (Jenkins & Terjeson, 2011). Our own hypotheses need not apply to the context of goal pursuit in particular. And although we will identify certain antecedents of monitoring, our focus is on surreptitious monitoring's consequences for perceptions of progress, even when actual progress does not change.

That said, the question of what factors influence monitoring frequency is also relevant for understanding the implications of the MFE. More generally, we suggest that beliefs or expectations about different targets are likely to influence how frequently they are monitored. In that way, the MFE can prove to be a novel mechanism through which prior beliefs or expectations can affect perceptions of others. Previous research on self-fulfilling prophecies (e.g., Carlana, 2019; Figlio, 2005; Gentrup et al., 2020; Hill & Jones, 2021; Papageorge et al. 2020) and confirmation bias (e.g., Vedejová & Čavojová, 2021) have identified different ways by which expectations may influence perceptions. Research on the confirmation bias, for instance, has demonstrated how beliefs and expectations can change what information people seek out and what information they neglect (Klayman & Ha, 1987). The MFE proposes a new way in which these beliefs and expectations can color target evaluations even when target progress is held constant and the actual, objective meaning of the information to which perceivers are exposed is equivalent.

We present five studies that, collectively, document the MFE, identify causal antecedents that influence monitoring frequency, and distinguish among mechanisms that explain why monitoring frequency influences perceptions of progress. Study 1a asked participants about their intentions to monitor different employees more or less frequently in light of different beliefs or expectations about those targets. Study 1b then manipulated the frequency with which workers were monitored to probe implications for perceptions of worker productivity. Study 1c had participants consider how frequently they, as workers, would want to be monitored if their goal were to maximize others' perceptions of their own productivity. Thus, in the context of a single paradigm, we are able to identify determinants of monitoring frequency, the consequences of monitoring frequency (i.e., the MFE), and the potentially counterproductive preferences of those being monitored. Study 2 tested for the MFE in a distinct domain, one in which fast progress is instead negative (i.e., the spread of a contagious disease). Study 3 distinguished two mechanistic accounts of the MFE, whether it emerges due to perceivers' failure to track information about targets or to make use of the temporal information that they do recall. Materials, data, and analysis code can be accessed online: https://osf.io/kt8rj/?view_only=3722dd9d1edd442aab4fc2b24a20de27.

Study 1a

Study 1a sought to establish whether there would be systematicity—based on beliefs and expectations about targets—to how frequently people would choose to monitor targets' progress. Participants were asked to consider being a workplace manager whose job was, in part, to monitor the performance of their employees. Participants considered employees who varied in the importance of their job, the length of time they had spent on a workplace team, and their reputations as particularly good or bad employees. We expected that managers would display a bias toward being particularly vigilant of those engaging in important work, those who were new to the job, and those who were suspected to be slackers. Such differential monitoring preferences would be important—and at times insidious—if Study 1b finds support for the monitoring frequency effect in this context.

Method

Participants. Fifty-five undergraduates (76% female, 20% male, 4% non-binary; M_{age} =

20.16, SD_{age} = 6.15) at a European university participated in exchange for course credit.

Procedure. Participants considered being the manager of a work team. As manager, their responsibilities included monitoring the employees by occasionally checking-in on them. At each check-in, the manager would take a look at how much work the employee had completed since the previous check-in. Participants were also told that they had limited time to monitor their employees; they could not monitor every employee every day.

More specifically, participants learned about three pairs of employees. Participants' task was to allocate 10 check-ins between the two employees in each pair. These 10 check-ins would span two five-day workweeks, during which only one check-in per day was possible. At a minimum, one check-in had to be allocated to each employee. The employees composing each pair differed along different dimensions: the alleged importance of their project, their tenure on the team, and the valence of rumors that spoke to each worker's productivity (see Table 1). The three pairs were presented in a counterbalanced order.

Table 1

Employee Descriptions (Study 1a)

Dimension	Employee	Description
Importance of task	I	"This employee was assigned a particularly important part of the project."
	U	"This employee was assigned a relatively less important part of the project."
Time on the team	N	"This employee just joined the team, so you don't yet know a lot about him."
	L	"This employee has worked for your company for many years, and you have known him for a long time."
Valence of rumors	B	"You have heard negative rumors that this employee can sometimes be slow and inefficient, which makes you nervous about the employee's productivity."
	P	"You have heard positive rumors that this employee is fast and efficient, which makes you confident about the value of this employee as a team member."

Results and Discussion

For each pair of employees, participants tended to allocate more check-ins to one of the employees over the other. More specifically, participants allocated more check-ins to the employee who was completing a particularly *important* part of the project (I; $M = 6.69$, $SD = 0.77$) instead of a relatively *unimportant* part of the project (U; $M = 3.31$, $SD = 0.77$), paired $t(54) = 16.36$, $p < .001$, $d = 2.21$. They also allocated more check-ins to the *new* employee (N; $M = 6.91$, $SD = 1.22$) instead of the *longtime* employee (L; $M = 3.09$, $SD = 1.22$), paired $t(54) = 11.59$, $p < .001$, $d = 1.56$. Finally, participants assigned more check-ins to the employee about whom *bad* rumors were swirling (B; $M = 6.40$, $SD = 1.20$) compared to the one about whom there were *positive* rumors (P; $M = 3.60$, $SD = 1.20$), paired $t(54) = 8.68$, $p < .001$, $d = 1.17$.[1] Each of these findings suggests that monitors are unlikely to allocate their monitoring time evenly or randomly among targets. We next turn to whether such differential monitoring frequency will have a causal effect on how productive equally generative employees are seen to be.

Study 1b

Building on our findings that people show a systematic preference for monitoring certain targets more frequently than others, Study 1b tests the consequences of monitoring (what are actually equally productive targets) more versus less frequently. Participants completed a manager simulation task that essentially matched the experience described to Study 1a participants. More concretely, participants monitored four employees, checking in on their incremental progress over a 10-day period. The two employees who were checked in on rarely (4

---

[1] Note that because the responses within each pair were constrained to sum to 10, then these statistical tests and effect sizes are equivalent to what would be observed from one-sample t tests against 5, which would reflect an equal allocation of check-ins between the two targets.

times) and frequently (10 times) were actually equally productive, each completing 51 units of output over the 10-day period. But we expected that participants would show evidence of the monitoring frequency effect by rating the rarely monitored worker as more productive than the frequently monitored one. Two other employees who were checked in on a moderate amount (7 times) differed in their actual productivity, completing either 50 or 52 units of output. A comparison between these two will show whether participants have (at least) *some* sensitivity to actual productivity.

Method

Participants. One hundred thirty-two undergraduates (88% female, 11% male, 1% who chose to not disclose; $M_{\text{age}} = 20.59$, $SD_{\text{age}} = 5.13$) at a European university participated in exchange for course credit.

Procedure. Participants took part in a workplace simulation. They considered being the manager of a factory that made watch parts. Each day, a (supposedly random) selection of employees was checked-in on, at which point the participant would learn their output since their last check-in. After the monitoring period, participants were asked to evaluate how productive each employee was. Crucially, the employees were said not to be aware of these inspections; thus, their productivity could not have been affected by them.

Overall, participants monitored 4 distinct employees during a simulated 10-day period. As participants proceeded through each day of the simulation, they received feedback about how many watch parts some or all of the employees had completed since the last check-in. To avoid ambiguity, each time feedback was provided, it was explicitly labeled as reflecting the employee's progress since the last check-in. When an employee was checked-in on, the employee’s face, as well as the number of watch parts completed since the last check-in, were presented on screen for 4 seconds. We counterbalanced which White male face—all drawn from the Chicago Face Database (Ma et al., 2015)—was used to represent each employee. Table 2 presents the check-in schedule, as well as how many incremental watch parts each employee was shown to have completed at each check-in.

The critical targets were employee F[requent] and employee R[are]. These two employees were equally productive; they varied only in how frequently they were monitored—frequently (F) or rarely (R). We also included two additional employees: H[igh] and L[ow]. These employees were monitored moderately frequently, though one (H) had slightly *higher* productivity than the two critical targets, whereas the other (L) had slightly *lower* productivity.

