Tennessee v. Meta
Case No. 23-1364-IV
PX-03081

Sentiment and Engagement: Towards a New Approach to Goaling for Facebook App

By David [redacted], Denise [redacted] & Anneke Buffone[1]

Published 2022

We thank these collaborators for their insights, edits,and critiques: Ana [redacted], Anand [redacted], Nicole [redacted], Jake [redacted], Jacky [redacted], Colin [redacted], and Xiuhua [redacted]. In addition, thank you to Katie [redacted], Jacqueline [redacted], Conrad [redacted], Chris [redacted], Katie [redacted], Jason [redacted], Charlie [redacted], Franz [redacted], and Olivia [redacted] for your thoughtful advice and comments.

Everyone working at Meta wants to build products that both improve people's lives and form the basis of a sustainable business. As the metaverse comes more into focus, our responsibility to build healthy products becomes ever more important. The purpose of this series of notes is to lay out the argument that we need to improve our goaling strategy to continue to effectively serve our users and foster the long-term health of our business. It is based on the views and inputs of a mix of long-term employees and newer hires, and a range of functions including UXR, DS, Engineering, Product Management, and Design.

This research presented here is based on a collection of studies on sentiment and engagement, each investigating, and shedding light on, one piece of a larger puzzle. As such, we synthesize, theorize, and strategize in this space with incomplete insights. This note is intended to start a discussion, not be the final word on how to solve user needs related to sentiment and engagement.

Summary

Facebook App has invested over the last several years in a suite of survey-based sentiment measures of user experience, including direct measures such as Facebook App Value (FAV), Feed Satisfaction Survey (FSS), Tracking Reach of Integrity Problems Survey (TRIPS) and Interest Relevance Survey (IRS); and survey-based behavioral proxy measures such as Meaningful Social Interactions (MSI) and Meaningful People (MP).

However, despite this investment, direct behavioral metrics (particularly sessions) are still the most important factors that determine what changes we launch and where we invest resources. There are several reasons for this:

[1] Authors contributed equally

- Inadequacies of survey measures which do not adequately capture concrete user experience, but instead measure satisfaction (closer to brand sentiment)
- Difficulties operationalizing the existing survey metrics
- Recent company focus on optimizing for sessions due to [HYPERLINK "https://fb.workplace.com/work/file_viewer/410679217406421/?surface=POST_ATTACHMENT" \h]
- Preference for simple metrics has led to less emphasis on survey-based behavioral proxy measures such as MSI

Unfortunately, there are issues with this state of affairs. We need to capture user needs, wants, and product evaluations and use these explicit signals to build and evaluate our ranking system, to better optimize our product for good user experience. Much of the research we have reviewed indicates that existing behavioral metrics are weakly correlated, if not negatively correlated, with sentiment measures. In the first half of this note, we summarize existing evidence that the metrics we currently optimize for are not necessarily well-aligned with sentiment.

In the second half of this note, we propose a set of changes to increase how accurately we measure user experience and thus how well we are able to evaluate product work that aims to improve that experience:

- Conduct more research to evaluate the association between explicit signals (e.g., survey measures) and implicit measures (e.g., behavioral measures) that feed into ranking and experimentation.
- Conduct research to improve our approaches for understanding user evaluations of our product.
    - Develop survey measures that are product-, session-, or content-level specific, or capture specific concepts (e.g., TRIPS for perceived integrity problems).
    - Design surveys that better facilitate user recall of actual experiences on app.
    - Develop multi-question scales (as opposed to single-question measures)
- Build measures of user value that blend behavior and survey data (e.g., quality sessions).
- Place greater importance on existing measures that incorporate these principles (e.g., TRIPS, Interest Relevance Survey).

The ultimate goal is to create a research org that helps us better inform company and product strategy. We should pursue research that a) measures concrete user pain points and delights, (b) allows for driver analysis to detect how to move brand sentiment vs. product experience satisfaction, (c) identifies the most important markets and user segments for growth, (d) informs optimal relative investment in safety/integrity vs. product growth, (e) identifies strategy for long-term user growth and retention using longitudinal and repeated measures research

One important topic left for future writing is whether better optimizing for user value would build a more sustainable business long-term. One [HYPERLINK "https://fb.workplace.com/groups/164146357389025/posts/1319206745216308/" \h] is that the long-term News Feed Integrity holdout shows a **drop in sessions but an increase in Daily Active People**.

3

The Challenge With Behavioral Optimization

The Role of Behavioral Metrics

The most important FB App goals are behavioral. The H2 2021 goal related to "Creating value for every day lives" states that "the FB App strategy will continue to focus on increasing the value that our app provides to users. To track progress to this goal, we will once again use DAU and Core Sessions and set goal targets both globally and for US&CA." The survey-based measure Facebook App Value is listed as a target. Other top goals include CEDAU (engagement with groups) and user-to-user interactions (such as U2U comments), which have replaced previous goals based on the sentiment-aware metrics Meaningful People and Meaningful Social Interactions, respectively ([HYPERLINK "https://fb.workplace.com/groups/567552929959993/permalink/6439809502734277/" \h]).

