In Suspense About Suspensions? The Relative Effectiveness of Suspension Durations on a Popular Social Platform

Jeffrey Gleason
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA
Roblox
San Mateo, California, USA
gleason.je@northeastern.edu

Alex Leavitt
Roblox
San Mateo, California, USA
aleavitt@roblox.com

Bridget Daly
Roblox
San Mateo, California, USA
bdaly@roblox.com

Abstract

It is common for digital platforms to issue consequences for behaviors that violate Community Standards policies. However, there is limited evidence about the relative effectiveness of consequences, particularly lengths of temporary suspensions. This paper analyzes two massive field experiments ($N_1 = 511,304$; $N_2 = 262,745$) on Roblox that measure the impact of suspension duration on safety- and engagement-related outcomes. The experiments show that longer suspensions are more effective than shorter ones at reducing reoffense rate, the number of consequences, and the number of user reports. Further, they suggest that the effect of longer suspensions on reoffense rate wanes over time, but persists for at least 3 weeks. Finally, they demonstrate that longer suspensions are more effective for first-time violating users. These results have significant implications for theory around digitally-enforced punishments, understanding recidivism online, and the practical implementation of product changes and policy development around consequences.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; • **Applied computing → Psychology**.

Keywords

consequences, suspensions, duration, user behavior, moderation, trust & safety, social media, gaming, field experiment

ACM Reference Format:
Jeffrey Gleason, Alex Leavitt, and Bridget Daly. 2025. In Suspense About Suspensions? The Relative Effectiveness of Suspension Durations on a Popular Social Platform. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3706598.3713163

1 Introduction

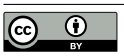
Throughout the evolution of the internet, digital social technology companies have addressed problems on their platforms with a variety of product interventions and policies. Most platforms have adapted Community Standards — the "rules" for participation — that administrators use to determine what counts as appropriate or inappropriate content, behavior, or actors [13, 28, 66]. When a user violates these rules, platforms typically take an action against the user as punishment or to encourage correction. From educational warnings, to feature rate-limiting, to temporary suspensions, to permanent bans, issuing consequences for violating behaviors is a core tenet of the moderation process [4].

Consequences have become a standard practice against violating actors, behavior, and content online because of their perceived effectiveness [29]. Within the academic literature on content moderation, there have been many studies that suggest the effectiveness (or ineffectiveness) of different types of consequences, and some of these studies may differ from the expectations of studies done offline within criminology and related disciplines [53]. For example, permanent bans have causally resulted in significant changes to user behavior or the variety of content in a platform's ecosystem [17, 20]; similarly, warnings causally reduced problematic behavior of users [74]. However, field experiments that compare the relative effectiveness of different implementations of consequences in digital spaces are rare.

More specifically, few studies look at the impact of the timing of consequences — such as frequency, duration, or sequence. While there may be key differences between a "softer" intervention like an ephemeral warning compared to a "harder" consequence like a permanent ban, there are still limited empirical insights about the relative effects of different consequence types. For example, is a warning more effective when sent twice rather than when sent only once? Is a suspension for a week more impactful than a suspension for a day?

In addition to the specifics of the intervention's impact, there is the related question of how it changes behavior over time. Many interventions in the literature have tried to measure the sustaining or waning effects of such interventions over long periods after the intervention occurs [26, 36, 47]. But as there have been limited studies on the relative effects of different consequences, there are also subsequently limited studies on how long the effects of a given consequence might last.

Finally, there is also the question of heterogeneous effects of duration on different user populations. For example, does the length of a consequence affect first-time violators differently than repeat violators? Understanding differential impacts directly informs platforms' strategies to not only try to reform individual users, but also to create strategies against violators who might be more dedicated in their repeat behaviors.



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713163

Investigating the relative impact of consequences is important because we need to know what types of consequences are most effective to combat problematic actors, behavior, or content. In the field of Trust & Safety [33], many product and policy teams at platforms have designed and implemented consequence frameworks [5]. These teams usually desire to reduce the amount of problematic actors, behavior, or content in general. Perhaps more importantly, they strive to prevent repeat offenses from people who are allowed to remain on the platform. These desires are particularly important as platforms begin to move away from outright banning participants to adopting newer perspectives around restorative justice, ones that encourage behavioral reform [18, 49, 73]. These policy-to-enforcement frameworks, however, may be grounded more in intuition or hypotheses [6, 15], and not necessarily backed by applied empirical evidence of consequence effectiveness. Therefore, creating evidence of consequence effectiveness can help policy and product teams determine best practices for establishing more successful systems to reduce initial and repeat violations.

This paper describes the result of two field experiments that test differences in consequence durations. Specifically, the experiments vary the duration of temporary suspensions — enforced in response to a range of violations on a popular social platform — in order to test the relative effectiveness of different consequence durations on key safety- and engagement-related user behavior outcomes. Measuring both of these outcomes implies a trade-off between reducing future harmful behavior and excessively punishing users, especially users who have a strong potential to reform.

This study investigates these trade-offs through the following research questions:

- **RQ1**: What is the effect of suspension duration on offending users' subsequent offending and engagement?
- **RQ2**: How does the effect of suspension duration on offending users' subsequent offending evolve over time?
- **RQ3**: What is the effect of suspension duration on first-time vs. repeat violators, in terms of both subsequent offending and engagement?

2 Related Work

2.1 Suspensions as Consequences

Consequences are one key piece of a spectrum of interventions for platform governance or "content moderation" processes [15, 56]. Generally, a consequence acts as a punishment for an offense, in a general theory of incentives towards behavioral change [53]. While a platform may enable different types of interventions against violating actors, behavior, or content, a consequence specifically is an action taken by a platform in response to a violation, usually of the platform's rules, policies, or Community Standards [66]. Interventions may be "proactive" (i.e., addressing the problem before it appears widely to other users across the platform; identified, for example, through machine learning classification) or "reactive" (i.e., addressing the problem after it appears to other users; identified, for example, through reporting mechanisms filled out by users) [35]. In both cases, if a problem is identified as violating the platform's rules, a consequence may be issued against the user's account.

For online platforms, consequences — sometimes referred to as "enforcement action" [4] or "remedies" [32] — are an integral part of the Trust & Safety field in the contemporary technology industry, to address a range of policy-violating issues. And internet technology sites, apps, and platforms have executed a wide range of consequences, ranging from warnings to removals [32]. Suspending — or the ability to "remove content temporarily [or] prevent users from accessing their accounts temporarily — from anywhere between minutes and forever" [32] — remains one of these key interventions.

Suspensions then act as a temporary intervention either against an entire account or against use of a particular feature of the platform (e.g., temporary inability to view, upload, or edit content; ability to make money; etc.). And suspensions have become so essential that even the Santa Clara Principles on Transparency and Accountability in Content Moderation (2018) recommends that platforms "publish the numbers of posts removed and accounts permanently or *temporarily suspended* [authors' emphasis] due to violations of their content guidelines" [1]. In practice, though, suspensions may be enacted in any particular form by the administrators of the technology platform.

2.2 Effects of Consequences

The effect of consequences on digital platforms in general is clear: a range of different consequences can reduce offenses, as well as reduce recidivism. Warnings work to reduce hateful language [74]. Restricting access reduces new user acquisition in hateful communities [16]. Comment deletion on Facebook decreased subsequent rule-breaking, while just hiding comments didn't have an effect [36]. And comment deletion on Reddit reduced immediate noncompliance rates [67]. In general, a variety of different interventions — especially permanent ones — result in positive outcomes.