Following the 10 days of monitoring, participants completed a two-item measure of perceived productivity about each of the four targets. The targets were identified both by their employee letter and their photo. One item directly asked participants to *rate* each employee's productivity: "How would you rate each employee's productivity?", anchored at 1(*very low*) and 9(*very high*). The second item asked participants to make an *estimate*: "If you checked-in on each employee on the next day, how many parts do you think they would have completed?" The

Table 2

Check-in Schedule and Observed Output at Each Check-In (Study 1b)

Employee	Day										Total
	1	2	3	4	5	6	7	8	9	10	
F	5	6	5	4	6	5	4	6	4	6	51
R	5	-	-	15	-	-	15	-	-	16	51
L	6	4	-	10	5	-	9	6	-	10	50
H	5	-	10	6	-	11	5	-	9	6	52

Note. On days an employee was not monitored, a dash has been inserted.

F = Frequent, R = Rare, H = High, L = Low.

order of the items was counterbalanced. We standardized and averaged these items to create a two-item *perceived productivity* composite ($r = .21$, $p < .001$).

Results and Discussion

In order to determine whether and how the employees were perceived to differ in their productivity, we conducted a mixed model analysis. *Employee* was a categorical predictor that identified which of the four employees (F, R, H, or L) was being judged. We included random effects of participant and target face. These account for the non-independence of trials completed by the same participant or that concern the same face.

A significant effect of employee emerged, $F(3, 390.68) = 15.63$, $p < .001$, which indicated systematic heterogeneity in the productivity perceptions (see Table 3). Suggesting that participants were sensitive to actual differences in productivity, the high-productivity worker (H) was judged to be more productive than the low-productivity worker (L), $t(389.03) = 2.14$, $p = .033$. Providing direct evidence of the MFE, the rarely monitored employee (R) was judged to be

Table 3

Average Judgments (and SDs) for Each Measure, by Target (Study 1b)

Measure	Employee			
	F	R	L	H
Rating	5.81 (1.99)$_{b}$	6.83 (1.57)$_{a}$	6.09 (1.54)$_{b}$	6.48 (1.37)$_{a}$
Estimate	9.09 (11.91)$_{c}$	13.05 (16.35)$_{a}$	10.08 (12.74)$_{bc}$	10.92 (14.45)$_{b}$
Composite	-0.21 (0.78)$_{c}$	0.24 (0.85)$_{a}$	-0.09 (0.68)$_{c}$	0.06 (0.72)$_{b}$

Note. The composite averages the standardized perceived productivity measures: the Rating of target productivity and the Estimate of how many parts they would complete in an upcoming period. Means within each row that do not share a subscript differ at the $p < .05$ level.

F = Frequent, R = Rare, H = High, L = Low.

much more productive than the frequently monitored one (F), $t(391.03) = 6.48$, $p < .001$. It is notable that this (larger) gap emerged despite no actual difference in these employees' productivity.

One illustration of the practical value of these results comes from Study 1a, which identified different factors that can encourage a monitor to engage in more or less frequent monitoring of targets. That is, employees who are new, those about whom negative rumors have been heard, and those whose work is deemed more important, may all be judged to be less productive simply by virtue of being monitored more frequently. This list is not meant to be exhaustive. Instead, these findings can be easily extended by identifying other variables that connect to the preference for monitoring frequency. Study 1c thus continues in this vein. Whereas Study 1a examined the monitoring frequency preferences of those who took on the role of the monitor (i.e., the manager) in our simulation, Study 1c provides a complementary examination by probing the monitoring frequency preferences of those who adopted the perspective of the monitored (i.e., an employee).

Study 1c

Monitors often have control over just how frequently they decide to check-in on a target's progress. But the monitored can also play a role. An employee can try to attract their boss's attention and thus have their work monitored frequently, or they can try to work in the shadows, outside of their boss's vigilant eye. Whereas monitors themselves are most likely to have an information-gathering goal, the monitored themselves will typically have an impression management goal. In this case, that would be a desire to be viewed as a good, productive worker.

In Study 1c, we again described the same workplace context used in the previous two studies, but this time we asked participants to consider being an employee. We asked participants—if their goal were to achieve a positive impression as a productive worker—whether they would prefer to be monitored more or less frequently than their coworkers. Study 1b suggests that people should prefer to be monitored less frequently in this context to achieve this goal. But we suspected that the MFE may be counterintuitive, such that people may think that as an employee they would want to be monitored more frequently, having their production remain the focus of their boss's attention. The hypotheses, methods, sample size, exclusion criteria, and analysis plan were preregistered: https://aspredicted.org/ZWJ_MSR.

Method

Participants. We requested 1,000 CloudResearch-approved American participants from AMT, though received one more than requested. Per our preregistered exclusion criteria, we excluded 43 participants who could not report that their goal was to come off as a particularly productive employee. This left 958 participants (60% female, 39% male, 1% other; M_{age} = 40.45, SD_{age} = 11.98) in all analyses reported below.[2]

Procedure. Participants were offered information about a workplace context that paralleled that used in the previous studies. Except in this case, participants considered being one of the four employees instead of the supervisor. They learned that the manager would occasionally evaluate the employees' incremental output since their last check-in ("the supervisor counts how many watch parts an employee has completed since the last time they checked-in on that particular employee.") Furthermore, we explained that the manager would occasionally evaluate all employees by "rating how productive the employee is and estimating

---

[2] We first conducted exploratory Supplemental Study A (N = 102), which provided initial evidence that those adopting the monitored's perspective would counterproductively prefer to be monitored *more* frequently. We preregistered a much larger sample size for Study 1c because we made a number of methodological improvements (see Supplemental Materials) that had the possibility to reduce the initially observed sample size. This large sample size would give us more ability to distinguish no true systematic preference for monitoring frequency from a systematic preference.

how many parts they can complete in a day." Note that these evaluations parallel those used in Study 1b to demonstrate the MFE. Finally, participants were asked, "If your goal were to try to convince the supervisor that you are a particularly productive employee, would you want your supervisor to perform check-ins on your progress more or less frequently than they perform check-ins on your fellow employees?" Participants responded on a 7-point scale. The order of the 7 response options was counterbalanced, but responses were always coded in this way: *much less frequently* (1), *somewhat less frequently* (2), *a little less frequently* (3), *equally frequently* (4), *a little more frequently* (5), *somewhat more frequently* (6), *much more frequently* (7).

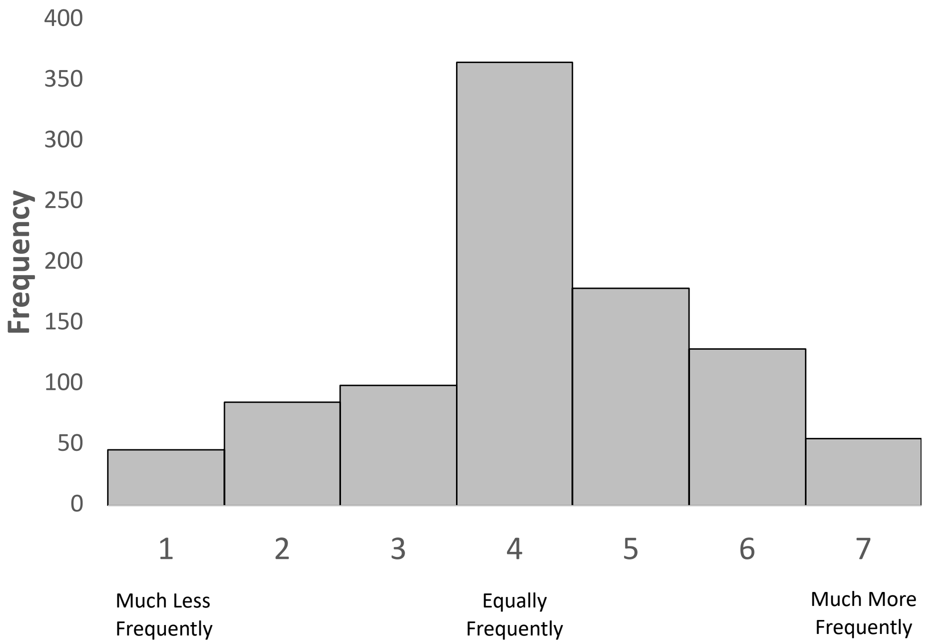
Results and Discussion

We tested whether those taking the perspective of eager employees—those who want to be judged as particularly productive—would prefer to be monitored more or less frequently than other workers. They tended to show a preference for being monitored *more* frequently than other workers, $M = 4.20$, $SD = 1.45$, $t(957) = 4.35$, $p < .001$, $d = 0.14$. Even though Study 1b suggested that employees would come off as more productive to the extent they are monitored less frequently, participants thought that greater monitoring would help them to achieve this goal. This suggests that the MFE is counterintuitive.