Often, product launches are mainly evaluated based on behavioral metrics. Conceptually, product launches have a requirement that they need to demonstrate user value. In many cases, –both for simplicity's sake and to align multiple teams–user value is measured through behavior, particularly sessions. We rarely launch changes that regress sessions, even if there is a demonstrated change in sentiment. This is encoded in the contracts between teams -- the Social Responsibility (SR) team currently has a 0% sessions guardrail for integrity changes and also a 0% sessions guardrail for sentiment changes. Integrity issues are one exception, but sessions remains an important consideration. In the last year, a notable launch from Social Responsibility that was allowed to regress sessions was one that reduced the prevalence of civic content, based on explicit guidance from Mark earlier in the year. As an example of how a 0% sessions guardrail affects sentiment-related work, the Responsible Ranking & Legitimacy team within SR has been working to increase the influence of post-level sentiment models such as p(Worth Your Time) and p(Inspire). In order to get a potentially launchable candidate, we had to go through [HYPERLINK "https://fb.workplace.com/groups/592508124224763/posts/2058064217669139/" \h], primarily to find candidates that minimized sessions' regressions; this process both weakens the resulting changes and imposes a much larger burden of time and effort for sentiment-related changes than for ones based on behavioral metrics.

The recent discussion around unconnected edge stories (UCEs) is one example of how prioritization of engagement over sentiment could impact product development. In H1 2021, the News Feed ranking team ran a suite of experiments designed to drive improvements in the FSS surveys ([HYPERLINK "https://fb.quip.com/pM7FASyB5UU9" \h], jump to "The Results" for the most relevant section to this discussion). Because of the lack of power, not many of the outcomes were statistically significant. However, a few of experiments did seem promising, most notably the removal of UCEs, with a statistically significant improvement in FSS of 0.75%. This improvement came at a cost of 0.2% sessions, and also seemed to vary by region (though likely was underpowered for the latter conclusion). Because this session drop was judged to be unacceptable, no further experimentation was run along these lines. Since then, UCEs have emerged as a major cause of high prevalence of engagement bait content in the [HYPERLINK

"https://transparency.fb.com/data/widely-viewed-content-report/" \h]. It's also [HYPERLINK "https://fb.workplace.com/photo/?fbid=412372199571256&set=pcb.406718093434010" \h] that UCEs score much lower than other content (at least 1 point lower on a 5 point scale on average). Dan Day has a thorough [HYPERLINK "https://fb.workplace.com/notes/955136095385791" \h] of the discussion around UCEs to date -- see particularly the "did not work" section to get a sense of the decision-making process (which focuses on the impact to sessions and MSI). The discussion around UCEs continues, but regardless of the ultimate outcome, it's very clear that the burden of proof for making sentiment-based changes at the cost of behavioral metrics is extraordinarily high.

Another [HYPERLINK "https://fb.workplace.com/groups/446124469463153/posts/818777335531196/" \h] that illustrates this dynamic is the disabling of a long-standing filter that prevents birthday posts from friends to non-friends from being eligible for feed. The launch was motivated by sessions gains but was recently rolled back because of employee complaints about the resulting stories being a bad experience.

When ranking systems are primarily based on behavior. We've [HYPERLINK "https://about.fb.com/news/2019/04/people-publishers-the-community/" \h] about p(Worth Your Time) for a while now, and feature it prominently when describing how ranking works. For example, Nick Clegg's [HYPERLINK "https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2" \h] from earlier this year gives "likes" and "worth your time" as examples of signals we use for ranking, before going on to say that we don't try to keep users glued to their smartphones. But the reality of how News Feed works is very different. While we have tried to [HYPERLINK "https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2" \h], we have had mixed success. p(Worth Your Time) has yet to become a key factor in ranking, and its importance has actually declined the last couple years, as has the importance of p(MSI). These patterns suggest renewed attention to users' attitudes, and finding mechanisms to accurately measure those attitudes is necessary to maintain the centrality of user value in our ranking models.

Currently, the dominant model in ranking, by far, has become the "linear VPVD" (Viewport View Duration) model, which predicts how long a user will spend looking at a post. The primary way we measure importance of different factors is using the [HYPERLINK "https://www.internalfb.com/intern/unidash/dashboard/feed_analyzer/main/" \h], which runs offline experiments which measure the degree to which News Feed ranking would change if we were to remove various ranking factors. The [HYPERLINK "https://pxl.cl/1SCj8" \h] shows Normalized Discounted Cumulative Gain, which assigns greater weight to changes near the top of News Feed. According to this measure, the clean linear VPVD term alone (only one of many terms that capture time spent) has 50% more importance than the next highest factor, MSI. Meanwhile, the most prominent term that involves Worth Your Time (linear_vpvd, which combines time spent prediction with other factors) has about 1/10th the importance of clean linear VPVD, while the most important survey-based model (group_irs_value, where irs stands for Interest Relevance Survey) has less than 1/4th the importance of clean linear VPVD. (Note:

according to "[HYPERLINK "https://www.internalfb.com/intern/px/p/1SCkF/" \h]", which weighs all positions equally, Worth Your Time fares a bit better in terms of importance, but the high-level picture remains the same.)

The two direct causes for the current balance of behavioral vs. survey-based models are 1) the difficulty of launching survey-based changes, discussed above, and 2) the focus on sessions as a top-level metric over the last couple years leading to a series of launches that greatly increase the importance of time-based terms in ranking.

