Internal Instagram and Facebook Documents

These documents were created during my time working at Instagram as a consultant from 2019 to 2021. I had earlier worked at Facebook as a Director of Engineering from 2009 to 2015. I was the senior engineering and product leader for efforts to keep users safe and supported. I returned to Instagram in 2019 to work exclusively on user experience and well-being.

The documents here in this Google drive are in chronological order. I've also attached two folders. One contains the earlier work Facebook published to the public on the subject matter of teens. It includes findings from research and product development from 2011 to 2015. The second folder contains, for your background, a document from the Facebook Files that discusses a similar pre-existing survey that is mentioned in some of the other documents.

The following notes serve as a guide or crib sheet to help you make sense of the documents.

Document labeled '1 - "Bad Experiences" Measurement - Plan for a Plan - Nov 19 2020 WB review.pdf'

The first document is a set of slides prepared by me and members of the Instagram Well-being team in November 2020, after I had been at Instagram for about a year. My team and I had come up with a new framework with which to measure what we started to call "bad experiences" for users. Our hope was to let users tell us about those experiences and then develop tools to support them.

Along with product managers, researchers, and others in the company, I prepared this slide set for the Well-Being leadership team at Instagram. We were proposing to formally set goals for the reduction of bad experiences, as defined by the users themselves, as well as measuring the effectiveness of the support tools we planned to introduce. Did users think the tools were helping them deal with the issues? Since data drives almost all work at Instagram and Facebook, we were arguing for the creation of data sets that could be measured and reported.

Note especially the slide called "Examples of bad experiences often in policy gaps," which helps explain why so many bad experiences did not violate existing company policies.

Internal Instagram and Facebook Documents

(Some slides appear twice so the full slide can be visible as well as comments various company employees made on top of them.)

Document labeled '2 - AI + FAI Workshop_ bad experiences.pdf'

By early 2021 the idea of working on "bad experiences" had begun spreading to other parts of the company. This is a presentation prepared by Facebook Research, pertaining to its own work on related issues. I had nothing to do with preparing this document.

Note in particular, though, the slide labeled "bad experiences are common and frequent," which indicates that two out of five users on Facebook had an experience in any given week that they considered "bad." Another important slide is the one labeled "Users in this study rated borderline content **as harmful** as violating content." Note that this data set was compiled in 2018. This shows that the company had known for a long time that what users were experiencing as harmful did not match the company's definition of content that violated policy.

I don't know if this work at Facebook continued. However, it is my understanding that approximately half of the Facebook Research Team that did this kind of work was eliminated in 2023 as part of the "year of efficiency."

Notes on the following three documents, all with the term BEEF in their title

While our recommendations for regular measurement, reporting, and goal setting were not all adopted following our initial presentation at Instagram, we did get some traction with the "bad experiences" model. We were given more resources and created BEEF—the Bad Experiences and Encounters Framework.

The first two documents illustrate the thought process behind BEEF, which was intended to help employees better understand the bad experiences people were having inside Instagram. In the surveys conducted as part of this work, we began asking people about unwanted sexual advances.

These documents explain the research plan—how many people got the surveys, the methodologies, etc. They demonstrate that the company was running a very thorough and methodical survey program around this work.

The thumbnail slide presentation is the only form I have that in—this was the full presentation of the results of the BEEF work. Apologies that it is partly illegible.

Document labeled '7b - BEEF by Age (attachment to Gap in understanding e-mail).png'

This table is the one piece of detail I have from the full BEEF presentation. It comes from slide 19, entitled "Issues by all age groups." I attached this chart in the email I sent to Mark Zuckerberg, which is included in this drive.

Note: the data for all users is an accurate representation of what people reported experiencing. The data in the columns listed by age groups, however, is unadjusted. That's important to recognize. The survey was conducted in two parts, and only those users who indicated they had had some sort of bad experience in the past seven days were asked to continue, including by indicating their age. So these age-based tables are the percentage of all users that age who had one or more bad experiences, who had this exact bad experience. In other words, the percentages in the age-based tables do not represent the total number of people that age who had that experience in the past week.