Two types of "account access" consequences — bans and suspensions — can also be effective. Multiple studies support the same conclusion for bans: they can have significant impact on violations and recidivism. For example, closing subreddits on Reddit resulted in many accounts stopping activity, and remaining accounts decreased hate speech usage by 80% [17]. In another study of Reddit, removing 2000 communities led to 15.6% of those communities' participants leaving Reddit. However, 6.6% of remaining participants reduced toxicity, while 5% increased toxicity [20]. Deplatforming on Twitter not only reduced conversation about the deplatformed users, but also overall activity and toxicity of supporters [39].

Suspensions, of course, are another type of consequence that are stronger than a warning but less severe than a permanent ban. Suspensions allow platforms to take temporary enforcement actions against users that restrict access to all features of the platform for a predetermined amount of time [52]. Reactivation of the account may also involve the completion of some action, like verification, modifications, or acknowledgement.

Suspensions can also be effective, but there is limited work on these temporary consequences. While this paper has already referenced studies that demonstrate the impact of permanent bans in digital platforms, there is only one paper that specifically investigates differences in temporary suspensions (and only with observational data). This study finds that "all suspensions, including 24-h, 48-h and permanent suspensions … decreased the offense

probability of suspended offenders after suspension and increased the latencies of their offenses after suspension" [75].

As one addendum, it is important to note that the digital environment poses a unique challenge to "access"-type consequencing: the ability for someone to migrate. Users who create an account on one platform, when suspended or banned, may migrate to either another account on the same platform, to other communities within the same platform (e.g., [14]), or to other similar platforms entirely [37]. Multiple accounts are not uncommon on social technology platforms [42, 43, 55]. Migration is found to be common [31], and some studies find that user migration can lead to increased toxic behavior [38, 64].

2.3 Relative Effects of Suspensions and Different Suspension Durations

While suspensions can be effective, it's less clear how the duration of a suspension impacts recidivism. In offline criminology research, reviews of the literature suggest that differences in punishment severity do not appear to be effective at deterring crime [53]. Further, incarceration length's impact on recidivism "appears too heterogeneous to draw universal conclusions," even if some studies point to longer sentences potentially being more likely to deter crime [10].

While criminology is helpful for developing initial hypotheses, some literature from smaller, temporary offline behavioral interventions might suggest some additional directions. For example, a study of timeouts on children found that 15 and 30 minute timeout periods "produced a 35% decrease in deviant behavior," far more than a "1 min [period which] resulted in an average increase of 12%" [71]. However, the study also found that there was "little difference between the effectiveness of 15 and 30 min" [71]. Similarly, longer suspension periods (91-180 days) for driving offenses led to lower offense ratios than shorter suspension periods (1-30 days; [27]). However, school suspensions on students that were more severe did not deter those students from misbehaving in the future (and might have made reoffense worse for younger students; [45]).

Unlike offline interventions, in an online platform environment, users may be affected by digital interventions in very different ways. Digital platforms can implement interventions in more precise, more timely, and more targeted ways. Platforms might consider very short suspensions, or consequences targeted to specific feature use, or — at the other end of the consequence spectrum — permanent bans. Therefore, temporary suspensions might have different outcomes in online contexts than in offline applications. In the previously mentioned, observational suspension duration study [75], the authors do not note any key differences between the one- and two-day suspension periods, so there is still a significant opportunity to collect evidence for this question.

Uniting both approaches, then, one might surmise that a longer digital suspension duration could be more effective than a shorter one:

- **RQ1**: What is the effect of suspension duration on offending users' subsequent offending and engagement?
- **H1**: Longer suspensions are more effective than shorter suspensions at reducing reoffense.

2.4 Lasting Effects of Different Suspension Durations

What is even less known about digital consequences is how long those effects last, and if there is any relative difference in how much the effects persist.

Given there are limited papers on suspensions, the literature on digital consequences and other online interventions points to some evidence of lasting effects. Warnings, for instance, have been shown to be effective across some periods of time. One experiment on Twitter found that "the act of warning a user of the potential consequences of their behavior can significantly reduce their hateful language one week after receiving the warning" [74]. Other papers in the misinformation studies space point to similar lengths. For instance, one study of inoculation found intervention effects that lasted up to 3 months [47]. In another misinformation study, where eight different interventions were compared, the authors found that "the interventions also differed in the duration of their effects . . . interventions focused on teaching new skills (inoculation and media literacy) showed less decay than interventions that labeled specific pieces of content as reliable or unreliable (preemptive fact checking, source credibility, warnings)" [26]. Deletion of comments were effective, and "the effect on . . . lowered rule-breaking behavior lasted longer than the effect on continued commenting behavior" [36].

For suspension-type consequences in particular, there are no empirical studies that look at the relative duration of potential lasting or expiring effects. Therefore, there are limited priors that would suggest any hypothetical directional differences. However, if one assumes that a "stronger" (longer) intervention is more effective initially, one might also assume that its effectiveness persists at a higher level for some period of time. Because there is limited evidence on this topic overall, the paper characterizes how long the effect lasts and how strong it remains over time.

- **RQ2**: How does the effect of suspension duration on offending users' subsequent offending evolve over time?
- **H2**: The increased effectiveness of longer suspensions at reducing reoffense persists at a higher level.

2.5 Heterogeneous Effects of Different Suspension Durations

Finally, still due to the limited evidence about suspension effectiveness, there is limited knowledge about the heterogeneity of how different suspension durations might affect different types of people. In the literature around offline consequences, some studies have examined recidivism rates specifically for first-time violators vs. repeat violators (e.g., [54, 65]), which suggest that there are some differences across the demographic and behavioral profiles of first-time vs. repeat violators.

Moving to online environments, there is little evidence of differences around the behavior of first-time vs. repeat violators, as well as the potential differential impact of interventions on recidivism across both groups. Some descriptive research demonstrates that "a small number of individuals typically accounts for the vast majority of the behavior" [57]. However, research on reactions to content moderation processes have shown that users — and one may surmise it is the case especially for not-yet-violating users — often

do not fully understand reasons for removal [50], the consequence itself, or how to contest the system's decision [69]. Therefore, it may be that first-time violators are more-strongly impacted by consequences, and especially so by longer suspensions compared to shorter ones.

- **RQ3**: What is the effect of suspension duration on first-time vs. repeat violators, in terms of both subsequent offending and engagement?
- **H3**: Longer suspension durations impact first-time violators more than repeat violators, in terms of both subsequent offending and engagement.

2.6 Safety- and Engagement-Related Outcomes and Consequences

As previously described, consequences have largely been examined through the lens of recidivism, or the likelihood that a violator will reoffend and commit another violation. In addition to recidivism-related safety outcomes, this paper also highlights two additional factors: reporting-related safety outcomes and general engagement outcomes.

While consequence-related behaviors are the main outcome measured for the experiment, user reports are another key safety-related outcome. Digital platforms rely significantly on user reports to understand the prevalence and types of safety violations that occur amongst users [21, 44]. With some variability, users will report violators when problems occur [70]. Therefore, one should expect that if consequences like suspensions impact violators' behaviors, then they will also impact how often the violating player is reported by others. This makes report filing another useful outcome to measure.

Additionally, engagement is a basic measure that most technology platforms use for business purposes. Scholars have noted the importance of engagement-based metrics to understand user activity across various platforms (e.g., [8, 25, 68]). Engagement — or disengagement [51] — therefore is also a key indicator of the impact of platform mechanisms and interventions. Related to suspensions, general user activity of the violator may additionally be impacted in addition to violation recidivism.