To further explore this data, we identified how many participants expressed an interest in being monitored with a different frequency than the other workers (see Figure 1). Of these 593 participants, only 230 (39%) preferred what Study 1b suggested is the better route to maximizing one's reputation as productive—i.e., to be monitored less frequently than the others. A binomial test showed that this was significantly less than 50%, $p < .001$. More than 1.5 times as many participants displayed the counterproductive preference.

Figure 1

Count of Participants Who Expressed Different Preferences for Being Monitored (Study 1c)



Note. For the purpose of this figure (and all analyses), responses from participants who saw a reversed scale were reverse-coded to match this scale.

In summary, Study 1a showed that monitors' beliefs and expectations lead them to monitor some targets more frequently than others. Study 1b provided initial evidence of the monitoring frequency effect—i.e., that more frequently (vs. more rarely) monitored targets seem to progress more slowly. Study 1c showed that the targets of monitoring—to the extent that they may have control over how frequently they are monitored—do not leverage the monitoring frequency effect (and may actually be hurt by it). Of course, the MFE need not only apply to the monitoring of workers. Study 2 tests it in a qualitatively different domain.

Study 2

On June 4, 2021, the U.S. state of Florida became the first in the nation to move from a daily report of new COVID-19 infections to a weekly one. The announcement came from the press secretary of Ron DeSantis, Florida's Republican governor who has built a national profile in part due to his opposition to pandemic-related precautionary measures. The governor's office explained that progress against the virus meant that there was no longer a need to issue the daily reports (Paz, 2021). But is a shift to infrequent monitoring a way to reinforce perceptions that the threat posed by COVID is waning?

In Study 2, participants completed a simulation in which they took on the role of a regional epidemiologist. We tested whether infrequently (vs. frequently) monitoring the spread of a contagious disease—in the form of occasional (instead of daily) reports of new infections—may make the disease seem to be progressing through a community more quickly. Whereas Study 1b tested for this monitoring frequency effect in a context in which progress was good (i.e., worker productivity), Study 2 tests for the MFE in a context in which progress is bad (i.e., disease spread). We expected to find support for the MFE, such that more infrequent updates on the number of new cases would encourage a sense that the disease was progressing quickly. In so doing, we could address an alternative explanation for Study 1b that more infrequent monitoring simply encourages more positive impressions of targets. The hypotheses, design, sample size, exclusion criteria, and analysis plan were preregistered: https://aspredicted.org/LDH_27X.

Method

Participants. We requested 100 Americans from Amazon Mechanical Turk from CloudResearch's approved participant pool.[3] At the study's conclusion, we asked participants whether they received information about both cities on every day of the simulation. Per our preregistered exclusion criteria, those 78 participants (59% female, 41% male; M_{age} = 39.69,

[3] We first conducted exploratory Supplemental Study B (*N* = 200), which provided initial support for the basic idea tested here. Based on the large effect observed in that study, we preregistered a smaller sample size for Study 2 and made a number of methodological improvements (see Supplemental Materials).

SD_{age} = 9.62) who accurately reported that they always received information about one of the cities, but only sometimes received information about the other city, were included in the analyses.

Procedure. Playing the part of a regional epidemiologist, participants learned that part of their job entailed monitoring the spread of a novel virus in two similarly sized cities, Washington and Franklin. They would receive one or two reports on each day, from one or both cities. The report detailed how many new infections had been reported in that city since the last report. We explained that "the two cities simply have different official policies regarding what days of the week they issue these reports." So that participants would understand there was a direct correspondence between infection rates and hospital demand—a detail that was particularly relevant to one of our two key measures—participants learned that approximately 2.1% of those infected would require hospitalization. We quizzed participants on whether each report would include the number of new infections that had been recorded since the last report (as opposed to on that day in particular). Just before the simulation began, this detail was reinforced to all participants.

The simulation lasted 12 days. At the start of each day, the screen was cleared except for a simple timestamp ("Day X"). Next, one or two new-case reports were offered in sequence. For one of the cities, a report was issued every day. For the other city, a report was issued every three days. Which city was rarely or frequently monitored was counterbalanced across participants. Despite variation in monitoring frequency, 1,171 citizens of each city were infected over the course of the simulation (Table 4).[4]

[4] The values themselves were based on actual COVID new-infection reports released by a large U.S. city just prior to the period when this study was run.

Table 4

Monitoring Schedule and New Cases Reported at Each Check-in, by City (Study 2)

City	Day												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
F	105	104	104	97	88	89	81	90	98	102	102	111	1,171
R	-	-	313	-	-	274	-	-	269	-	-	315	1,171

Note. On days a city report was not received, a dash has been inserted.

F = Frequently monitored city, R = Rarely monitored city.

After completing the 12 days of monitoring, participants completed two measures designed to probe (mis)perceptions that the disease was progressing more quickly in one community than the other. The *progress* measure read “To what extent is the disease progressing quickly through the population of…?” Participants judged each city—presented in a counterbalanced order—on a 9-point scale anchored at 1(*Not at all*) and 9(*Extremely*). The *hospital* measure instead asked participants to make a policy recommendation:

> “Although both cities have the same number of permanent hospital beds, you have the ability to authorize the supply of temporary hospital beds to one or both cities. Now that you have a sense for how quickly the disease is spreading in each city, which city in your opinion has a greater need for more hospital beds?”

The 9-point scale was anchored at 1 (*Definitely* [City X]) and 9 (*Definitely* [City Y]). The midpoint (5) was labeled “*They both have the same need.*” We counterbalanced which city occupied which endpoint. We recoded all responses so that higher numbers reflected an opinion that the rarely monitored city should be prioritized for more beds.

Results and Discussion

We test whether the MFE emerged even when progress—here, community disease progression—is a negative quality. The progress measure showed that the disease was perceived be progressing more quickly through the rarely monitored city ($M = 6.53$, $SD = 1.82$) than the frequently monitored city ($M = 5.88$, $SD = 1.91$), paired $t(77) = 2.96$, $p = .004$, $d = 0.33$. The hospital measure—through a direct comparison with the scale midpoint (5)—revealed a preference for diverting hospital resources to the rarely instead of the frequently monitored community ($M = 5.79$, $SD = 1.90$), $t(77) = 3.70$, $p < .001$, $d = 0.42$. Of the 46 (of 78) participants who expressed a preference for supplying one community with more additional hospital beds than another, 35 (76%) of them offered the opinion that the rarely monitored community needed more beds. A binomial test showed that this was significantly greater than 50%, $p < .001$.

These results provide another demonstration of the MFE. Frequent monitoring appears to minimize perceptions of progress not merely in domains in which slow progress is a negative (e.g., work output), but when it is a positive (e.g., disease spread). For our final study, we return to our original context in an effort to more clearly identify what information-processing factor or factors are responsible for the MFE.

Study 3

We have provided support for the monitoring frequency effect in two contexts. Even when the rate of progress (and thus the total observed progress) was actually constant, increasing the frequency of monitoring led perceptions of progress to decrease. By our reasoning, people display a sort of temporal neglect: They are more sensitive to the greater incremental progress that infrequent monitoring begets without fully adjusting for the background passage of time. That said, there remains ambiguity regarding exactly what form this temporal neglect takes. Our studies have yet to distinguish whether the MFE emerges due to (1) biased tracking of temporal information that is then reasonably applied, or instead (2) that the target information (as recalled) is not synthesized normatively to fully incorporate the implications of the different temporal frequency of monitoring.