For the correct adjusted numbers for some extremely important categories affecting teens, see the final document—my email to Adam Mosseri. The numbers in that email *do* reflect the total percentage of users that age who had that bad experience.

Email to Mark Zuckerberg and M-team '7a - Gap in our understanding of harm and bad experiences.pdf'

I sent this to express my concern about what we had been learning as well as to make suggestions for steps the company could take to respond on behalf of users. Before I sent it, I vetted it carefully with multiple people inside the company, including some who were quite senior. In other words, I followed the normal procedure for flagging issues to executive leadership in a technology company, consistent with communications I participated in during my earlier stint at Facebook, 2009-2015.

Regarding the statistics in the email to Zuckerberg for bullying, experienced negative comparisons, and sexual harassment: these represent the unadjusted figures (see note above), which had been given to me at the time by company researchers. The adjusted figures for these categories, which I obtained later, are all in the next document, the email I sent to Adam Mosseri.

Email to Adam Mosseri ‘8 - WSJ published Mosseri Pre-Read ffpreread110223.pdf’

I later had a meeting with Adam to discuss these findings and recommendations. This is the email I sent him in advance of our meeting, to “pre-brief” him. Note that this email was published by the Wall Street Journal on November 2. The statistics in this email are all the adjusted numbers.

Folder ‘2011-2015 Published work by Facebook on Teen Bullying’

This folder has presentations that were made available to the public by Facebook between 2011 and 2015. The data was the result of close collaboration with Marc Brackett and Robin Stern from the Yale Center for Emotional Intelligence. Of special note is slide 3, titled ‘Why are we here’ in ‘CRD2_Yale Team_Compassion Day 2 Presentation_FINAL copy.pptx’.

The document ‘CRD3_Yale_Team_Compassion Research Day 3__1_23_2013_FINAL.pptx’ shows the results of a product development process for helping teens with the issues they experience on social media.

Folder ‘From Facebook Files’

For background, TRIPS was a pre-existing survey in which users told Facebook about their bad experiences. The historical numbers are within range of the BEEF findings.

WELL-BEING LEADS REVIEW · NOV 19, 2020

“Bad Experiences” Measurement

Plan for a 2021 plan

Meeting Objectives

Goals

1. Share ambition and key examples of problems to solve
2. Share high-level approach
3. Feedback on next steps

Non-Goals

1. Follow up on Oct 15 review of TRIPS proxy success criteria
2. Roadmapping-level detail
3. Product strategy for bad experiences
4. “Positive Experiences” measurement (ex. promoting positive well-being)

Our mission is to create the safest and most supportive global community.

We envision an Instagram where, everyone, especially teens, creators and underserved communities, feel safe and supported to express themselves and to push culture forward.

Supportive - Helping people with bad experiences, by reducing and resolving them. Not generally focused on enforcement.

Policy Violating - Problems defined in the company's Community Standards content policies

Borderline - Non-violating content policy intended to address part of policy narrowness

Bad Experiences - Problems defined by people's perception, as implemented in survey taxonomies like TRIPS and Hard Life Moments

Legitimacy - People and external stakeholders believe that our integrity work is effective at reducing harm and that our enforcement is defensible and fair. Both 1st and 3rd party perceptions

Reach - Measures the size of bad experiences (number of people affected by them or proxy)

Support Effectiveness - Measures how effectively we're resolving a bad experience. For people experiencing the problem (1st party), intersecting with 1st party part of legitimacy

Our company integrity mission is 'protect the community and its voice'. Our policies have to balance both protecting the community and its voice.

People feel they’re having a bad experience or they don’t. Their perception isn’t constrained by policy balance.

Understanding “bad experiences” enables us to see the fuller people-first picture of how people experience safety and support concerns, agnostic of policy.