Note that different features within FB App do vary in the extent to which they incorporate sentiment, and so the situation within News Feed may not be representative of every area -- though as discussed above, FB App often goals on behaviors and sessions plays a prominent role for many products. Different features within FB App vary in the extent to which they incorporate user evaluations, so the situation in Feed isn't a complete picture of product development within FB App. However, most product teams do place primary importance on behavioral metrics, and especially sessions. Combining the prominence of Feed with the general approach to goaling, the overall picture is currently one in which behavioral metrics play a much larger role than survey-based measures.

The Reasons for Behavioral Metrics

The most popular behavioral metrics have several attractive properties:

- (Often very) large volume -- yielding metrics with high statistical power
- Relative sensitivity to product changes -- and thus more actionable
- Stability -- a comment means more or less the same thing it did 10 years ago
- Simplicity and intuitiveness -- easy for stakeholders and leadership to understand (though potentially a false sense of understanding)

The reason for the importance of sessions in particular is a bit less obvious. Sessions was [HYPERLINK "https://fb.workplace.com/notes/338798390550983/" \h] as a metric in 2018, primarily to address issues with the existing time spent metric. At the time, optimizing for time spent resulted in a content mix that was particularly video-heavy and appeared to be unhealthy as evaluated through a range of other metrics, including sessions, DAP, and other engagement metrics. A key motivating argument for sessions was an [HYPERLINK "https://fb.workplace.com/photo/?fbid=434231740344990&set=pcb.1849026315145975" \h] between number of sessions and Daily Active People (though the same relationship was also observed for time spent).

These factors explain the existence of sessions as an important metric, but don't fully explain the central role of sessions in FB app decision making. There are likely at least three reasons we turn to sessions to make product decisions:

- [HYPERLINK "https://fb.quip.com/VSgcAjrbKsFy" \h] between sessions and user sentiment in 2018 suggested user value.

- As FB App has diversified its product offerings over the last five years, there is a perceived need for a "common currency" between different efforts / products (see [HYPERLINK "https://fb.workplace.com/notes/1094159301063654" \h]).
- Usage is good for the business, as users will come back to products they find valuable, and sessions is one of the most direct measurements of usage.

Through some combination of these three factors, sessions is now commonly taken to be, by definition, the best way to capture user value. The typical reasoning goes, "The reason why product change X leads to more sessions is because after experiencing the new version of the product, people are more satisfied and/or excited with their experience. As a result, they choose to open the FB App more often." Undoubtedly this story is true in some cases -- for example, a less-engaged user who newly starts interacting with the dating portion of the app is quite likely to visit more often. But in other cases, this story is neither proven nor particularly logical (see the next section for further discussion).

One final note on MSI. At a superficial level, the shift (in 2018) towards and then (the last few years) away from MSI as a primary metric appears to be a move towards and then away from sentiment-based approaches. And in a limited fashion, this is true -- if nothing else, this tracked a move towards and then away from user-to-user interactions. But though MSI is indeed "research-informed" (the research on active vs. passive consumption) and "survey-informed" (surveys measuring meaningfulness of different interactions), the most important decision in MSI -- namely, the ratio of importance placed on comments vs. likes -- has always been based more on other factors -- namely, sessions and original posting. The survey work suggested comments being roughly two times as meaningful as likes; the MSI metric places fifteen times more on comments. Thus, MSI is better understood not as an outlier in terms of Facebook putting user sentiment first, but rather as one of many metrics developed over the years that aim to capture the success of the product that Facebook was trying to build at the time. And even with this understanding, the link between MSI and sentiment continues to fray -- the News Feed team just launched [HYPERLINK "https://fb.workplace.com/groups/592508124224763/posts/2188066928002200/" \h] to remove the friend closeness multiplier from the MSI score, despite [HYPERLINK "https://fb.workplace.com/groups/726657910872920/posts/763918310480213/" \h] clearly indicating that people find interactions with close friends much more meaningful.

The Complex Relationship Between Behavior and Sentiment

The sessions metric is not always well-aligned with user evaluations. While part of the motivation for sessions is [HYPERLINK "https://fb.quip.com/VSgcAjrbKsFy" \h], alignment between sessions and sentiment is not always positive. Mixed results have been found in a

number of contexts (see examples in table below).

Sampling of Past Investigations of Sessions and User Evaluations

User Behavior	User Evaluation	Relationship
sessions (+)	Reported ease of use; FB is fun	Strong, positive (2018)
sessions (+)	Satisfaction w/ FB; good for the world; trustworthiness; usefulness	Weak, positive (2018)
Marketplace sessions (+)	FAV	Strong, positive (2020)
Events sessions (+)	FAV	Strong, positive (2020)
sessions (-)	Connection; useful; fun; trust	Negative (2018)
sessions (+)	Problematic use; loneliness	"U"shaped trend (2018)
sessions (+)	Spend too much time on FB; uncivil comments	Positive (2018)
sessions (+)	Composite sentiment score	Negative (2021)
sessions (+)	Feeling sad	Positive (2020)
Groups sessions (+)	FAV	Negative (2020)
sessions	Value	None (2021)
sessions	Worth your time	None (2021)
sessions	MSI	None (2021)

Sessions increase related to positive attitudes or sessions decrease related to increase in negative attitudes

Sessions increase related to increase in negative attitudes or decrease in positive attitudes

Sessions increase or decrease not related to attitudes

It is understandable that not all sessions will necessarily result in positive user attitudes. For example, not finding what you're looking for can increase sessions with neutral or negative user value. Our product design can contribute to these mechanisms unintentionally either through direct means -- such as, excessive badging -- or indirectly -- for example, by creating (although inadvertently) a [HYPERLINK "https://www.simplypsychology.org/schedules-of-reinforcement.html" \h], where a user encounters good content occasionally at unpredictable intervals, but stale or irrelevant content most of the time, would likely translate to high sessions but low user value.