By the first account, the challenge with monitoring progress may simply be that people are less adept at accurately tracking (and thus accurately distinguishing between targets with regard to) how often a target is monitored. That is, given incremental progress is directly observed, it may be simpler to distinguish the (greater) incremental progress of a rarely (from a frequently) monitored target. Keeping up with how frequently monitoring is occurring at all—given each check-in is experienced less by its frequency and more by the salient output information that is presented during such check-ins—may be a more difficult tracking task. When actual and perceived frequency decouple, judgments are likely to show more evidence of regression to the (sample) mean (Fiedler & Unkelbach, 2014). Under this possibility, the MFE would emerge due to a tendency to differentiate targets more based on how much incremental output they displayed than on the different frequency with which they were monitored, even though perceivers leaned on this tracked (but biasedly recalled) information in a normative manner when drawing conclusions about target productivity. That said, we were *a priori* skeptical of this account's plausibility given previous work suggesting that people can automatically track the frequency of events (Hasher & Zacks, 1984). In this sense, it seemed unlikely that monitoring frequency itself would be especially plagued by the sort of poor tracking that could give rise to the MFE.

Note that the second account—that perceivers would relatively neglect the temporal component when forming productivity impressions—is most consistent with the logic developed in the Introduction, particularly the idea of evaluation by moments (and the temporal neglect that implies) that seems to characterize people's evaluations of affective experiences (Fredrickson &

Kahneman, 1993; Kahneman, 2000; Kahneman et al., 1993; Redelmeier & Kahneman, 1996; Varey & Kahneman, 1992). Of course, the MFE is not about affective evaluations. But just as people neglect the duration of affective experiences and instead prioritize the intense affective experiences of single focal moments, the MFE itself may reflect a more robust reliance on the feedback gleaned from single moments (e.g., observed output) without a full consideration of how monitoring frequency should also guide interpretation of these data points.

Study 3 sought to test these two accounts by returning to the worker monitoring paradigm used in Study 1b. Notably, these two accounts are not mutually exclusive; each may contribute to the MFE. In addition to a few other modifications, we added measures that asked participants to recall how much incremental output each target tended to display at each check-in and how often those check-ins occurred. This allowed us to replicate the MFE and test whether biased recall about the targets and/or relative neglect of the temporal information contribute to the MFE. The materials, sample size, hypotheses, exclusion criteria, and analysis plan were preregistered: https://aspredicted.org/83J_16G.

Method

Participants. Three hundred seventy-two Americans were recruited from Amazon's Mechanical Turk. We included two memory-based attention checks at the study's conclusion. One required participants to respond that not every target was monitored every day. The second required participants to recognize that the amount of progress seen in each worker's output box was a reflection of how many parts had been completed since the last check-in, not simply on that day. Per our preregistered exclusion criterion, the 97 participants who failed to answer correctly either or both of these two questions were excluded from all analyses. This left a total of 275 participants (70% female, 28% male, 1% non-binary, 1% genderqueer; M_{age} = 38.77, SD_{age} = 12.26) in the analyses reported below.

Procedure. The procedure took a similar form to that of Study 1b, but with the following four changes. First, to make it even easier for participants to internalize that the progress that they observed reflected incremental progress since the last check-in, we added more detail to the background information that explained why the manager saw the output that they saw. More specifically, we emphasized that when a worker completed a part and placed it in their output box, it was only the manager who would be able to empty the box. At that point, we quizzed participants on these details—"When you inspect the output box of a given employee, what do you observe?"—and then reinforced the key instructions—"Every time you monitor an employee, you will see what that employee has completed *since* the last time you monitored them. Only then is the box emptied."

Second, we included only two employees (the rarely and frequently monitored), and thus omitted the two control employees used in Study 1b. Third, we expanded the monitoring period from 10 to 12 days (see Table 5). Fourth, and most critically, we expanded our list of measures to allow us to distinguish between the mechanistic accounts:

Table 5

Check-in Schedule and Observed Output at Each Check-In (Study 3)

Target	Day												Total
	1	2	3	4	5	6	7	8	9	10	11	12	
F	5	3	4	3	3	4	5	5	4	4	3	5	48
R	-	-	12	-	-	10	-	-	14	-	-	12	48

Note. On days an employee was not monitored, a dash has been inserted.

F = Frequent, R = Rare.

Perceived productivity. One item—asking participants to *rate* how productive the employee is—was retained from Study 1b. The *estimate* item asked, "After the next two days, how many parts do you think each employee will have completed?" Note that we asked about 2 days because it was equally different from the interval at which participants monitored the two targets (i.e., 1 and 3 days). Participants first completed one item for both targets before proceeding to the other. We counterbalanced the order of the items and the order of the targets within each item. We standardized each item and averaged them (r = .37, p < .001) to create a two-item *perceived productivity* composite.

Recalled target information. Next, we had participants report, for each target, how many completed parts they tended to observe each time they surveyed the output box and how many times the target was monitored:

Recalled observed output. For each target, participants responded to "On the days you monitored the employee, how many parts (on average) did you find in their box?" The actual averages were 4 and 12. In an effort to avoid the contaminating influence of outlier responses, we imposed a constraint that participants had to answer between 0 and 20 parts, inclusive.

Recalled monitoring frequency. For each target, participants were asked "On how many days, out of the 12, did you monitor each employee?" Here, we imposed the constraint that responses had to be between 0 and 12, inclusive.[5]

Just like with the perceived productivity items, participants first completed one of the recalled items for both targets before proceeding to the other. We counterbalanced the order of the items as well as the order of the targets within each item.

Results and Discussion

[5] Because the frequently monitored target was monitored every day, responses could not err in a positive direction. This essentially guaranteed that we would observe significant negative bias in such recall.

We begin by testing whether the MFE replicated in this somewhat modified paradigm. At that point, we move to our new measures in an effort to test the potential mechanistic accounts:

Monitoring frequency effect. In order to test for the MFE, we defined a mixed model predicting the perceived productivity composite with employee (+1: rarely monitored, -1: frequently monitored) as a fixed-effects predictor. Given each participant offered these judgments about two targets, we included a random effect of participant. A significant effect of employee, $B = 0.28$, $SE = 0.03$, $t(548) = 8.35$, $p < .001$, showed that the rarely monitored target was seen as more productive than the frequently monitored target ($M = -0.28$, $SD = 0.69$). The untransformed means and between-target comparisons by item are provided in Table 6.

Should the MFE have followed from the target details as recalled? For each target, we multiplied how many times the participant remembered checking-in on a target by the average amount of output that they remembered observing at each of those check-ins. This product reflects a *normative productivity calculation*, an index of how productive a worker would be seen to be if participants optimally synthesized their own recollections about both observed output and monitoring frequency into a productivity perception. Using this index, the

Table 6

Effect of Monitoring Frequency on the Separate Measures (Study 3)

Item	Employee				
	Rare	Frequent	*B* (*SE*)	*t*	*p*
Rating	7.21 (1.74)	6.32 (2.05)	0.45 (0.08)	5.50	< .001
Estimate	10.13 (4.90)	7.38 (2.75)	1.38 (0.14)	9.61	< .001
Composite	0.28 (0.86)	-0.28 (0.69)	0.28 (0.03)	8.35	< .001

Note. In the Employee columns, values are means (and standard deviations). The inferential statistics, with betas and standard errors, describe a test of the effect of employee in that row.

rarely monitored worker should not have been seen as any more productive (*M* = 49.09, *SD* = 27.68) than the frequently monitored worker (*M* = 48.41, *SD* = 22.59), *B* = 0.34, *SE* = 1.08, $t(548) = 0.31$, $p = .754$. This clearly contradicts the first mechanistic account.