3 Methods

This paper uses two field experiments to measure the relative impact of suspension duration on Roblox.

3.1 Roblox as a Platform

Roblox is a global immersive platform for connection and communication where millions of people come to create, play, work, learn, and connect with each other [12, 46]. These experiences are all built by Roblox's global community of creators, and they range from competitive action games, to creative role playing games, to casual social hangouts [62]. Users can design and customize avatars, make connections to befriend other users, and communicate via text and voice chat within experiences. As of June 30, 2024, Roblox has 79.5 million global daily active users [60].

Roblox also provides a layer of safety features, such as content moderation, reporting tools, and — important to this work — interventions and consequences. Roblox's moderation systems mirror

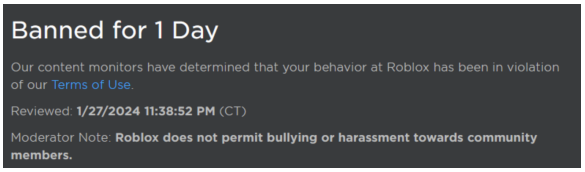


Figure 1: Example of a message sent to a user following an enforcement action, after they were reported for Bullying or Harassment.

those of many other social technology platforms. Roblox has extensive Community Standards that span safety, civility, integrity, and security, with the goal of "always ma[king] it a key priority to ensure . . . community members can connect, create, and come together in a space that is welcoming, safe, inclusive and respectful" [62, 63]. Violations of these policies can result in enforcement actions (i.e., consequences), such as warnings, content removal, account- or feature-level restrictions, or permanent deletions. Violations are identified and consequences are issued based on a combination of user-submitted reports [61], automatic classifiers [11], and human moderators. Roblox's transparency report has more details on the platform's moderation and enforcement processes [58].

3.2 Experiment Design

3.2.1 Experiment 1. Experiment 1 compared 1-day suspensions to 1-hour suspensions for users on their first policy violation in the last month. Experiment 1 utilizes a 1-hour suspension for the "short" duration, because it matches the experience of a "timeout," where play might be interrupted, but where a player might re-enter experiences later in the day. A 1-hour suspension is stronger than an ephemeral warning (which might be presented as a pop-up and then immediately dismissed), but short enough that a user could return in the same or an adjacent play session.

3.2.2 Experiment 2. Experiment 2 compared 3-day suspensions to 1-day suspensions, specifically for users on their second policy violation in the last month. Aligning with the deterrence literature for repeat violators, Experiment 2 utilizes a 3-day suspension for the "long" duration, because stronger consequences could potentially provide a larger incentive for not re-offending.

3.2.3 Ethical Considerations. Some of Roblox's community standards deal with extremely serious violations, which were excluded from these experiments. Further, the experiments described in this paper only include users aged 13 years old or over. The platform's Terms of Service dictates participation in product experiments, and data used in the study was de-identified and analyzed in aggregate. The benefits of the experiment greatly outweigh any possible harms, given the relatively short duration of the consequences issued (varying suspension times by definition changes the consequence length per user, but the consequences used in this study are not permanent nor particularly lengthy). Field experiments are especially valuable for designing tools to understand online communities better that can help improve platforms across the internet as a whole, since platforms can utilize this data and information to inform their own product decisions or policy enforcement effectiveness.

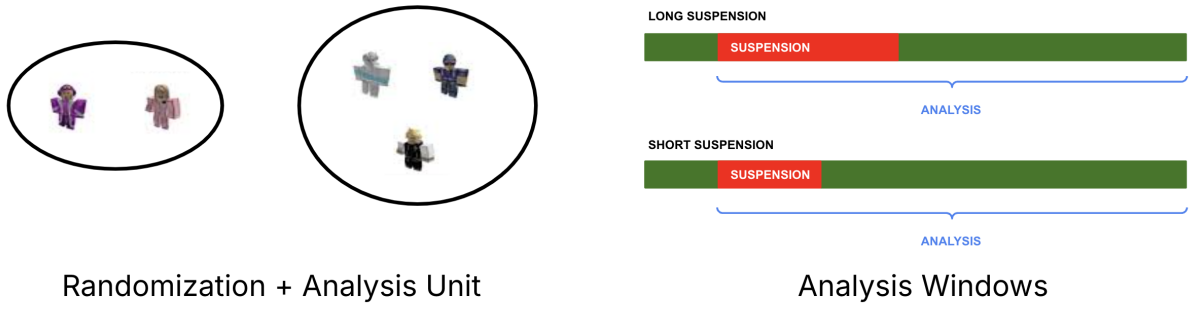


Figure 2: The left sub-figure depicts the randomization and analysis unit: a user, which represents a cluster of one or more platform accounts associated with a specific person. The right sub-figure depicts the analysis window, which begins at the time the user's suspension starts on a given account.

3.2.4 Experiment Implementation. Users were divided into mutually exclusive subsets, such that some users were only eligible for Experiment 1 and others were only eligible for Experiment 2. Participants were randomized and analyzed at the user-level, where a "user" represents a cluster of one or more platform accounts associated with an individual person, as defined by Roblox's internal detection systems [59]. The experiment analysis window for each user begins at the time the user's suspension starts on a given account, after a moderator files the enforcement action. Importantly, the analysis window does not begin at the time that the account's suspension expires, because users with multiple accounts [43] could still conceivably access the platform on alternate accounts during the suspension period.

Relatedly, a suspension's operational period begins when a suspension is issued (by platform moderators), regardless of whether the account is in use at the time. Even if the account might be offline during this suspension period, the user — upon logging back in — will see a suspension notice, and they must interactively acknowledge the notice to regain access to the account and continue playing. Finally, all consequences in these experiments originated from user-submitted reports, which were reviewed by a combination of automatic classifiers and human moderators.

In both experiments, the assignment probability for each variant was 50%. Both experiments ran for 26 days, from July 19 – August 14, 2024. Experiment 1 included 511,304 users, and Experiment 2 included 262,745 users.

3.2.5 Estimation and Inference. The experiment data was analyzed with CUPED (Controlled Experiment Using Pre-Experiment Data; [24]) to improve the sensitivity of the experiments. CUPED incorporates pre-experiment data to reduce variance in the outcome metric. Variance is reduced by a factor of R^2, where R^2 is the proportion of variance explained by a regression of the outcome metric on pre-experiment covariates. Specifically, for each outcome, one covariate is used: the same metric over a 7-day pre-experiment window. This approach reduced variance substantially for count-based outcomes (i.e. by a factor of 0.830 for number of consequences in Experiment 1 and a factor of 0.985 for total time spent in Experiment 2). Using CUPED, both experiments were powered to detect a 1.5% change in reoffense rate, as well as a 1.2% change in total time spent in Experiment 1, and a 1.4% change in time spent in Experiment 2.

In the results that follow, the delta method is used to generate confidence intervals on the relative scale [23]. P-values correspond to absolute comparisons and are adjusted for multiple comparisons using Benjamini-Hochberg [9]. Confidence intervals are not adjusted for multiple comparisons.

3.2.6 Interference. Spillover across alternate accounts s accounted for by cluster-randomizing at the user-level. This correction is important because social network users often maintain multiple accounts, especially in online gaming settings [7, 43]. Furthermore, the treatment (suspension duration) is likely to strongly affect a user's propensity to switch to a different account. However, it's important to note that this clustering does not incorporate any information about peer relationships, such as friend group structures or co-play communities. Thus, if a user's potential outcomes are affected by the treatments of their friends, we have a SUTVA violation. For example, there may be peer comparison effects: i.e., a 1-day suspension may be less of a deterrent if Roblox issued my friend a 1-hour suspension. Alternatively, interference might spread through the discussion of punishments on public forums, which is known to be a popular topic.