Even research showing a positive relationship between sessions and user attitudes reveals a complex relationship. The positive association between sessions and various outcomes, but [HYPERLINK "https://fb.quip.com/IZHOAdPXww9Y" \h]; past that point the same outcomes began declining (the current average number of sessions per user is around eighteen). To solve this problem, in 2021 session optimization was capped through the [HYPERLINK "https://fb.workplace.com/notes/1119991598435664/" \h] metric at 50 visits per user per day and only measures sessions longer than 3 seconds. However, these limits were quite conservative, and analysis showed that Core Sessions could still lead to product work that focused on high session-count users ([HYPERLINK "https://fb.workplace.com/groups/558331381190793/posts/1485257618498160/" \h]). Even visiting Facebook 50 times in a single day is quite a lot. Recently a new metric has been introduced that caps sessions at 15. Since sessions was introduced, there have been several years of steady optimization for sessions, and from a purely **mathematical** standpoint, this suggests that further optimization for sessions is unlikely to correlate with "true" user value -- this general phenomenon is often referred to as [HYPERLINK

HIGHLY CONFIDENTIAL (COMPETITOR)

"https://en.wikipedia.org/wiki/Goodhart%27s_law" \h]. In other words, optimization for a metric can lead to the degradation of that metric over time.

Importantly, when the sessions metric was originally introduced, a number of "informational metrics" were introduced in order to avoid launches that were likely to "game" sessions or increase non-valuable sessions. For example capping sessions at 30 per day and only counting sessions at least 15 seconds in length were intended to support positive sessions ([HYPERLINK "https://fb.workplace.com/notes/330286781612578/" \h]). However, unintentionally over time these counter-metrics were gradually marginalized and are no longer part of standard launch evaluation (if they exist at all). Currently, there are no broadly useful taxonomies of sessions that can be used to understand the types of visitation that a particular change is driving. However, a UXR workstream started in 2020 has worked to understand what makes a quality session. It has shown that individual sessions vary greatly in quality ([HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=429012491454518&id=100030373125590" \h] [HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=388562886218953&id=100051957353177" \h]). This workstream has developed a [HYPERLINK "https://fb.workplace.com/groups/375376650972558/" \h] metric and has had a [HYPERLINK "https://docs.google.com/document/d/10yxrUE08Z7MUSx_eIoO1ZPA8LYdX99TS0tjElkb6KKE/edit" \h]. See also [HYPERLINK "https://fb.workplace.com/notes/2666602090244583" \h] for more on this topic.

While sessions and user evaluations are not always misaligned, it's important to recognize that in cases where they are, engagement is often prioritized. As discussed above, in the "Role of Behavioral Metrics", behavioral metrics play a more important role in decision-making than survey metrics. As a result, a number of product launches which show positive gains in user attitudes have been blocked due to declines in sessions metrics. This means that only when sessions are positive (with either positive user evaluations or negative user evaluations) do we realistically launch products. This makes user evaluations inconsequential to product launch decisions, although they are still used to inform product design.

Other investigations into user engagement and attitudes. The few existing studies that attempt to measure the overall impact of Facebook usage also show mixed results. For example, [HYPERLINK "http://web.stanford.edu/~gentzkow/research/facebook.pdf" \h] found that reduced Facebook use increases positive outcomes for people, such as increased self-reported subjective well-being, and more time spent socializing offline, although there are also important costs such as feeling less informed. Active engagement, particularly liking a post, [HYPERLINK "https://www.internalfb.com/intern/anp/view/?id=147556" \h] to have moderately positive association with post-level Worth Your Time survey scores. On the other hand, many passive consumption events, most notably Viewport View Duration (VPVD), show little to no correlation with Worth Your Time (see also [HYPERLINK "https://fb.workplace.com/groups/2097929377105943/posts/2248759225356290/" \h]). There are many reasons why a user might stop to view a post -- some are good (the post is interesting), some are bad (the post is offensive). Similarly, our current ranking terms assume

Commented [1]: https://fb.workplace.com/notes/395124579217778 there is some work on the communities side that has seen some association here

that a comment is a positive signal, whether the interaction is good (offering condolences for a lost loved one) or bad (debunking a hoax). Overall, research shows that engagement either active, passive, or simply opening the app (a session) can be a positive, negative, or neutral experience for users, suggesting that goaling on number of engagement alone is unlikely to consistently create positive user experiences.

Commented [2]: there is discounting in both ranking models + accounting (msi deltoid metrics) for many types of useless/harmful comments - deeptish mukherjee would have the most up to date info, but fct_comment_quality has a quick overview of discounted types

… But Sentiment Measures Have Issues as Well

In the previous section we explored the issues with the status quo. But the answer isn't just paying more attention to existing sentiment metrics -- in order to move beyond a development process dominated by behavioral metrics, we will need to improve the existing approaches for measuring sentiment.