To be clear, the fact that the normative productivity calculation does not differ between the targets does not mean that participants tracked the output and monitoring frequency variables accurately. In fact, as can be seen in Table 7, the actual recall for each attribute for each target was significantly biased toward the overall mean. But this bias in recall was essentially identical in magnitude for each attribute. As a formal test of the recall errors' statistical distinguishability, for each participant and for each recalled feature, we took the difference score in the recollections for the two targets and divided that by the actual difference. In this way, a quotient of 1 would indicate unbiased (relative) recall. Although both quotients were less than 1 (*M*s = 0.63 and 0.64), they did not significantly differ, $t < 1$. The near equivalence of this bias helps to

Table 7

Actual and Recalled Target Attribute Values (Study 3)

Measure Employee	Actual value	*M* (*SD*)	*t*	*p*	*d*
Recalled observed output					
F	4	4.83 (2.16)	6.40	< .001	0.39
R	12	9.88 (3.19)	-11.02	< .001	0.66
Recalled monitoring frequency					
F	12	10.19 (2.38)	-12.59	< .001	0.76
R	4	5.05 (2.37)	7.36	< .001	0.44

Note. The test statistics, significance values, and effect sizes correspond to one-sample t-tests comparing the recalled attribute values against the actual value for a particular target.

explain why the product of the two variables (i.e., the normative productivity calculation) did not differ between the targets.

Were the recalled target details normatively applied? If the MFE cannot be traced to bias in how the observed output and monitoring frequency were recalled, then this suggests it may be *how* these target recollections were leaned upon that may explain the MFE. In particular, we tested our second account by examining whether perceived productivity was more a function of recalled observed output than recalled monitoring frequency. Note that as each of these (recalled) variables increases—holding the other one constant—we should see an accompanying increase in perceived productivity. But if participants display the evaluation-by-moments approach that is core to our second mechanistic account, then we should see bias in which of these attributes guides productivity judgments.

To test this second account, we entered the two recalled variables—recalled observed output (per check-in) and recalled monitoring frequency—as fixed effects into the model that already included employee as a predictor of the perceived productivity composite. This model showed that as recalled observed output increased—holding recalled monitoring frequency constant—perceived productivity increased, B = 0.24, $SE = 0.04$, $t(546) = 5.26$, $p < .001$. But as recalled monitoring frequency increased—holding recalled observed output constant—there was no tendency for perceived productivity to go up, B = 0.07, $SE = 0.05$, $t(546) = 1.50$, $p = .135$. A test of whether these betas were different in magnitude was significant, $t(546) = 2.65$, $p = .008$.

General Discussion

When tracking progress, one does not merely observe the final result of the progression. Instead, one gets updates along the way, which offer a sense of whether everything is staying on track. But depending on whether one maintains a watchful eye or a more inattentive one, how much change will be observed at each check-in will vary, independent of any actual differences in the rate of progress. Such monitoring frequency is likely to be influenced by specific beliefs or expectancies (Study 1a). And as demonstrated by the monitoring frequency effect (Studies 1b, 2-3, Supplemental Study B), more (vs. less) frequent monitoring produces an illusion that the rate of progress slows down, which can have predictable negative or positive consequences—depending on the meaning of quick progress within a domain—for judgments of the target. The monitored may have faulty intuitions, and actually counterproductive preferences, regarding how much they prefer to be monitored (Study 1c, Supplemental Study A). This suggests the MFE is counterintuitive. The MFE is not driven by biases in how information about a target is tracked and recalled. Instead, the MFE stems from a failure to fully incorporate information about monitoring frequency (and instead disproportionately lean on information about how much incremental progress was recalled at the individual check-ins) when determining how quickly the target has progressed (Study 3).

Considering the MFE in Light of Related Literatures

We found that biases in tracking and recalling information about the targets—(somewhat distorted) recollections of how much output tended to be observed at check-ins and how frequently they were monitored—was not a key mechanism that could explain the MFE. That said, especially in light of research that suggests that frequency tends to be automatically processed (e.g., Flexser & Bower, 1975; Hasher et al., 1987; Howell, 1973; for a review, see Hasher & Zacks, 1984), readers may be surprised that we found systematic errors in these recollections. But being able to track such information automatically and thus efficiently does not mean that it will be tracked perfectly. When one is monitoring the progress of several targets, then even minor errors of source confusion will lead the targets to be recalled as more similar than they would have been otherwise. In fact, when for each recalled attribute one sums the two recalled values, we actually see stronger evidence of recall accuracy. For example, participants on average recalled monitoring the workers 15.24 times, which shows little systematic bias from the 16 total check-ins. Participants instead simply recalled the targets as being slightly more similar on these attributes (observed output, monitoring frequency) than they actually were. But because this source confusion was equally pronounced for each attribute, such biases essentially canceled each other out and thus did not—in and of themselves—lead to the MFE.

Instead, the MFE was localized to a failure to integrate the recalled attribute values in a normatively defensible way. More specifically, judgments of productivity were independently predicted by the recalled amount of progress that was observed at the check-ins, but not by how often participants remembered completing such check-ins. Had participants synthesized this information accurately, each should have been a positive predictor when controlling for the other. In this way, the MFE has similarities to research on the ratio bias and denominator neglect (e.g., Reyna & Brainerd, 2008), in which judgments are sensitive to focal top-line attribute indices but fail to incorporate the less salient attributes by which these salient markers must be divided to understand their meaning. Such denominator neglect explains why Kokis et al. (2002) found that on 43% of trials participants chose to draw from an urn that had 8 or 9 winning balls out of a 100 instead of an objectively superior one that had just 1 winning ball out of 10. Denominator neglect reduces with age and intellectual ability (Toplak et al., 2014), but is still prevalent among adults (Acredolo et al., 1989; Reyna & Brainerd, 1993).

We think the MFE is compatible with, but distinct from demonstrations of ratio bias. After all, ratio bias studies actually present participants with ratios (e.g., 9 winners out of 100 balls) and test whether the denominator is relatively neglected. As our normative productivity calculation made clear, progress in our paradigms is perhaps most intuitively captured as a sum (of the amounts observed at each check-in) or a product (i.e., the amount observed at a typical check-in X the number of such check-ins), a representation that is not possible in the classic ratio-bias paradigms. Furthermore, in our paradigms, we held constant the total duration that targets were monitored. Had we instead varied this duration—for example, having participants monitor one target for 12 days and another for 24 days—productivity would need to be represented mentally (even if not by the experimenter) as a quotient (total output / days of work). In that case, the worker monitored for longer might appear more productive, because the implicit numerator (total output) may not be adequately adjusted in light of the implicit[6] denominator (total days worked).

Despite these differences, we think the MFE and the ratio bias highlight complementary ways in which certain attributes are naturally salient and thus exert disproportionate influence on more complex judgments. The representativeness heuristic (Kahneman & Tversky, 1972)—with its emphasis that people judge complex features by leaning on judgments of superficially similar features—hints at a useful umbrella theoretical concept that applies to both the MFE and denominator neglect. Much as the perceived attractiveness of a gamble may be disproportionately guided by the number of opportunities to win (and neglect the total number of balls that shape the likelihood that those winners will be drawn), judgments of productivity seem to disproportionately weight observations of worker production (and neglect the time scope over which such observations were collected). In all cases, perceivers are sensitive to the attribute that is most similar to the characteristic being judged (i.e., desirability, productivity) and neglect the background feature that aids in interpreting this focal attribute.

---

[6] We refer to these as implicit to communicate that they need not be presented as a numerator and denominator, but their normative application would require them to be used in these roles.

The monitoring frequency effect, which documents the consequences of dividing monitoring into subperiods, may appear to be at odds with effects of subadditivity and unpacking, whereby breaking down larger categories into subcategories increases the perceived magnitude or frequency of those overarching categories (e.g., Fiedler et al., 2009; Mulford & Dawes, 1999). For example, Tversky and Koehler (1994) found that participants estimated that 58% of American deaths are due to natural causes. When the packed category "natural causes" was unpacked for participants into component parts—"heart disease, cancer, and other natural causes"— the estimate went up by 15 percentage points. But with the MFE, breaking up total progress into more component parts leads the overall progress to seem *smaller*. We suspect the superficial discrepancy is attributable to a difference in the paradigms used to study (and thus the mechanisms that give rise to) these distinct phenomena. In the unpacking literature, unpacking an amorphously large category into component parts helps to clarify and make concrete just how many varied components serve to define the overarching (packed) category. In contrast, in our paradigms, participants were always directly exposed to the full quantity of progress (e.g., the number of widgets produced, the number of infection cases). In that sense, the entire category (e.g., overall productivity) was always fully unpacked; what was varied was merely how many of these unpacked pieces were summed up to characterize the feedback at a single check-in. In this sense, the MFE and unpacking are not antagonistic, but merely share superficial features that at first glance can mask the distinct underlying processes that give rise to them.