In order to assess the strength of evidence for interference across friends, we employ Aronow's ex post test for interference in randomized experiments [2]. This test measures the dependence between outcomes for a fixed subset of users and the treatment statuses of other users, and compares it to the null distribution that would occur if there were no indirect effects. Specifically, our fixed subset of users is a 50% sample of the 1-day suspension group, and our test statistic is the Spearman rank correlation between the reoffense outcomes of these users and their number of Roblox friends who received a 1-hour suspension. If having more friends who receive 1-hour suspensions reduces the deterrence of a 1-day suspension, the observed correlation should be greater than the vast majority of draws from the null distribution. However, we find that this is not the case: the p-value is 0.152. Thus, we do not see strong evidence for interference across friends in experiment 1.

Table 1: Descriptive statistics for experiment 1 outcome variables (1-hour suspension group).

Outcome Variable	Mean	Standard Deviation
Reoffense rate	0.27	0.44
Number of consequences	0.67	7.29
Number of reports against	4.57	98.33
Time-to-reoffense (hours)	137.15	134.65
Days active	11.24	7.32
Total time spent (hours)	134.37	3005.28

Table 2: Descriptive statistics for experiment 2 outcome variables (1-day suspension group).

Outcome Variable	Mean	Standard Deviation
Reoffense rate	0.47	0.50
Number of consequences	1.64	20.12
Number of reports against	10.70	201.34
Time-to-reoffense (hours)	125.61	120.77
Days active	12.39	7.29
Total time spent (hours)	333.83	8065.01

3.3 Data

3.3.1 Outcome Variables. This study includes 6 total outcome variables, 4 of which are related to violations/consequences and 2 of which are related to engagement.

Violation-Related Outcomes.

(1) *Reoffense*: A binary variable indicating whether the user violated any rules included in Roblox's Community Standards between the start of their suspension and the end of the experiment.
(2) *Number of Consequences*: The total number of consequences issued to the user between the start of their suspension and the end of the experiment. Again, consequences can relate to any Community Standard violation.
(3) *Number of Reports (Against Violating User)*: The total number of user-submitted reports against the user between the start of their suspension and the end of the experiment.
(4) *Time to Reoffense*: For users who reoffended, the difference (in hours) between the start of their suspension and their next violation.

Engagement-Related Outcomes.

(1) *Days Active (of Violating User)*: The number of days that the user is active on Roblox between the start of their suspension and the end of the experiment.
(2) *Time Spent (of Violating User)*: The total time (in hours) that the user spent on Roblox between the start of their suspension and the end of the experiment.

3.3.2 Descriptives. The median age in Experiment 1 was 20 (95% percentile: 53) and in Experiment 2 was 22 (95% percentile: 55). More specifically, 30.4% and 30.6% of users in Experiments 1 and 2 were 13–16, while 40.4% and 40.3% were 17–24, respectively. 54.3% and 55.2% of users in Experiment 1 and 2, respectively, identified as male. Users are from all countries where Roblox operates, the largest of which is the US.

Table 1 shows the means and standard deviations for each outcome variable over the 26-day experiment period (in the shorter suspension group), in Experiment 1. Table 2 shows the corresponding values for Experiment 2. As expected, users who committed a second violation in a month's time — and were thus included in Experiment 2 — have a higher baseline reoffense rate, number of consequences, and number of reports against them. They also have a lower baseline time-to-reoffense. Users in Experiment 2 were also more engaged, with a higher baseline number of days active and total time spent.

4 Results

4.1 Hypothesis 1

Figure 3 and Table 3 show that a 1-day suspension (relative to a 1-hour suspension), issued after a first policy violation, reduces reoffense rate by -6.7% ($p < 0.0001$), number of consequences by -6.4% ($p < 0.0001$), and number of reports against by -4.6% ($p = 0.0017$). It also increases time-to-reoffense by 7.0% ($p < 0.0001$), i.e., people take longer to reoffend. On the other hand, the effect of the longer suspension on our two engagement metrics — days active and total time spent — is not significant.

Figure 4 and Table 4 tell a similar story. The 3-day suspension (relative to a 1-day suspension), issued after a second policy violation, reduces reoffense rate by -8.1% ($p < 0.0001$) and number of consequences by -3.4% ($p = 0.0112$). The 3-day suspension also increases time-to-reoffense by 9.3% ($p < 0.0001$) — a larger relative amount than in Experiment 1. The effect of the 3-day suspension on days active is significantly negative (-1.7%; p < 0.0001), but smaller than the effect on reoffense rate. Finally, the effect on total time spent is not significant.

Based on these results, H1 is supported. Longer suspensions are more effective at preventing reoffense and consequences compared to shorter suspensions. Furthermore, longer suspensions have a larger relative effect on violation metrics than they do on engagement metrics (the latter of which were often insignificant).

4.2 Hypothesis 2

Figures 5 and 6 illustrate how reoffense rates progress over time. The top sub-plots show the cumulative reoffense rate N days after the start of a user's suspension. The bottom sub-plots show the relative effect on reoffense rate N days after the start of a user's suspension. Tables 5 and 6 show these statistics at 7, 14, and 21 days from the start of a user's suspension.

Both figures demonstrate that longer suspensions have a large negative effect on reoffense rate in the short-term. In other words, when the shorter suspension has expired, but the longer suspension is still in effect. Specifically, the 1-day suspension (relative to the 1-hour suspension) reduced reoffense rate by -33% ($p < 0.0001$). The 3-day suspension (relative to the 1-day suspension) reduced reoffense rate by -26% ($p < 0.0001$).

While there are relative effects of -33% to -26% in the short-term, it's important to contextualize this finding given the existence of multiple account usage. For example, an impassable suspension

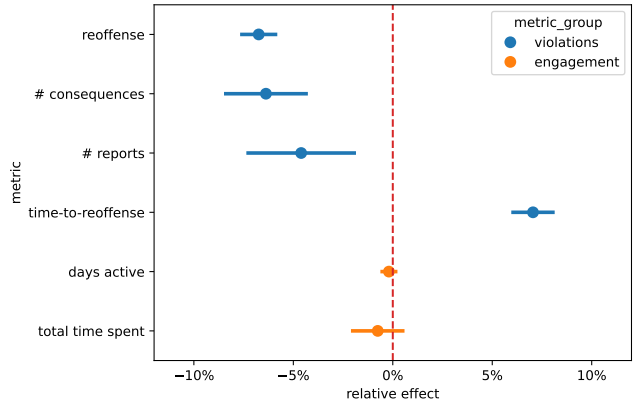


Figure 3: Effects of 1-day suspension vs. 1-hour suspension (baseline) after first policy violation.

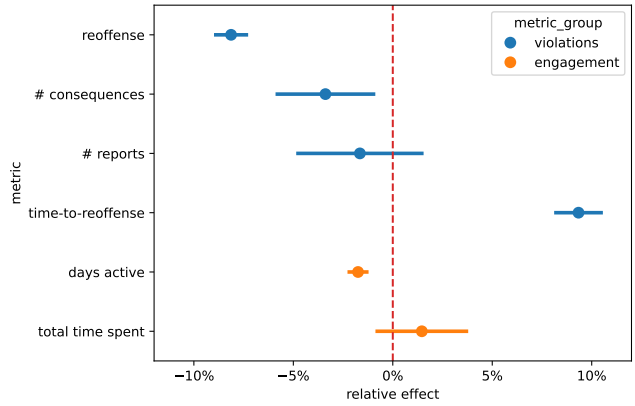


Figure 4: Effects of 3-day suspension vs. 1-day suspension (baseline) after second policy violation.