100%

of Community Standards policy focuses on enforcement (delete, MAD, or resources)

~100x

Incremental reach for B&H witness in TRIPS compared with Community Standards B&H policy violating impressions

[redacted] 11/19/2020

I agree with the spirit of this fact - the perceived problem is likely much bigger than the CS-violating problem - but I think it may be a bit misleading to compare % of impressions that are bad vs. % of people who have a single bad experience (in a week's worth of impressions)

100%

of Community Standards policy focuses on enforcement (delete, MAD, or resources)

~100x

Incremental reach for B&H witness in TRIPS compared with Community Standards B&H policy violating impressions

[redacted] 11/19/2020

I think that it's generally true if we compare content prevalence impressions to TRIPS DAP.

Our policies are narrow (CS violations are very narrow, given consequences of overenforcement when we get it wrong, borderline somewhat less narrow).

[redacted] 11/19/2020

Is this true across problems or predominantly on bullying and hate speech where context matters a lot?

Types of gaps between policy and bad experiences

Policy can be effective, but ...

1. incident doesn't meet the policy bar though clearly bad
2. incident doesn’t fit current policy definitions
3. incident is hard to detect (by human or automated review)

Policy isn’t the right tool when ...

4. “bad” is impossible to determine with any consistency
5. enforcement actions (hard or soft) cannot address the problem

Examples of bad experiences often in policy gaps

Mass harassment

Very intense, low reach. Disproportionately affects creators. Portion isn't CS policy violating but still felt as intense.

Non-credible or non-violent threats

Doesn't meet CS violating policy bar, but felt intensely with moderate reach.

Thinspiration

Easy to find content adjacent to eating disorder part of SSI and to broader social comparison and body image issue-triggers.

Billie Eilish Admits That Reading Instagram Comments Was “Ruining” Her Life

"The cooler the things you get to do are, the more people hate you."

Chrissy Teigen • Rita Wilson • Wilfred Zaha • Dua Lipa • Milana Vayntrub

HOME > CULTURE

The actress who plays AT&T's Lily is facing waves of online sexual harassment, including manipulated images and objectifying memes

- Only some is policy violating or borderline
- The frequency and volume makes a very intense bad experience
- Low reach