Finally, the monitoring frequency effect may also shed light on previously examined phenomena that relate to monitoring frequency. Chang, Webb, Benn, and Reynolds (2017) had participants take part in a financial management simulation with two distinct goals in mind. One was a promotion-focused goal (to reach a certain high monetary threshold), while the other was a prevention-focused goal (not to fall below a low monetary threshold). When participants felt the promotion-focused goal was particularly important, a sense that they were doing poorly reduced monitoring frequency (of their own standing). When instead participants thought the prevention-focused goal was important, poor performance increased monitoring. By the MFE, poor performers who were focused on promotion (and thus were infrequent monitors) may have come to see themselves as getting out of their slump. Poor performers who were focused on prevention (and thus were frequent monitors) may have continued to feel mired in it. This may have produced (unjustified) optimism or pessimism about their prospects. Ultimately, how the confidence-boosting (vs. confidence-reducing) effect of monitoring frequency balances against the information that frequent monitoring provides to self-regulatory efforts likely depends on the performance domain. If this tradeoff were understood in a particular performance context, people could perhaps be nudged to adopt the *same* goal but in promotion- or prevention-focused terms, based on which sort of monitoring approach would prove most adaptive.

Theoretical Contribution

The counterintuitive nature of the MFE. The monitoring frequency effect joins a set of psychological phenomena that people readily display, but less readily intuit. In general, such counterintuitive findings fall in two categories. Some findings—like affective forecasting errors (Wilson & Gilbert, 2005) and perceptions of intergroup polarization (Westfall et al., 2015)—are those for which people understand the general direction of an effect (e.g., Christmastime is a joyous season for its celebrants…), but misperceive its magnitude (…but not as joyous as people think it will be; Buehler & McFarland, 2001). For other phenomena—like the satisfaction that comes from forging connections with random strangers (Epley & Schroeder, 2014) or the greater propensity to return lost property that is of greater monetary value (Cohn et al., 2019)—people misintuit even the direction of the effect. The present paper offered preliminary evidence that the MFE falls into this latter category. This finding not only helps to further delineate the classes of psychological phenomena into which people have especially poor insight, but it identifies a likely source of impression mismanagement. Those who hope to one day achieve the spotlight may do best by working in the shadows. Of course, once the fruits of one's labor are achieved, social attention should be beneficial, assuming those fruits are indeed sweet.

A novel way beliefs and expectations guide performance evaluations. The monitoring frequency effect identifies a previously unidentified route by which prior expectations and beliefs about targets may change how those targets' actions are evaluated. Although Study 1a identified a handful of target features that change how frequently targets are monitored, recall that one of those related to whether people started with positive or negative expectations about the target. There are many factors—including social category cues like ethnicity (Papageorge et al., 2020), socioeconomic status (Figlio, 2005), and gender (Carlana, 2019)—that are known to guide expectations about a target's abilities and performance. Sometimes such expectancies can produce actual changes in performance. When teachers have higher (vs. lower) expectations, students seem to perform better (vs. worse; e.g., Carlana, 2019; Figlio, 2005; Gentrup et al., 2020; Hill & Jones, 2005; Papageorge et al. 2020; for reviews, see Jussim & Harber, 2005; Wang et al., 2018). This causal influence can emerge when expectations shape, for example, the frequency or nature of the feedback that teachers provide to a student for whom they do or do not have high hopes (Gentrup et al., 2020). In this way, changes to actual performance are a first way in which expectations can change performance evaluations.

Other routes, like the one identified by the MFE, do not require an actual change in performance. For example, beliefs and expectations can encourage people to seek out information that has the potential to confirm those priors (e.g., Rajsic et al., 2015). Due to this confirmation bias, perceivers reinforce their initial beliefs. By a separate process, beliefs and expectations can lead people to evaluate or interpret the *same* evidence in a way that reinforces their prior beliefs. Such effects do not simply reflect the stickiness of a prior, for these expectancies color judgments only or especially when they are set *before* perceivers actually sample the evidence. For example, Lee et al. (2006) found that people liked the taste of beer better when a few drops of a secret ingredient had been added. However, if participants knew before tasting that the addition was balsamic vinegar, then this disgusting-sounding supplement contaminated the subsequent tasting experience. But crucially, learning this feature afterward did not spoil the memory of the already-completed trial. Similarly, Critcher and Dunning (2009) showed more directly how *a priori* expectations on a performance task can color one's bottom-up experience while completing the task (e.g., fueling a sense that one was struggling less on each question or solving the problems more quickly), thereby explaining why prior self-views can color performance evaluations (Ehrlinger & Dunning, 2003).

Consider how the MFE reflects a qualitatively distinct way in which beliefs or expectations can color perceptions of a target. Previously identified mechanisms highlight how these priors can bring about actual changes in a target, can modify what information one attends to, or can change how such ambiguous information is interpreted, all serving to confirm one's existing priors. The monitoring frequency effect instead identifies a more indirect route, one by which beliefs or expectations can change the pace at which the same information is gathered, which in turn affects perceptions. More generally, to the extent that different beliefs and expectations can influence the frequency of monitoring a target, then the MFE establishes why these factors can serve as causal precursors to perceptions of targets' rate of change and thus propensity for progress.

Broader implications. The monitoring frequency effect may offer insight into how factors beyond beliefs and expectations color evaluations of progress. For example, motivation and personal relevance could also influence monitoring frequency, with predictable consequences for evaluations. We observed one example of this in Study 1a, in which participants indicated a preference for more frequent monitoring of more important projects. The perverse effect whereby task importance encourages more frequent monitoring, which (by the MFE) should slow perceptions of progress, presumably extends to self-monitoring as well. This suggests a perverse trap. The dieter who is eager to lose weight, the person who is desperate to gain a social media following, and the academic who aspires to be influential in the field may find themselves frequently checking how many kilos they have lost, how many followers they have acquired, and how many new citations they have amassed. Such eagerness may ultimately beget a sense of failure and inadequacy. We ourselves will be frequently checking to see whether our own work inspires many direct tests of these ideas.

References

Acredolo, C., O'Connor, J., Banks, L., & Horobin, K. (1989). Children's ability to make probability estimates: Skills revealed through application of Anderson's functional measurement methodology. *Child Development, 60*(4), 933–945. https://doi.org/10.2307/1131034

Alves, H., Vogel, T., Grüning, D., & Mata, A. (2023). Why leading is (almost) as important as winning. *Cognition, 230*, 105282.