But the answer isn't just paying more attention to existing sentiment metrics – current sentiment measures have a number of issues that need to be addressed. This section lays out the issues, while the following section explores potential solutions.

General Questions Don't Work Well for All of Facebook

Operationally, having one sentiment metric to "[HYPERLINK "https://fb.workplace.com/notes/338798390550983/" \h]" would be very convenient, just as sessions has become the default behavioral metric. Unfortunately, high-level sentiment metrics haven't worked out very well in practice.

It is difficult to meaningfully summarize opinions about different Facebook features. Facebook App Value is a well-developed survey, but it asks a user to take all sorts of different experiences from all sorts of different products and condense them down into a single rating. "Value" from News Feed is likely quite different from "value" for Dating -- and to get a final "value" for Facebook, the user also has to mentally weigh the importance of different features. As a result, the cognitive effort required to accurately answer this question is much higher than most users are likely to make. Consider trying to evaluate sentiment towards Unilever, a conglomerate focusing on food and home goods. One could easily write a survey on how much a given person values Lipton tea, but it would be much harder to accurately measure perceived value of Unilever's entire product suite. You might love Ben and Jerry's but hate Popsicles. Trying to calculate your average sense of quality of all Unilever products from Vegan Mayo, to Dove soap to Domestos household cleaners feels nigh impossible. Facebook's suite of products (News Feed, Marketplace, Watch, Groups, Dating, …) is in some ways even more varied.

People can't remember what they see and do on Facebook. Even if people could meaningfully summarize their experiences with Facebook products, they probably won't be able to recall those experiences when asked in a survey. Continuing the Unilever analogy, some of us might have a strong opinion about Ben and Jerry's Half-Baked ice cream -- but what about Hellman's mayo? What about their vegan version? Does their mayo taste any different from Miracle Whip? If we want measures that accurately capture people's experiences, we will

probably need to remind them of those experiences. A lot of what happens on Facebook is simply not memorable enough to lead to the kinds of memories that users can easily recall when answering surveys. For user behaviors that are more habitual (e.g., passive scrolling, opening up the app for a notification), there may be no recall at all ([HYPERLINK "https://www.ssoar.info/ssoar/bitstream/handle/document/55960/?sequence=1" \h]). News Feed is optimized for fast, bite-sized content which seems more in line with an automatic, more shallow, less deep ([HYPERLINK "https://www.scientificamerican.com/article/kahneman-excerpt-thinking-fast-and-slow/" \h]) cognitive processing that is less likely to lead to formation of specific memories. A recent cognitive testing study asking users to recall content from News Feed suggested that users have trouble remembering the composition of their Feeds ([HYPERLINK "https://docs.google.com/presentation/d/1Ee_OKnXxwhchpUudE7uxARcB7S2fxOmI/edit" \l "slide=id.p1" \h]). In a qualitative diary study, users did not recall various actions they took on Facebook, even right after having taken these actions ([HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=144440816663337&id=100032922579333" \h]). In the same study, users also appeared surprised when viewing their activity log, suggesting there may be multiple types of actions users forget.

Complex concepts can't be captured with single questions. Questions have to be relatively vague and high-level in order for one question to represent an entire social concept, for example community. But, if a question is broad enough to encompass multiple facets of one concept, then it is often so vague that product work is unlikely to move it substantially. Our [HYPERLINK "https://fb.workplace.com/groups/sentimentxfn/" \h] contains multiple one-question survey measures intended to represent a concept (such as [HYPERLINK "https://fb.workplace.com/notes/facebook-app-fyi/introducing-the-facebook-app-value-fav-metric/3072829962765598" \h] and [HYPERLINK "https://www.internalfb.com/intern/wiki/Sentiment-data/introduction-to-sentiment-metrics/" \h]). FAV asks "How much value do you get from using Facebook?" This question does not indicate what makes up value or, if the metric declines, what we can do at Facebook to increase user value. [HYPERLINK "https://medium.com/facebook-research/the-satisfaction-trap-35f94ee9d9d8" \h]) discusses this issue in more detail.

People fall back to brand sentiment when asked general questions. Because questions like "How much value do you get from using Facebook?" or "Overall, how satisfied are you with the posts you see in your News Feed?" likely do not trigger specific memories, they tend to function as association tests that capture general positive/negative brand sentiment related to Facebook overall, rather than specific in-product experiences ([HYPERLINK "http://repository.essex.ac.uk/12368/1/Albaghak.JSSAM.GreenAccessv1.pdf" \h]). General questions increase the salience of larger corporate issues that might be unrelated to someone's everyday experience. Even if I like Ben & Jerry's Chunky Monkey ice cream, that doesn't mean I approve of Unilever as a corporation.