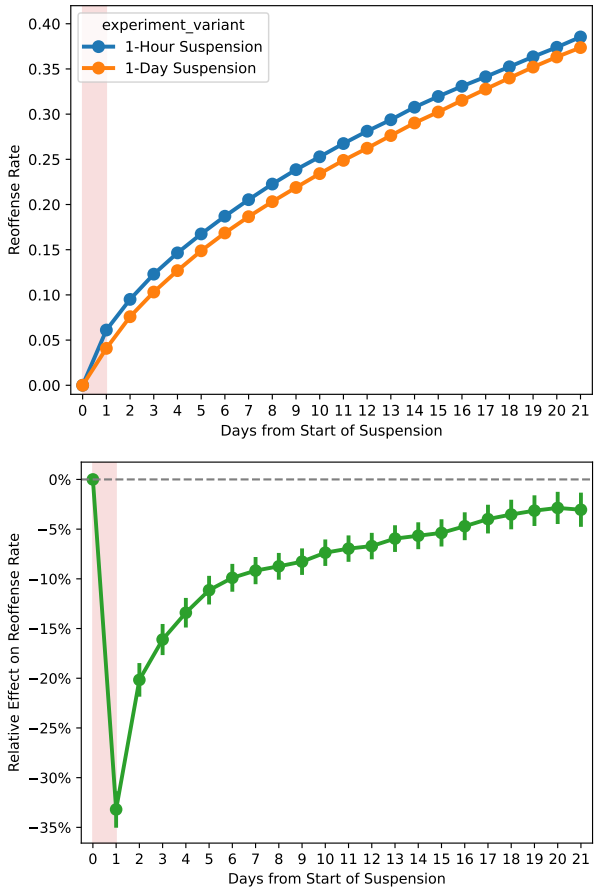


Figure 5: Top: Reoffense rate over time in experiment 1. The x-axis shows days from the start of a user's suspension. Each point references users who were observed for at least that many days. The red region represents the length of the longer suspension. Bottom: Relative effect on reoffense rate over time (baseline = 1-hour).

implies a relative effect of -100% for as long as the longer suspension is in effect. Therefore, users likely circumvent suspensions by switching to alternate accounts.

After the longer suspension expires, the magnitude of the effect decreases quickly over the next few days. This trend implies that the initial deterrence from the longer suspension wanes quickly. However, the relative effect in Experiment 1 is still -5.7% ($p < 0.0001$), fourteen days later. Then, the relative effect in Experiment 1 remains at -3.1% ($p = 0.0001$), twenty-one days later. In Experiment 2, the relative effect persists at -7.2% ($p < 0.0001$), fourteen days later. Then, the relative effect remains at -4.6% ($p < 0.0001$), twenty-one days later.

Based on these results, H2 is supported. Longer consequence durations persist at a higher level of effectiveness over time, compared to shorter suspensions. Overall, longer consequences are strongest in the short-term and wane relatively quickly, but they still persist at least 21 days later in both experiments.

4.3 Hypothesis 3

Figures 7 and 8 segment users according to their number of historical violations during their entire tenure on the platform. We bucket them into "first time" (0 historical violations), "infrequent" (1-4 historical violations), and "frequent" (5 or more historical violations) violator groups. The groupings were split at 5 violations, because it was the median, conditional on having at least one violation. In Experiment 1, there were 28% "first time" violating users and 31% "infrequent" violating users. In Experiment 2, there were 28% "infrequent" violating users.

Figure 7 and Table 7 show that the 1-day suspension (relative to the 1-hour suspension) has the largest relative effect on reoffense rate for the "first time" violators group (-12.6%; $p < 0.0001$). The relative effect is also larger in the "infrequent" violators (-9.0%; $p < 0.0001$) than in the "frequent" violators group (-4.4%; $p < 0.0001$). We observe a similar pattern for total time spent, with the largest relative effect on the "first time" violators group (-5.1%; $p < 0.0001$)

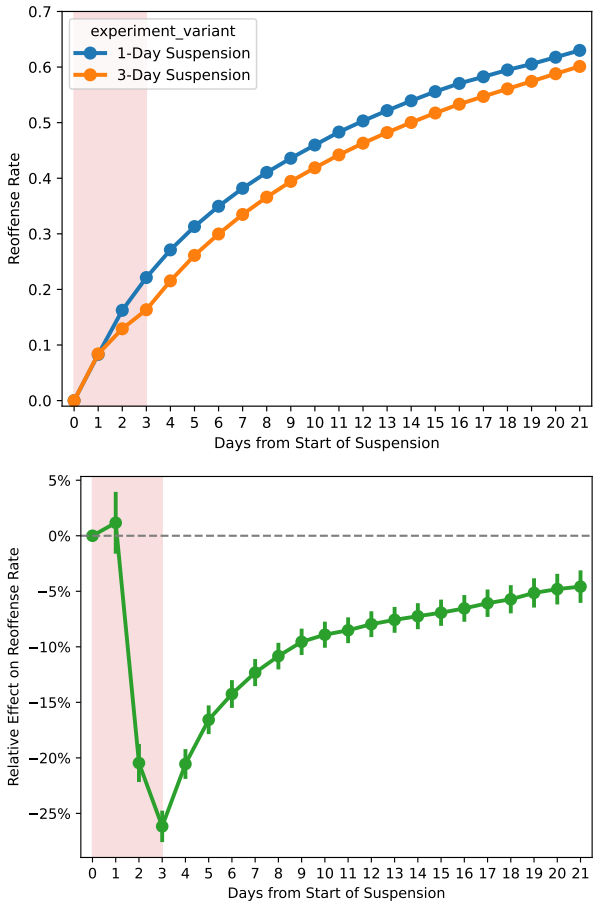


Figure 6: Top: Reoffense rate over time in experiment 2. The x-axis shows days from the start of a user's suspension. Each point references users who were observed for at least that many days. The red region represents the length of the longer suspension. Bottom: Relative effect on reoffense rate over time (baseline = 1-day).

and the second largest effect on the "infrequent" violators group (-3.5%; $p < 0.0001$). We don't observe a significant effect on time spent in the "frequent" violators group.

Figure 8 and Table 8 tell the same story. The 3-day suspension (relative to the 1-day suspension) has a larger relative effect on the "infrequent" violators group than the "frequent" violators group, for both reoffense and total time spent. Again, we don't observe a significant effect on time spent in the "frequent" violators group.

Based on these results, H3 is supported. Longer suspensions are most effective for first-time violators and least effective for frequent violators, compared to shorter suspensions. Furthermore, longer suspensions reduce time spent for first-time violators, while having no discernible effect on time spent for frequent violators.

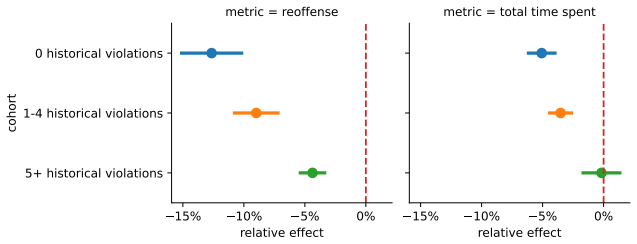


Figure 7: Users segmented by number of historical violations (of any policy). Effects of 1-day suspension vs. 1-hour suspension (baseline) after first policy violation.