Alves, H., & Mata, A. (2019). The redundancy in cumulative information and how it biases impressions. *Journal of Personality and Social Psychology, 117*(6), 1035–1060. https://doi.org/10.1037/PSPA0000169

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*(4), 446–458. https://doi.org/10.1037/0096-3445.121.4.446

Buehler, R., & McFarland, C. (2001). Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin, 27*(11), 1480-1493. https://doi.org/10.1177/01461672012711009

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics, 13*4(3), 1163-1224. https://doi.org/10.1093/qje/qjz008

Chang, B. P. I., Webb, T. L., & Benn, Y. (2017). Why do people act like the proverbial ostrich? Investigating the reasons that people provide for not monitoring their goal progress. *Frontiers in Psychology, 8*, 152. https://doi.org/10.3389/fpsyg.2017.00152

Chang, B. P. I., Webb, T. L., Benn, Y., & Reynolds, J. P. (2017). Monitoring personal finances: Evidence that goal progress and regulatory focus influence when people check their balance. *Journal of Economic Psychology, 62*, 33–49. https://doi.org/10.1016/J.JOEP.2017.05.003

Chang, B. P. I., Webb, T. L., Benn, Y., & Stride, C. B. (2017). Which factors are associated with monitoring goal progress? *Frontiers in Psychology, 8*, 434. https://doi.org/10.3389/FPSYG.2017.00434

Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science, 365*(6448), 70-73. https://doi.org/10.1126/science.aau8712

Critcher, C. R., & Dunning, D. (2009). How chronic self-views influence (and mislead) self-assessments of task performance: Self-views shape bottom-up experiences with the task. *Journal of Personality and Social Psychology, 97*(6), 931-945. https://doi.org/10.1037/a0017452

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*(1), 5-17. https://doi.org/10.1037/0022-3514.84.1.5

Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. *Journal of Experimental Psychology: General, 143*(5), 1980-1999. https://doi.org/10.1037/a0037323

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. *Psychology of Learning and Motivation - Advances in Research and Theory, 57*, 1–55. https://doi.org/10.1016/B978-0-12-394293-7.00001-7

Fiedler, K., Hofferbert, J., & Wöllert, F. (2018). Metacognitive myopia in hidden-profile tasks: The failure to control for repetition biases. *Frontiers in Psychology, 9*, 903. https://doi.org/10.3389/fpsyg.2018.00903

Fiedler, K., & Unkelbach, C. (2014). Regressive judgment: Implications of a universal property of the empirical world. *Current Directions in Psychological Science, 23*(5), 361-367. https://doi.org/10.1177/0963721414546330

Fiedler, K., Unkelbach, C., & Freytag, P. (2009). On splitting and merging categories: A regression account of subadditivity. *Memory & Cognition, 37*(4), 383-393. https://doi.org/10.3758/MC.37.4.383

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom—A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*(1), 527-561. https://doi.org/10.1006/obhd.2001.2981

Figlio, D. N. (2005). Names, expectations and the black-white test score gap. https://doi.org/10.3386/w11195

Flexser, A. J., & Bower, G. H. (1975). Further evidence regarding instructional effects on frequency judgments. *Bulletin of the Psychonomic Society, 6(*3), 321-324. https://doi.org/10.3758/BF03336675

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45–55. https://doi.org/10.1037//0022-3514.65.1.45

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction, 66*, 101296. https://doi.org/10.1016/j.learninstruc.2019.101296

Grüning, D. J., Alves, H., Mata, A., & Fiedler, K. (2023). Reversing the cumulative redundancy bias to demonstrate metacognitive flexibility in cue utilization. *Journal of Experimental Social Psychology, 107*, 104471.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*(12), 1372-1388. https://doi.org/10.1037/0003-066X.39.12.1372

Hasher, L., Zacks, R. T., Rose, K. C., & Sanft, H. (1987). Truly incidental encoding of frequency information. *The American Journal of Psychology, 100*(1), 69-91. https://doi.org/10.2307/1422643

Hill, A. J., & Jones, D. B. (2021). Self-fulfilling prophecies in the classroom. *Journal of Human Capital, 15*(3), 400-431. https://doi.org/10.1086/715204

Howell, W. C. (1973). Storage of events and event frequencies: A comparison of two paradigms in memory. *Journal of Experimental Psychology, 98*(2), 260-263. https://doi.org/10.1037/h0034380

Jenkins, J., & Terjeson, K. J. (2011). Monitoring reading growth: Goal setting, measurement frequency, and methods of evaluation. *Learning Disabilities Research & Practice, 26*(1), 28–35. https://doi.org/10.1111/J.1540-5826.2010.00322.X

Jussim, L., & Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review, 9*(2), 131-155. https://doi.org/10.1207/s15327957pspr0902_3

Kahneman, D. (2000). Evaluation by moments: Past and future. In D. Kahneman & A. Tversky (Eds.), *Choices, Values and Frames*. New York: Cambridge University Press and the Russell Sage Foundation.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401–405. https://doi.org/10.1111/J.1467-9280.1993.TB00589.X

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454. https://doi.org/10.1016/0010-0285(72)90016-3

Klayman, J., & Ha, Y. W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211-228. https://doi.org/10.1037/0033-295X.94.2.211

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. Journal of Experimental Child Psychology, 83, 26–52. doi:10.1016/S0022-0965(02)00121-2

Lee, L., Frederick, S., & Ariely, D. (2006). Try it, you'll like it: The influence of expectation, consumption, and revelation on preferences for beer. *Psychological Science, 17*(12), 1054-1058. https://doi.org/10.1111/j.1467-9280.2006.01829.x

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122-1135. https://doi.org/10.3758/s13428-014-0532-5

Mulford, M., & Dawes, R. M. (1999). Subadditivity in memory for personal events. *Psychological Science, 10*(1), 47-51. https://doi.org/10.1111/1467-9280.00105

Papageorge, N. W., Gershenson, S., & Kang, K. M. (2020). Teacher expectations matter. *Review of Economics and Statistics, 102*(2), 234-251. https://doi.org/10.1162/rest_a_00838

Paz, I. G. (2021, June 4). Florida will no longer publish daily coronavirus reports. *The New York Times*. https://www.nytimes.com/2021/06/04/world/florida-covid-19-cases-vaccine.html.

Rajsic, J., Wilson, D. E., & Pratt, J. (2015). Confirmation bias in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 41*(5), 1353-1364. https://doi.org/10.1037/xhp0000090

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Reyna, V. F., & Brainerd, C. J. (1993). Fuzzy memory and mathematics in the classroom. In Anonymous. In G. M. Davies & R.H. Logie (Eds.), Memory in everyday life (pp. 91−119). Amsterdam: North Holland Press.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and individual differences, 18*(1), 89-107. https://doi.org/10.1016/j.lindif.2007.03.011

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology, 50(4),* 1037-1048. https://doi.org/10.1037/a0034910

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547-567. https://doi.org/10.1037/0033-295X.101.4.547

Unkelbach, C., Fiedler, K., & Freytag, P. (2007). Information repetition in evaluative judgments: Easy to monitor, hard to control. *Organizational Behavior and Human Decision Processes, 103*(1), 37-52. https://doi.org/10.1016/j.obhdp.2006.12.002

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making, 5*(3), 169–185.

https://doi.org/10.1002/BDM.3960050303

Vedejová, D., & Čavojová, V. (2021). Confirmation bias in information search, interpretation, and memory recall: Evidence from reasoning about four controversial topics. *Thinking & Reasoning, 28*(1), 1-28. https://doi.org/10.1080/13546783.2021.1891967

Wang, S., Rubie-Davies, C. M., & Meissel, K. (2018). A systematic review of the teacher expectation literature over the past 30 years. *Educational Research and Evaluation, 24*(3-5), 124-179. https://doi.org/10.1080/13803611.2018.1548798

Wang, Y., Baum, S. M., & Critcher, C. R. (in press). Needing everything (or just one thing) to go right: Myopic preferences for consolidating or spreading risks. *Journal of Personality and Social Psychology*.

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology, 92*(5), 821–833. https://doi.org/10.1037/0022-3514.92.5.821

Webb, T. L., Chang, B. P. I., & Benn, Y. (2013). 'The Ostrich Problem': Motivated avoidance or rejection of information about goal progress. *Social and Personality Psychology Compass, 7*(11), 794–807. https://doi.org/10.1111/SPC3.12071

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the United States: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science, 10*(2), 145-158. https://doi.org/10.1177/1745691615569849

Wilson, T. D., & Gilbert, D. T. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science, 14*(3), 131-134. https://doi.org/10.1111/j.0963-

7214.2005.00355.x

Supplemental Materials

Supplemental Study A

This study served as an exploratory pilot study for Study 1c. Unlike Study 1c, which was run using CloudResearch-approved participants recruited from AMT, Supplemental Study A recruited from the same participant population from which participants were drawn from Studies 1a and 1b (psychology undergraduates at a European university). We predicted that participants would not indicate a preference for being monitored less frequently (and maybe even a preference for being monitored more frequently) than their fellow employees.

Supplemental Study A: Method

Participants. One hundred two undergraduates (76% female, 23% male, 1% non-binary; M_{age} = 20.29, SD_{age} = 5.38) at a European university participated in exchange for course credit.