In large part due to these issues, attempts to optimize for sentiment are often unsuccessful. While there have been some successes improving metrics such as CACTUS,

FSS, TRIPS, CAU, and [HYPERLINK "https://fb.workplace.com/groups/groupsresearchannouncements/permalink/2635363793414160/" \h], in general, these sorts of metrics have been hard to move. It often takes either a very strong product change or a multi-year holdout to detect changes in sentiment ([HYPERLINK "https://fb.workplace.com/groups/groupsresearchannouncements/permalink/2635363793414160/" \h]). For example, we have recently attempted to optimize for FAV (and product-specific counterparts such as FSS), but it has proven to be relatively insensitive and thus of limited use as a guide for decisions -- a general problem for goaling on sentiment metrics ([HYPERLINK "https://fb.workplace.com/notes/375033606961548/" \h]). Brand sentiment is generally very difficult to move either in the short-term or through individual product changes, and is influenced by factors external to actual Facebook experiences ([HYPERLINK "https://docs.google.com/presentation/d/1nG_KrkIypCPzf6CYx0okrewN_wRvpOt157y2aeoq7rM/edit?usp=sharing" \h]). Other teams intended to move sentiment have been deprecated (e.g. the [HYPERLINK "https://fb.workplace.com/groups/333174774260627" \h] team 2019 and the [HYPERLINK "https://www.internalfb.com/intern/wiki/Goodwill/" \h] team 2014-2019 to move CAU).

Potential Solutions

Making improvements will require taking the misalignment of engagement and sentiment seriously and deciding, top down, to make this an area of focus for Facebook. We will need to better understand the associations between user attitudes, engagement, and sessions, and use our new-found understanding to inform our choice of metrics. These are complex issues and won't have simple solutions. We recommend a multi-prong strategy:

- Conduct research to inform further improvements to our approaches for measuring user attitudes:
    - Develop survey measures that are
        - product-level, session-level, or content-level focused, and/or
        - capture specific concepts (e.g., TRIPS for perceived integrity problems)
    - Design surveys that better facilitate user recall of actual experiences on app
    - Develop multi-question scales (as opposed to single-question measures)
- Build measures of user value that blend behavior and sentiment (e.g., [HYPERLINK "https://fb.workplace.com/groups/375376650972558/permalink/396699315506958/" \h], [HYPERLINK "https://docs.google.com/document/d/1au7SBqf_yH8SYUbMuKTOrkZUd3kWJ6MetpKO4wCeCZ0/edit" \l "heading=h.cj8ylcz8ssvl" \h], multi-method measures)
- Place greater importance on existing measures that incorporate these principles (TRIPS, Interest Relevance Survey)

Include user evaluations more intentionally in product launch criteria

We want to ensure that the products we launch: provide user value; are as safe as possible; are inclusive; function as we expect; are understood by users; and do not create problematic experiences when used in conjunction with existing products. To promote these standards we

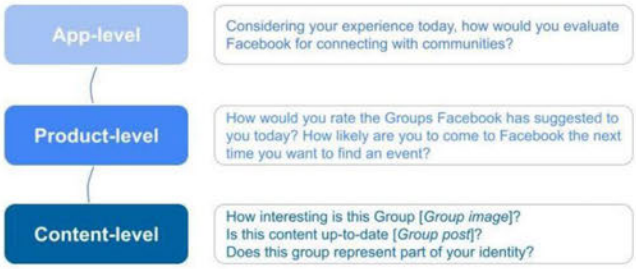
have developed an [HYPERLINK "https://docs.google.com/document/d/1-Mruiqj6R0zxef4apCfBkSAeOitGfxujffEB9JiVvkk/edit?usp=sharing" \h] for including research in product launch decisions.

Improve our survey measures

As discussed above, high-level sentiment metrics are often quite difficult to move, and thus can be difficult goals for product teams. Below, we describe a host of strategies for UXR to design more movable survey metrics for product teams.

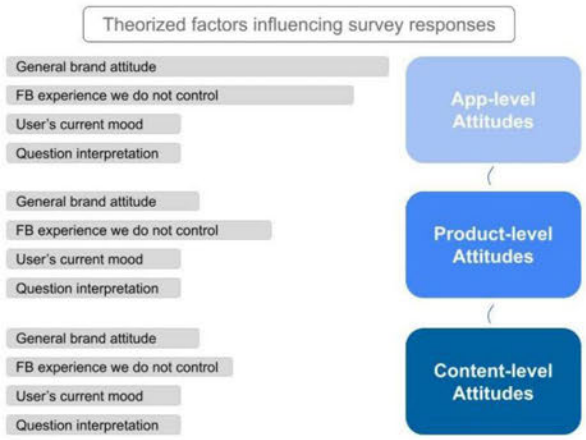
Encourage a tiered strategy for measuring attitudes. We should consider an alternative model of survey development that relies on concrete measures at the product level and include a clear understanding of how these ladder up to broader, well-defined sentiment measures. Hypothesized processes and theoretical dimensions of survey constructs should connect across product-level and app-level measures.

We should think of user attitudes as pertaining to the app, the product, and the content (e.g., Facebook-level, group-level, and post-level). Teams have had success developing more focused sentiment metrics, either through concentrating on specific features, content, and problems, or through focusing on more detailed in-product experiences, such as opinions about individual posts ([HYPERLINK "https://fb.workplace.com/groups/2346488082237140/permalink/2435884969964117/" \h]; [HYPERLINK "https://fb.workplace.com/groups/331605127611974/posts/463060641133088" \h]).[2]



There are various factors that influence survey responses. We typically think of these factors collectively as error, but they can be more or less influential depending on the type of questions we ask. The figure below shows several salient factors for Facebook--represented by the gray bars. The size of the gray bars represents how strongly we theorize each factor is related to survey responses based on the level (app, product, or content) of the question we ask.