When I received rape threats in my DMs, Instagram offered me no help at all

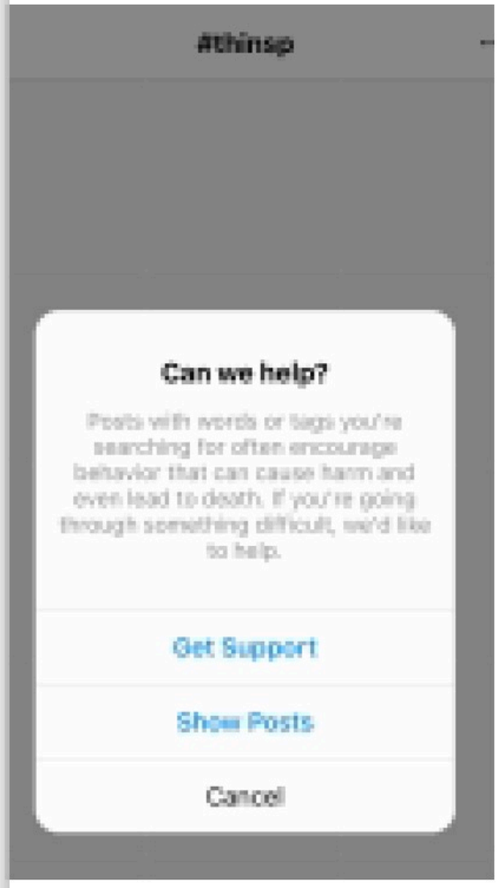
Comment

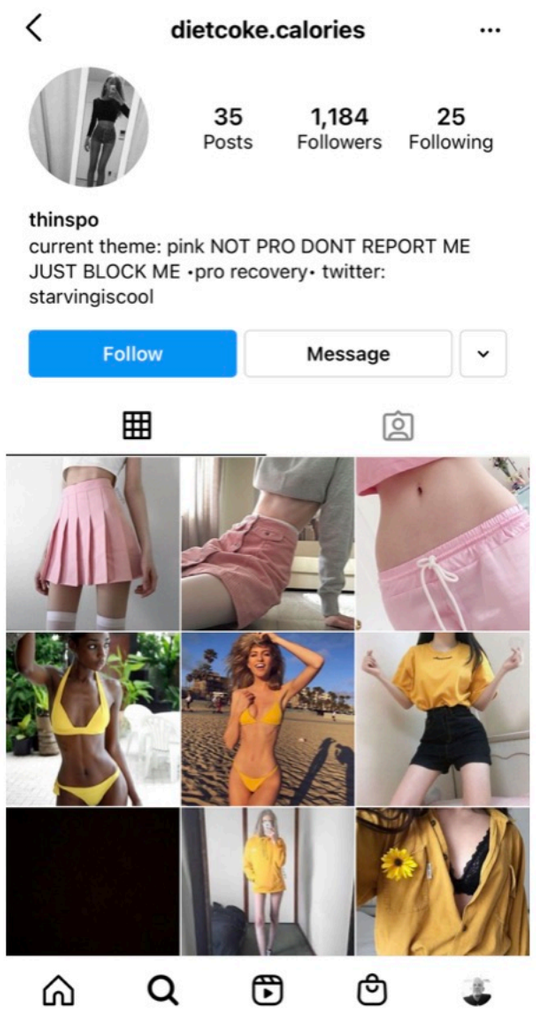
FIRST PERSON

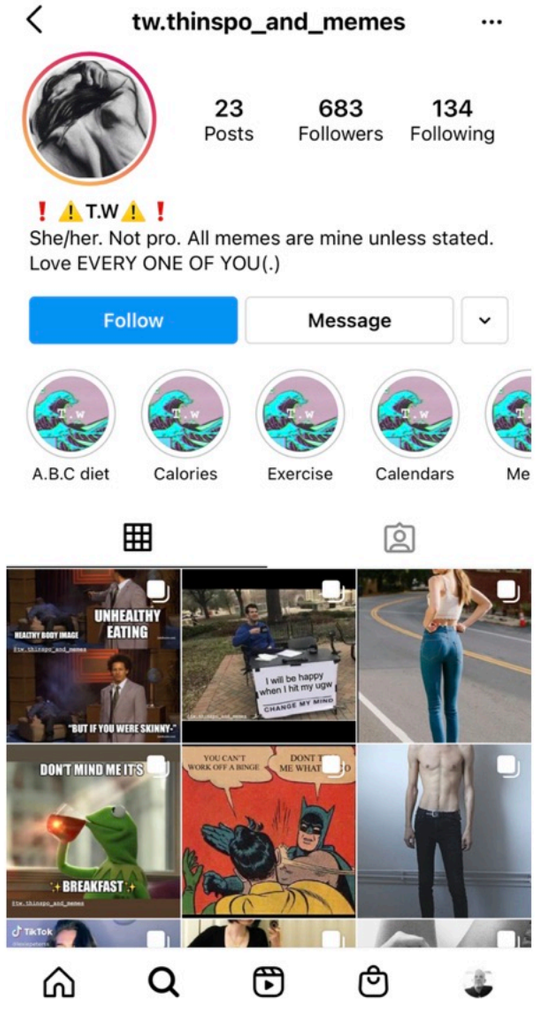
[redacted] Friday 24 Jul 2020 3:52 pm

I've been on the receiving end of an unsolicited dick pic or two in my time. I laughed them off at first.

Then the same man started sending explicit descriptions of violent sexual fantasies he had about me. This included rape, choking, putting me in crutches and forcing my mother and father to watch.







"We're not teaching [others] to say, hey are you okay? We're teaching them to click a report button or to call the cops and try and trace the phone or trace the location based on a Tweet. I'm kind of afraid of the resources because of that."