Procedure. Participants were offered information about a workplace context that paralleled that used in the previous studies. Except in this case, participants considered being one of the employees instead of the manager. Knowing that the manager would occasionally count how many watch parts had "been completed since the previous check-in", participants considered how frequently they would want to be monitored if their goal were to demonstrate their productivity. More specifically, participants were asked, "If your goal were to show your manager how productive you are as an employee, would you rather your supervisor monitor your progress more or less frequently than they monitor others'?" Participants responded on a 7-point scale. We counterbalanced whether the scale went from 1(*less frequently*) to 7(*more frequently*) or the reverse ordering. The neutral midpoint (4) was always labeled "equally frequently." We reverse-scored some responses so that higher numbers always reflected a (counterproductive) interest in being monitored *more* frequently.

Results

We tested whether those taking the perspective of an employee who wanted to demonstrate how productive they were would prefer to be monitored more or less frequently than other workers. They tended to show a preference for being monitored *more* frequently than other workers, $M = 4.68$, $SD = 1.68$, $t(101) = 4.06$, $p < .001$, $d = 0.40$. To further explore this data, we identified how many participants expressed an interest in being monitored with a different frequency than the other workers. Of these 58 participants, only 14 (24%) preferred what Study 1b suggested is the better route to maximizing one's reputation as productive—i.e., to be monitored less frequently than the others. A binomial test showed that this was significantly less than 50%, $p < .001$. More than three times as many participants displayed the counterproductive preference.

Discussion

Having found preliminary evidence that the monitored's preferences for monitoring frequency may be counterproductive, we designed Study 1c to offer an even stronger confirmatory test:

First, in Study 1c, we provided more details that would allow the materials to more closely parallel those used in Study 1b. Namely, we explained that participants were one of four employees (much as participants in Study 1b were in charge of monitoring four employees). In addition, whereas in Supplemental Study A participants learned that the supervisor would evaluate the productivity of their employees, Study 1c more closely paralleled Study 1b by describing these evaluations using language that was more similar to the dependent measures used in Study 1b: "by rating how productive the employee is and estimating how many parts they can complete in a day."

Second, we made several modifications to address potential ambiguities. To make sure there was no confusion regarding what the supervisor would see at each check-in, we explained that the supervisor would see how many watch parts the employee had completed since the last check-in *of that particular employee*. Although this italicized portion was implicit in Supplemental Study A, we made it explicit in Study 1c. Furthermore, our subsequent studies (Studies 2 and 3) include additional checks to make certain there was no confusion on this point. In addition, instead of saying that the employee's goal was to show "how productive you are as an employee," we explained to Study 1c's participants that their goal "was to try to convince the supervisor that you are a particularly productive employee." Although the directionality of the goal was implicit in Supplemental Study A (especially given participants in Supplemental Study A had just learned that a positive evaluation would increase their chances of being promoted), it was made explicit in Study 1c. Finally, the 7-point scale used in Supplemental Study A had endpoints of "less frequently" and "more frequently." In that sense, there may be ambiguity about what scale responses that are between these endpoints (1 and 7) and the neutral midpoint (4) reflect, given the endpoints merely express a *direction* of different in monitoring, but not a *degree* of difference in monitoring. Study 1c addresses this by providing qualitative labels to each point on the response scale that express the degree to which one would wish to be monitored more or less frequently than others: much more/less, somewhat more/less, or a little more/less frequently.

Third, in order to get a more precise estimate of the effect size (and recognizing that the modifications described above might reduce the effect size observed in Supplemental Study A), we increased the sample size by almost tenfold in Study 1c. Furthermore, we added a memory-based attention check at the study's conclusion in order to screen out participants who did not internalize the key detail—i.e., that their goal was to convince their supervisor that they were a particularly productive employee. Study 1c, like Studies 2 and 3, was also preregistered.

Supplemental Study B

This study served as an exploratory pilot study for Study 2. Like Study 2, it asked participants to track the community progression of a contagious disease. Participants completed a 12-day simulation in which they monitored the spread of a new contagious disease in two remote villages. In actuality, the two villages had an equivalent rate of new infections (on average, 4 per day) that was stable across time. But we varied which village was monitored more frequently. We predicted that participants would think that the more rarely monitored village was more affected by the disease.

Supplemental Study B: Method

Participants. Two hundred English speakers were recruited from Prolific. Twenty-four participants were unable to answer at least one of two memory-based attention checks presented at the study's conclusion. One required participants to remember that they always received information about one of the villages each day, but only sometimes did about the second village. The other required participants to report that what they saw in the admittance room was how many new people had gotten sick since the last visit. After excluding these 24 participants, we were left with 176 participants (77% female, 22% male, 1% non-binary; *M* M_{age} = 41.22, SD_{age} = 13.38) in the analyses reported below.

Procedure. In the simulation, participants adopted the role of a doctor who was tracking the spread of a novel, contagious disease in two remote villages. Whenever a villager began to show symptoms of the disease, they were quarantined in an admittance room until the doctor could arrive to check-in on them and authorize their transfer to a treatment room. In this way, the number of patients found in the admittance room reflected the number of newly ill patients since the doctor's previous visit. We quizzed participants to make certain they understood that patients found in the admittance room were not simply those who had fallen ill that day, but since the doctor's previous visit. Feedback served to reinforce this key detail.

Participants monitored only two villages. This should have made the monitoring itself simpler (compared to Study 1b, in which participants had to keep track of four targets), thereby making our tests more conservative. Participants monitored the two villages over a 12-day period. Whereas the frequently monitored village was checked-in on every day, the rarely monitored village was visited every fourth day. Despite this variation in monitoring frequency, 48 villagers in total became ill in each of the two villages (see Table S1).

Following the simulation, participants completed two measures designed to probe the apparent severity of the outbreak. One measure asked, "How affected would you say each village

Table S1

Check-in Schedule and Observed Admittance-Room Patients at each Check-in, Study 2)

Target	Day											Total
	1	2	3	4	5	6	7	8	9	10	11	12	
F	5	3	4	3	3	4	5	5	4	4	3	5	48
R	-	-	12	-	-	10	-	-	14	-	-	12	48

Note. On days a village was not monitored, a dash has been inserted.

F = Frequent, R = Rare.

was?” For each village, participants responded on a 9-point scale anchored at 1(*Not at all*) and 9(*A lot*). The second item encouraged a direct comparison, “Which village was most affected by the disease?” This 9-point scale was anchored at 1(*[Frequently monitored village*]) and 9([*Rarely monitored village*]).

Supplemental Study B: Results

We tested whether the MFE held even when progress—here, community disease progression—was a negative result. Our first measure showed that the rarely monitored disease was perceived as having been more affected by the disease (M = 7.26, SD = 1.41) than was the frequently monitored village (M = 6.72, SD = 1.52), paired $t(175) = 4.67$, $p < .001$, $d = 0.35$. Our second measure showed that in a direct comparison against the midpoint, the rarely monitored village was judged as significantly more affected than the frequently monitored village, $t(165)$[1] = 4.50, $p < .001$, $d = 0.35$.

Supplemental Study B: Discussion

We believe that this study offers preliminary support for the MFE in a context in which progress is negative. That said, there were several shortcomings that we wished to address before running a confirmatory study.

First, due to a programming error, the endpoints of the direct-comparison (second) measure were not counterbalanced like we intended. Although this problem did not plague the first measure, we corrected it for Study 2. Second, an alternative explanation for the results, which asked which village was more affected by the disease, is that the village that had a doctor

[1] Ten participants did not complete this measure.

visit less frequently was more affected merely by dint of not having a medical expert on site as frequently. We improved the measures used in Study 2 to address this limitation.

That said, we did believe that this study is useful in giving us a better sense of the likely effect size that we would see in a similar paradigm. If we use the effect sizes observed in the present study to determine what sample size would be necessary for a well-powered study (a big assumption, given effect sizes in individual studies are estimated with considerable imprecision), then we would need approximately 79 participants. We decided to recruit 100 participants in Study 2, with the assumption that we would lose some due to inattention. Study 2 also moved to a preregistered, confirmatory test.