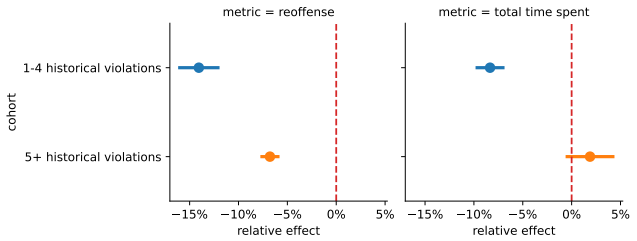


Figure 8: Users segmented by number of historical violations (of any policy). Effects of 3-day suspension vs. 1-day suspension (baseline) after second policy violation.

5 Discussion

5.1 Overview

To summarize the results again:

- *Longer suspensions are effective.* They reduce reoffense rate (-7% to -8%) and number of consequences (-6% to -3%) and increase time-to-reoffense (7% to 9%). In other words, users take longer to reoffend.
- *Longer suspensions reduce violations relatively more than they reduce engagement.* Specifically, days active and time spent are reduced by at most -2% across both experiments.
- *Effectiveness fades over time, but persists for at least 3 weeks.* Specifically, 3 weeks after the start of the suspension, reoffense rate is still down -3% and -5% in Experiments 1 and 2, respectively.
- *While longer suspensions are more effective across the board, longer suspensions reduce reoffense more among first-time violators.* For example, In experiment 1, the relative effects on first-time and frequent violators is -13% and -4%, respectively.

5.2 Platform Design Implications

Suspensions are a common moderation action across almost all social platforms today. The results demonstrated here may be easily applicable to other social technology platforms, and platform operators, designers, and policymakers should think about how these results might impact their own platform's enforcement decisions. For example, would user behavior on a given platform – in the context of the platform's use – be substantially affected by shorter or longer durations of consequences (e.g., if it impacted the ability

to write a chat message vs. stream a video)? How would changing the length of consequences impact the larger user community on a given platform? And how does differing consequence length intersect with other enforcement mechanisms to produce a strategic spectrum of interventions to address problematic user behavior? These are all questions that platform designers, engineers, and researchers should consider both theoretically and practically in their approaches to designing stronger platform consequence models.

While both experiments contain a significantly large sample size of participants, these results may be culturally specific to Roblox's platform. For example, the ease of creating alternate accounts is an important platform-specific factor that may influence the effects of suspension duration. Roblox is also an interactive environment that may comprise different affordances than, say, a text-based social media app, resulting in different user behaviors, expectations, and – then – effectiveness of some interventions. The demographics of the Roblox community compared to the audiences participating on other platforms may also impact potential behavior change. Further, implementation and effects may be dependent on the processes (such as reporting flows, moderation systems, delivery of interventions, etc.) that Roblox uses to identify and consequence on platform-specific violations. Platform designers can think critically about what aspects of Roblox's system features, user behaviors, behavioral incentives, and consequences and educational messaging might be parallel to another platform or where they might diverge. Ultimately, generalizability is a question of empirical theory testing, and systems designers should be encouraged to test what could work best in their own context and for their own users. In each of the sections below, we give specific recommendations on how platforms might approach the main takeaways from this paper.

5.3 Consequences vs. Engagement

Longer suspensions can reduce violations relatively more than they reduce engagement. This trend is illuminating for platform operators: it suggests that consequences can be implemented without a net impact on business metrics (where platforms may in theory want to avoid stronger consequences because they fear large, negative impacts to engagement). As previously recognized, alternative accounts exist on many platforms, though the ease of creating them varies across platforms. Longer suspensions could have an even greater engagement impact if users are unable to make alternate accounts or if they are less-commonly created on other platforms.

Formally combining multiple (potentially conflicting) metrics into an overall evaluation criterion is also a challenge in the design of online controlled experiments [34]. One approach to trade-off these metrics is to treat engagement metrics as guardrail metrics [34]. This implies selecting the suspension scheme that has the greatest negative impact on reoffense, but doesn't reduce engagement by more than a pre-specified amount. Future experiments could also measure spillover effects on users exposed to violations in order to capture total effects on the community, which would require more complicated experimental designs [3].

5.4 Effect of Consequences Over Time

The benefit of longer suspensions fades over time, and it does so relatively quickly in the first few days after the suspension ends. However, it's meaningful that reoffense rate remains reduced at least three weeks after the suspensions begin (-5% to -3%). Due to the drop-off, platform operators may want to pursue — in parallel with suspension — other types interventions that could theoretically persist for longer periods and/or at higher levels. For example, platforms could improve messaging about community norms, specificity about wrongdoing, clarity about one's standing in the community, and/or rapidity from violation to consequence [40]. The cumulative effect of different types of interventions is also a key area for future investigation.

5.5 Differential Impacts for First Time Violators

Both experiments demonstrate that heterogeneous effects are important to measure and consider when it comes to safety interventions. The large effects on first-ever violators, compared to repeat violators, suggest that there are opportunities for interventions that may be different for different cohorts of users. However, it's important to recognize that even for repeat violators, longer durations are more effective than shorter durations.

For first-time violators in particular, it may be important for platform designers and policymakers to emphasize education and transparency [48] or pursue community-based approaches [41, 72] that might help adapt restorative justice approaches for users new to violative behaviors. On the other hand, smaller effects on frequent violators implies that stronger approaches may need to be taken after a certain point. However, researchers are split on whether stronger punishments for repeat violators is more effective or even defensible [19, 30], and some research into punishment suggests that "penalty escalation" may be misguided [22]. Platform operators should think carefully about how various thresholds for defining user cohorts and intervention durations might have significant implications for consequence effectiveness. Researchers should investigate this area more deeply in future work.

5.6 Limitations

First, to reiterate, the enforcement of a suspension occurs at moderation time, not at re-login time, so a user's consequence (especially for a 1-hour suspension) might have happened when they were offline, reducing the impact of the shorter duration (even if the user was required to interactively acknowledge the violation to continue using Roblox). Overall, generalizability is important, and platforms and academics are encouraged to run more field experiments that explore duration differences.

Second, the experiments capped suspension duration at 3 days maximum. There may be different effects of consequences at higher durations (e.g., 5-, 7-, or potentially even longer durations), and platforms could experiment with these longer durations in the future. Researchers and platform operators should consider experimentation with longer durations to estimate the relative effects of even longer consequences.

6 Conclusion

We ran two field experiments on Roblox to measure the relative impact of suspension duration on key safety- and engagement-related outcomes. We learned that 1) longer suspensions were more effective, 2) longer suspensions reduced violations relatively more

than they reduced engagement, 3) the effectiveness of longer suspensions faded over time, but persisted for at least three weeks, and 4) longer suspensions reduced reoffense more among first-time violators.

Acknowledgments

Jeffrey Gleason was employed at Roblox at the time of this study. Thanks to the Roblox engineering team who made these experiments possible: Nicholas Ngai, Kit Tse, Eric Perez, Richard Li, and Johnny Zhou. Thanks also to the cross-functional team who provided feedback on the overall project and study design: Carly Villareal, Mia Paulsen, Dave Woolston, and GiGi Demmming.

References

[1] n.d.. The Santa Clara Principles On Transparency and Accountability in Content Moderation. Retrieved September 9, 2024 from https://santaclaraprinciples.org

[2] Peter M Aronow. 2012. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research* 41, 1 (2012), 3–16.