[2] The [HYPERLINK "https://fb.workplace.com/notes/406959400706404/" \h] was planning to use a similar method before the team was deprecated in 2019.



Additional research quantifying how these latent factors (general brand attitude, Facebook experience we don't control, etc.) load differently at the different granularity level of questions would help UXR decide which type of question to use in which context. For example, a session-level survey that has both a session-specific satisfaction and higher level satisfaction could be a place to start quantifying this ([HYPERLINK "https://fb.quip.com/aXjpACmorvYW" \h]).

This strategy has already shown some success. Specifically, to evaluate civic demotions at the beginning of 2021, a survey was developed to measure different aspects of civic experience. The strongest demotion we tested resulted in a statistically significant decrease among survey measures identifying negative user experiences ([HYPERLINK "https://fb.quip.com/woe2A6jbSamZ" \h]), specifically the Lorax items "too many political posts on FB", "news that are false or misleading". However, this improvement to civic-related pain points did not translate to [HYPERLINK "https://fb.quip.com/ywbYAk6irFXY" \h] (FSS improved, but non-stat-sig). This example demonstrates it is possible to measure sentiment improvements by focusing on concepts related to the change being tested.

Reduce the influence of brand sentiment. In survey construction, we should choose the survey items that are least affected by brand sentiment. We should also make it standard practice when validating surveys to carefully consider when in a survey flow brand sentiment is measured compared to product-specific questions. Moreover, controlling for brand sentiment in predictive models is essential to reduce the degree that brand sentiment is represented in survey measures intended to capture user evaluations of specific products.

Develop measures that people can answer. We need additional research with the explicit goal of investigating how to ask people about behaviors and attitudes in a way that better separates recalling a unique experience on Facebook versus a stable generalized attitude. Specifically, we need to know which types of attitudes about Facebook are variable and can change versus those that are stable and will not change as a result of product changes.

Variable attitudes are more likely attitudes tied to an actual experience. If the experience is positive, then the attitude is likely positive, if the experience is negative, the attitude is negative. Stable attitudes are likely based upon prior experiences and are not easily changed.

The first step in this direction is to design survey questions with a concrete experience in mind. No question should be written if we cannot, in concrete terms, describe what users would experience, how they would feel, and how they might behave if they answer "strongly agree" to a given question. Research from the field of psychology has long shown that memories can be quite unreliable ([HYPERLINK "https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.3542" \h], [HYPERLINK "https://psycnet.apa.org/buy/1995-42833-001" \h]). For example, a user might incorrectly feel that they have seen a certain piece of content if it feels "typical" to what one might see on Facebook. We currently have multiple resources to help UXR implement cognitive testing during survey development (a [HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=4079551536452 34&id=100032922579333" \h], a [HYPERLINK "https://fb.workplace.com/groups/cognitivetesting" \h] group, and specific [HYPERLINK "https://fb.workplace.com/notes/644550702928261/" \h] for cognitive testing interviews), but we could go a step further.

We hypothesize that one way to do this is to provide an anchor for memory. To facilitate this work, we can develop various survey question stems and questions that work well, then encourage teams to add in their product context where appropriate. Language that could be used across teams is underlined in the examples below. Please note that these are rough drafts and have not been cognitively tested.

- *In the last week, did you seen more or fewer posts on Facebook that were false or misleading compared to the week before?*
  - *a lot more*
  - *a little more*
  - *about the same*
  - *a little less*
  - *a lot less*

- *How slow or fast has your Facebook News Feed been loading in the last day compared to how its usual loading speed?*
  - *a lot slower*
  - *a little slower*
  - *about the same*
  - *a little quicker*
  - *a lot quicker*

- *In the time you spent on Facebook today, did you have a satisfying conversation or interaction with anyone?*
  - *Yes*
  - *No*

Story gallery surveys are a method we need to test further to show whether they can trigger memory (as external research about visual aids suggests [HYPERLINK "https://www.tandfonline.com/doi/abs/10.1080/19312458.2020.1784402" \h]).

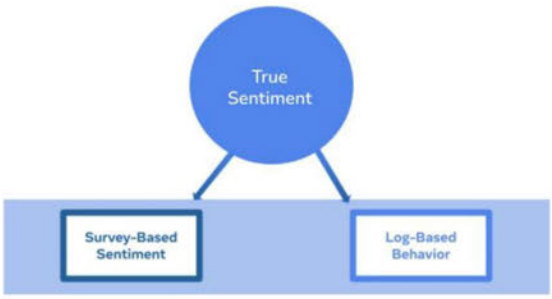
Use multiple questions for one concept. Attitudes about Facebook are latent constructs that we cannot observe directly and survey responses are an imperfect proxy to measure these attitudes (with varying degrees of error; [HYPERLINK "https://fb.workplace.com/notes/390355712128309/" \h]). This means concretely that we cannot understand a complex concept with a single question. Instead, we should rely on scales of at least three questions that can be combined into a composite. This gives us a chance to ask about various aspects of a construct (e.g., quality experience= quality of content, quality of interactions, quality of search functions, etc.).