- One type of triggering content related to body image issues
- Generally not policy violating, easy to seek, often leaks into general fitness discovery experiences
- Posters often suggest solution is self-support, not enforcement: “Block, don’t Report”, “Trigger Warning”
- Can trigger harassment for poster

2021 Proposal

What does success look like by end of 2021?

Operationalize “bad experiences” measurement for experiment launch decisions and goaling to empower teams to work on problems in policy gaps.

Measurement Tracks

Reach

Measure the number of people affected by bad experiences

Support Effectiveness

Measure how effectively we're resolving a bad experience

Not covered here: 3rd party ("reputational") reach or intensity separate from reach

Other definitions:
Intensity = felt experience by users

Severity = combines risk to business and risk of real world harm for people with felt intensity research

Measurement Tracks

[redacted] 11/19/2020

Why not intensity? Seems like it's as important as reach if we think working on something like mass harassment should be incentivised

[redacted] 11/19/2020

There's a pretty complete answer to this common question on the comment thread on this slide. cc'd you there. tldr:
- intensity as a pure metric separate from reach hasn't yet shown any applications.
- severity-bucketed reach and intensity-weighted reach are both variants inside the Reach track

Reach

Measure the number of peopl

Support Effectiveness

Measure how effectively we're resolving a bad experience

Not covered here: 3rd party ("reputational") reach or intensity separate from reach

Other definitions:
Intensity = felt experience by users

Severity = combines risk to business and risk of real world harm for people with felt intensity research

Reach measurement in 2020

Not covered here: Hard Life Moments for mental well-being concerns

Violating prevalence impressions

Labeled and classifier estimated metric for narrow policy violating definition. Available for a subset of problems based on CI + IGWB prioritization and operational cost.

FRX

Only negative user feedback with problem detail, with problems as defined by Community Standards policy. Available as aggregate for all reporting problems (dominated by spam) and by violation types and problem tags.

TRIPS

Survey for 13 integrity problems launched on IG in Dec'19. Also has Civic variant.

[redacted] 11/19/2020

since we're not talking about hard life moments and we're primarily looking at integrity+ (borderline experiences) - is it right to say this is primarily excluding measurement in the non-integrity well-being space for 2021?

[redacted] 11/19/2020

the logging right now is terrible, but long-term there's a lot of potential signals here -- even people viewing settings pages (people generally don't look at privacy settings just because they are curious) or even stuff like unfollowing/'not interested' accounts/content that lots of other people unfollow, or even people using the search box to try to find out

[redacted] 11/19/2020

the recent switch is a good idea. We only look at whether it is currently private as a feature

[redacted] 11/19/2020

No, not intended to exclude other 2021 paths. The intent of this deck is as a conversation starter. In 2020, HLM wasn't implemented as an ongoing metric like the other items on these slides.

Assuming that if we can get momentum here in the integrity space, it helps us leap forward with bad experiences in other areas (ie. mental well-being).

[redacted]r 11/19/2020

Blocks are in pTRIPS and USI. We (Foundation data and Bully eng) have been working to improve the attribution of account level actions so they're more predictive of problems.

I don't think we've considered switching from Public to Private recently.

[redacted] 11/19/2020

It would also make sense to look into whether we can get clear signal from self-remediation efforts -- what does it tell us when someone switches their feed from public to private, or blocks someone?

H2 learnings about Reach measurement

Violating prevalence impressions

While responsive to some changes, doesn't capture most personalized demotion impact in comments (similar on FB XI).

FRX

Per-product / per-surface cuts matter for seeing per-experiment impact. "FRX feedback" more inclusive than "FRX submit reports".

TRIPS

Responsive to IG Bullying H1 changes in comments slice (-5%).

Responsive to world events, including at sub-problem detail: Since George Floyd protests in June, in the US "threats" portion of B&H and "political beliefs" for HS are elevated (ss)

[redacted] 11/19/2020

Yes. Per the last WB review on pTRIPS, that's how we agreed to come back when we're focused on details to do a selection discussion. That level of detail was outside the

H2 learnings about Reach measurement

[redacted] 11/19/2020

One measure we could look at is sensitivity, as measured by the portion of experiments that are known negative/positive and are captured with metrics movement. Curious how well TRIPs scores on that dimension

Violating prevalence impressions

While responsive to some changes, doesn't capture most personalized demotion impact in comments (similar on FB XI).