[3] Peter M Aronow, Dean Eckles, Cyrus Samii, and Stephanie Zonszein. 2021. Spillover effects in experimental data. *Advances in experimental political science* 289 (2021), 319.

[4] Trust & Safety Professional Association. n.d.. Enforcement Methods and Actions. Retrieved September 9, 2024 from https://www.tspa.org/curriculum/ts-fundamentals/policy/enforcement-methods

[5] Trust & Safety Professional Association. n.d.. What is Content Moderation? Retrieved September 9, 2024 from https://www.tspa.org/curriculum/ts-fundamentals/content-moderation-and-operations/what-is-content-moderation

[6] Trust & Safety Professional Association. n.d.. What Is Policy and Why Does It Matter? Retrieved September 9, 2024 from https://www.tspa.org/curriculum/ts-fundamentals/policy/policy-development

[7] Katherine Avery, Amir Houmansadr, and David Jensen. 2024. The Effect of Alter Ego Accounts on A/B Tests in Social Networks. In *Companion Proceedings of the ACM on Web Conference 2024*. 565–568.

[8] Priyanjana Bengani, Jonathan Stray, and Luke Thorburn. 2022. What's right and what's wrong with optimizing for engagement. *Understanding Recommenders, Apr* (2022).

[9] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[10] Elizabeth Berger and Kent Scheidegger. 2021. Sentence length and recidivism: A review of the research. *Social Science Research Network* (2021).

[11] Kiran Bhat. 2024. Deploying ML for Voice Safety. Retrieved September 5, 2024 from https://corp.roblox.com/newsroom/2024/06/deploying-ml-for-voice-safety

[12] Manuel Bronstein and Dan Sturman. 2023. RDC 2023: Where Roblox is going next. Retrieved September 5, 2024 from https://corp.roblox.com/newsroom/2023/09/rdc-2023-roblox-going-next

[13] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 2008. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1101–1110.

[14] Bryce Cai, Sean Decker, and Crystal Zheng. 2019. The migrants of reddit: an analysis of user migration effects of subreddit bans. *preprint* (2019).

[15] Robyn Caplan. 2018. Content or context moderation? (2018).

[16] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[17] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* 1, CSCW (2017), 1–22.

[18] Neeraj Chatlani, Arianna Davis, Karla Badillo-Urquiola, Elizabeth Bonsignore, and Pamela Wisniewski. 2023. Teen as research-apprentice: A restorative justice approach for centering adolescents as the authority of their own online safety. *International Journal of Child-Computer Interaction* 35 (2023), 100549.

[19] CY Cyrus Chu, Sheng-cheng Hu, and Ting-yuan Huang. 2000. Punishing repeat offenders more severely. *International Review of Law and Economics* 20, 1 (2000), 127–140.

[20] Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. 2024. The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference*. 85–93.

[21] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[22] David A Dana. 2001. Rethinking the puzzle of escalating penalties for repeat offenders. *The Yale Law Journal* 110, 5 (2001), 733–783.

[23] Alex Deng, Ulf Knoblich, and Jiannan Lu. 2018. Applying the Delta method in metric analytics: A practical guide with novel ideas. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 233–242.

[24] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 123–132.

[25] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: conception, theory and measurement. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–39.

[26] Lisa Fazio, DG Rand, Stephan Lewandowsky, Mark Susmann, Adam J Berinsky, AM Guess, Panayiota Kendeou, Benjamin Lyons, JM Miller, Eryn Newman, et al. 2024. Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. (2024).

[27] James C Fell and Michael Scherer. 2017. Administrative license suspension: Does length of suspension matter? *Traffic injury prevention* 18, 6 (2017), 577–584.

[28] Casey Fiesler, Jialun Jiang, Joshua McCann, Kyle Frye, and Jed Brubaker. 2018. Reddit rules! characterizing an ecosystem of governance. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[29] Camille François. 2020. Actors, behaviors, content: A disinformation ABC. *Algorithms* (2020).

[30] Elaine Freer. 2013. First time lucky? Exploring whether first-time offenders should be sentenced more leniently. *The Journal of Criminal Law* 77, 2 (2013), 163–171.

[31] Zihan Gao and Jacob Thebault-Spieker. 2024. Investigating Influential Users' Responses to Permanent Suspension on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–41.

[32] Eric Goldman. 2021. Content moderation remedies. *Mich. Tech. L. Rev.* 28 (2021), 1.

[33] Shelby Grossman, Jeff Hancock, Alex Stamos, and David Thiel. 2021. Introducing the Journal of Online Trust and Safety. *Journal of Online Trust and Safety* 1, 1 (2021).

[34] Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, et al. 2019. Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* 21, 1 (2019), 20–35.

[35] Hussam Habib, Maaz Bin Musa, Muhammad Fareed Zaffar, and Rishab Nithyanand. 2022. Are Proactive Interventions for Reddit Communities Feasible?. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 264–274.

[36] Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM web conference 2023*. 2666–2676.

[37] Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. 2023. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS nexus* 2, 3 (2023), pgad035.

[38] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[39] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–30.

[40] Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. Procedural justice and self governance on Twitter: Unpacking the experience of rule breaking on Twitter. *Journal of Online Trust and Safety* 1, 3 (2022).

[41] Yubo Kou, Renkai Ma, Zinan Zhang, Yingfan Zhou, and Xinning Gui. 2024. Community Begins Where Moderation Ends: Peer Support and Its Implications for Community-Based Rehabilitation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[42] Alex Leavitt. 2015. " This is a Throwaway Account" Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 317–327.

[43] Alex Leavitt, Joshua Clark, and Dennis Wixon. 2016. Uses of multiple characters in online games and their implications for social network methods. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 648–663.

[44] Alex Leavitt and Kat Lo. 2023. A Comparative Analysis of Platform Reporting Flows. (2023). Trust & Safety Research Conference.

[45] Christina LiCalsi, David Osher, and Paul Bailey. 2021. Brief: An Empirical Examination of the Effects of Suspension and Suspension Severity on Behavioral and Academic Outcomes. https://www.air.org/sites/default/files/2021-08/NYC-Suspension-Effects-Behavioral-Academic-Outcomes-Brief-August-2021.pdf

[46] Tony Liu, Lyle Ungar, Konrad Kording, and Morgan McGuire. 2024. Measuring Causal Effects of Civil Communication without Randomization. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 958–971.

[47] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2021. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* 27, 1 (2021), 1.

[48] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[49] Carrie Menkel-Meadow. 2007. Restorative justice: What is it and does it work? *Annu. Rev. Law Soc. Sci.* 3, 1 (2007), 161–187.

[50] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[51] Heather L O'Brien, Ido Roll, Andrea Kampen, and Nilou Davoudi. 2022. Rethinking (Dis) engagement in human-computer interaction. *Computers in human behavior* 128 (2022), 107109.

[52] Digital Trust & Safety Partnership. 2023. Trust & Safety Glossary of Terms. Retrieved September 9, 2024 from https://dtspartnership.org/wp-content/uploads/2023/07/DTSP_Trust-Safety-Glossary_July-2023.pdf

[53] Raymond Paternoster. 2019. How much do we really know about criminal deterrence? In *Deterrence*. Routledge, 57–115.

[54] William J Rauch, Paul L Zador, Eileen M Ahlin, Jan M Howard, Kevin C Frissell, and G Doug Duncan. 2010. Risk of alcohol-impaired driving recidivism among first offenders and multiple offenders. *American journal of public health* 100, 5 (2010), 919–924.

[55] Joseph Reagle. 2023. Even pseudonyms and throwaways delete their Reddit posts. *First Monday* (2023).