Another multi-question strategy might be to use a tiered approach of questioning. We may start with asking people what the purpose of a given post is (e.g., to inspire, inform, entertain), and then focus on attributes related to that. For example, if a post is intended to inspire, is it doing that effectively, and does this person care about being inspired in the first place? This means asking multiple questions in a given survey that are each concrete and different, but collectively give insight to one concept.

Be more open to metrics evolving over time. To date, a strong premium has been placed on metrics that remain stable over time -- e.g., CAU, FAV, etc. The advantage of such a metric is that it allows for reliable longitudinal comparisons. However, as we have discussed, other metrics may be better-suited for evaluating product changes. Thus, for metrics designed to evaluate product work, and particularly metrics used for goaling, we should be open to improving these metrics regularly. Survey metrics may need to be updated if 1) survey methods change, 2) cultural shifts in language and understanding occur (e.g. see work on Communities; [HYPERLINK "https://fb.workplace.com/notes/833939057345888" \h]; [HYPERLINK "https://docs.google.com/presentation/d/1L4dLL2KFKJbyIPBmBWsjgYmoR6FBJ0c5KE0W4s2f5G8/edit" \l "slide=id.ge5b8d0d5eb_0_0" \h]), or 3) product priorities change.

Build measures that combine sentiment and behavior

Sentiment (as measured through surveys) and in-product behaviors are ultimately just different types of imperfect and noisy information that can be used to try to deduce a user's "true sentiment". In many cases, it can be beneficial to combine elements of both.



Investigate in which circumstances engagement and sentiment line up or are disparate. Then share these widely so product teams are familiar with them. For example, we see alignment between negative sentiment and low engagement. Users who show declines in FB use tend to have lower sentiment evaluations of FB ([HYPERLINK "https://fb.workplace.com/notes/369185774209186" \h]). In addition, the key correlates of churn and decreased FB usage is the belief that the things that they see on FB are largely negative ([HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=1571874787178 61&id=100032798297485" \h]). Users who report the lowest sentiment are more likely to delete their account in the following months ([HYPERLINK "https://fb.workplace.com/photo/?fbid=1044587466052498&set=pcb.1047144595774367" \h]). Users who report lower levels of perceived safety on platform are more likely to deactivate their accounts ([HYPERLINK "https://fb.workplace.com/permalink.php?story_fbid=140555893714353&id=100032798297485" \h]). We need to more fully understand why sentiment aligns with engagement in some cases and not others. Marketing has also done work to deeply investigate the correlation between engagement and sentiment ([HYPERLINK "https://fb.workplace.com/notes/405091380607517/" \h]; [HYPERLINK "https://docs.google.com/presentation/d/11sNCj9hvToy14HqvqBKQL-c6MRrrm1_Hp9KWyJXbaVo/edit" \l "slide=id.gcb45b6e01d_0_3214" \h]).

Create balanced measures. Behavioral measures are often favored because they are easy to understand -- e.g., "clicks have gone up." But there is often ambiguity -- not all clicks are equal. For instance, we know that 14% of clicks on ads are accidental ([HYPERLINK "https://fb.workplace.com/notes/155895289724653" \h]). Incorporating survey data can help improve the construct validity of our measures (i.e., the degree to which what we are measuring matches what we really care about). On the flip side, one of the core weaknesses of surveys is limited volume, so a sentiment measure (or sentiment proxy measure) that incorporates behavioral data will typically scale much better and thus lead to a more operational metric.

Sanity check our signals. With latent modeling approaches such as Item Response Theory ([HYPERLINK "https://fb.workplace.com/notes/190033862604470/" \h]) we can monitor the quality of data from survey questions alongside log behavior. For instance, we can understand

how cohesively these signals are providing a measure of sentiment and how well we're capturing the full spectrum of sentiment. Because changes to our product impact not only behavior and attitudes but the reasons for behavior and attitudes, we naturally expect shifts over time in the quality and value of these signals. And generally speaking, measures based on multiple signals (if appropriately combined) will give us a higher-quality estimate of the underlying concept.

Examples of Better Metrics

Engagement-based Interest Relevance Score (EIRS) predicts survey response based on the impression-level user engagement signals (normalized by presentation / user biases) and thereby 1) scales the survey's sample size and 2) bridges the gap between survey and user engagement.

Quality sessions metric uses links between what people do/see in a session and how they felt about the session (i.e., session-level sentiment) which can improve our understanding of session quality and the types of sessions we should optimize for ([HYPERLINK "https://fb.workplace.com/groups/375376650972558/permalink/396699315506958/" \h]). This helps us understand the drivers of high vs. low quality sessions by leveraging a sessions-level survey which captures user feedback about the session that a user just completed.

Next Steps

Successfully incorporating sentiment into our product development process will involve contributions from a variety of stakeholders. This will mean:

- Pushing on teams to invest more in understanding the association between implicit (e.g., sessions, engagement metrics) and explicit signals (e.g., survey-based measures) across product areas.
- Creating an FB-app-level taskforce to develop better goaling metrics and practices based on the recommendations outlined above.
- Developing a framework to make demonstrated user value (not just increase in engagement or visitation) an explicit criterion for all FB app product launches.

HIGHLY CONFIDENTIAL (COMPETITOR)