[56] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

[57] Ronald E Robertson. 2022. Uncommon yet consequential online harms. *Journal of Online Trust and Safety* 1, 3 (2022).

[58] Roblox. 2024. California AB 587 Roblox Transparency Report Q1-Q2 2024. Retrieved September 9, 2024 from https://oag.ca.gov/sites/default/files/Roblox%20AB%20587%20Report%20-%20Q1-2%202024.pdf/Roblox%20AB%20587%20Report%20-%20Q1-2%202024.pdf

[59] Roblox. 2024. Introducing the Ban API and Alt Account Detection. Retrieved September 5, 2024 from https://devforum.roblox.com/t/introducing-the-ban-api-and-alt-account-detection/3039740

[60] Roblox. 2024. Roblox Reports Second Quarter 2024 Financial Results. Retrieved September 5, 2024 from https://ir.roblox.com/news/news-details/2024/Roblox-Reports-Second-Quarter-2024-Financial-Results/default.aspx

[61] Roblox. n.d.. How to Report Rule Violations. Retrieved September 5, 2024 from https://en.help.roblox.com/hc/en-us/articles/203312410-How-to-Report-Rule-Violations

[62] Roblox. n.d.. Roblox Community Standards. Retrieved September 5, 2024 from https://en.help.roblox.com/hc/en-us/articles/203313410-Roblox-Community-Standards

[63] Roblox. n.d.. Safety & Civility at Roblox. Retrieved September 5, 2024 from https://en.help.roblox.com/hc/en-us/articles/4407444339348-Safety-Civility-at-Roblox

[64] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 742–753.

[65] Joseph P Ryan, Laura S Abrams, and Hui Huang. 2014. First-time violent juvenile offenders: Probation, placement, and recidivism. *Social Work Research* 38, 1 (2014), 7–18.

[66] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. " Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[67] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[68] Mariapina Trunfio and Simona Rossi. 2021. Conceptualising and measuring social media engagement: A systematic literature review. *Italian Journal of Marketing* 2021, 3 (2021), 267–292.

[69] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on human-computer interaction* 4, CSCW2 (2020), 1–22.

[70] Viktorya Vilk and Kat Lo. 2023. *Shouting into the Void; Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It.* Technical Report. PEN America.

[71] Geoffry D White, Gary Nielsen, and Stephen M Johnson. 1972. TIMEOUT DURATION AND THE SUPPRESSION OF DEVIANT BEHAVIOR IN CHILDREN 1. *Journal of Applied Behavior Analysis* 5, 2 (1972), 111–120.

[72] Sijia Xiao, Coye Cheshire, and Niloufar Salehi. 2022. Sensemaking, support, safety, retribution, transformation: A restorative justice approach to understanding adolescents' needs for addressing online harm. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[73] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing interpersonal harm in online gaming communities: The opportunities and challenges for a restorative justice approach. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–36.

[74] Mustafa Mikdat Yildirim, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. 2023. Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics* 21, 2 (2023), 651–663.

[75] Kenji Yokotani and Masanori Takano. 2022. Effects of suspensions on offences and damage of suspended offenders and their peers on an online chat platform. *Telematics and Informatics* 68 (2022), 101776.

Table 3: Experiment 1 results. Variant values are CUPED-adjusted means.

Outcome Variable	1-Hour Suspension	1-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate	0.270	0.252	(-0.021, -0.016)	(-7.584, -5.897)	<0.0001
Number of consequences	0.667	0.625	(-0.057, -0.028)	(-8.395, -4.36)	<0.0001
Number of reports against	4.462	4.257	(-0.329, -0.082)	(-7.271, -1.941)	0.00171
Time-to-reoffense (hours)	150.000	160.572	(9.121, 12.022)	(6.048, 8.048)	<0.0001
Days active	11.238	11.216	(-0.06, 0.016)	(-0.532, 0.143)	0.25916
Total time spent (hours)	131.951	130.953	(-2.664, 0.667)	(-2.012, 0.498)	0.25916

Table 4: Experiment 2 results. Variant values are CUPED-adjusted means.

Outcome Variable	1-Day Suspension	3-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate	0.472	0.434	(-0.042, -0.035)	(-8.897, -7.362)	<0.0001
Number of consequences	1.624	1.569	(-0.095, -0.015)	(-5.804, -0.969)	0.01117
Number of reports against	10.501	10.327	(-0.504, 0.157)	(-4.764, 1.458)	0.30376
Time-to-reoffense (hours)	128.439	140.432	(10.598, 13.387)	(8.203, 10.471)	<0.0001
Days active	12.384	12.168	(-0.271, -0.162)	(-2.187, -1.309)	<0.0001
Total time spent (hours)	327.031	331.804	(-2.481, 12.028)	(-0.781, 3.7)	0.23659

Table 5: Experiment 1 reoffense rate after 7, 14, and 21 days.

Outcome Variable	1-Hour Suspension	1-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate after 7 Days	0.205	0.186	(-0.022, -0.016)	(-10.426, -8.08)	<0.0001
Reoffense rate after 14 Days	0.308	0.290	(-0.021, -0.014)	(-6.842, -4.527)	<0.0001
Reoffense rate after 21 Days	0.386	0.374	(-0.018, -0.006)	(-4.606, -1.558)	0.00012

Table 6: Experiment 2 reoffense rate after 7, 14, and 21 days.

Outcome Variable	1-Day Suspension	3-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate after 7 Days	0.382	0.335	(-0.051, -0.042)	(-13.285, -11.189)	<0.0001
Reoffense rate after 14 Days	0.539	0.501	(-0.044, -0.033)	(-8.169, -6.175)	<0.0001
Reoffense rate after 21 Days	0.630	0.601	(-0.037, -0.021)	(-5.882, -3.303)	<0.0001

Table 7: Experiment 1 sub-group analysis. Variant values are CUPED-adjusted means.

Outcome Variable	Cohort	1-Hour Suspension	1-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate	0 violations	0.128	0.112	(-0.02, -0.013)	(-15.103, -10.182)	<0.0001
Reoffense rate	1–4 violations	0.209	0.190	(-0.023, -0.015)	(-10.773, -7.199)	<0.0001
Reoffense rate	5+ violations	0.415	0.397	(-0.022, -0.014)	(-5.369, -3.376)	<0.0001
Total time (hrs)	0 violations	29.107	27.630	(-1.802, -1.151)	(-6.163, -3.984)	<0.0001
Total time (hrs)	1–4 violations	49.221	47.486	(-2.183, -1.286)	(-4.42, -2.629)	<0.0001
Total time (hrs)	5+ violations	267.155	266.688	(-4.495, 3.56)	(-1.681, 1.331)	0.82011

Table 8: Experiment 2 sub-group analysis. Variant values are CUPED-adjusted means. Note that users with 0 historical violations were not eligible for this experiment, by definition.

Outcome Variable	Cohort	1-Day Suspension	3-Day Suspension	Absolute Diff. 95% CI	Relative Diff. 95% CI	Adjusted P-value
Reoffense rate	1–4 violations	0.315	0.271	(-0.051, -0.038)	(-16.01, -12.104)	<0.0001
Reoffense rate	5+ violations	0.532	0.496	(-0.041, -0.032)	(-7.597, -5.966)	<0.0001
Total time (hrs)	1–4 violations	47.706	43.724	(-4.642, -3.323)	(-9.667, -7.03)	<0.0001
Total time (hrs)	5+ violations	433.354	441.513	(-1.833, 18.151)	(-0.453, 4.218)	0.1095