FRX

Per-product / per-surface cuts matter for seeing per-experiment impact. "FRX feedback" more inclusive than "FRX submit reports".

TRIPS

Responsive to IG Bullying H1 changes in comments slice (-5%).

Responsive to world events, including at sub-problem detail: Since George Floyd protests in June, in the US "threats" portion of B&H and "political beliefs" for HS are elevated (ss)

H2 learnings about Reach measurement (continued)

USI

Most weight goes to FRX when predicting policy violations. Operationally challenging to add new signals in current central governance model.

pTRIPS

Super early results show some stat sig effects for 2 IG Bullying experiments (Remove Preview Filter Toggle, Personalized Demotions). TBD if surveys validate prediction.

Product cuts for both pTRIPS proxy and TRIPS surveys probably quite important.

FB XI pursuing similar approach in close collaboration with IGWB and CDS working group. Different in details.

[redacted] 11/19/2020

Is there any way to look at a white box description of both of these? I think it's pretty crucial for predictive metrics to be intuitive in order to gain intuition of the limitations

H2 learnings about Reach measurement (continued)

[redacted] 11/19/2020

Yes, in the last execution review that was narrowly focused on these we agreed to do a detailed follow up execution review for selection to land the H2 work. We'll need to do a deep dive there.

USI

Most weight goes to FRX when predicting policy violations. Operationally challenging to add new signals in current central governance model.

pTRIPS

Super early results show some stat sig effects for 2 IG Bullying experiments (Remove Preview Filter Toggle, Personalized Demotions). TBD if surveys validate prediction.

Product cuts for both pTRIPS proxy and TRIPS surveys probably quite important.

FB XI pursuing similar approach in close collaboration with IGWB and CDS working group. Different in details.

Support Effectiveness measurement in 2020

Adoption

See how many people use our new support experiences (ex. Support Requests, Restrict, Bulk Comment Management)

Supportiveness

IX-developed survey: “How supportive was Instagram during this experience?” triggered after completing a specific flow

LEGIT

Covers much more scope than effectiveness for people experiencing problems (1st party). Responsive to actor transparency changes on IG.

How to Get There

Major Next Steps

1: Finish H2 execution

TRIPS holdout read for IG bullying, pTRIPS validation, USI development. We'll have new data before Dec 17 roadmaps.

2: Coordinate with CI IMI, OC, IX

At least CI IMI is actively proposing roadmap priorities that would help advance work in this area.

3: Prioritize H1 steps to unblock IG teams

Which metrics would unlock the most impact? What needs Foundation and IMI help vs. being driven by teams?

4: Decide IG Bullying goal / launch metrics

H1'21 is a key inflection point for IG Bullying in its 5th half of working in this direction. WB is taking Bullying P0 goal.

[redacted] 11/19/2020

I believe the bullying p0 would be oriented towards creators. Does that changes things or same measure?

Major Next Steps

[redacted] 11/19/2020

I assume expanding the bullying focus to creators is more of a shift in focus than a complete change for measurement.

1: Finish H2 execution

TRIPS holdout read for IG bullying, pTRIPS validation, USI development. We'll have new data before Dec 17 roadmaps.

2: Coordinate with CI IMI, OC, IX

At least CI IMI is actively proposing roadmap priorities that would help advance work in this area.

3: Prioritize H1 steps to unblock IG teams

Which metrics would unlock the most impact? What needs Foundation and IMI help vs. being driven by teams?

4: Decide IG Bullying goal / launch metrics

H1'21 is a key inflection point for IG Bullying in its 5th half of working in this direction. WB is taking Bullying P0 goal.

Discussion

Feedback on ambition and examples?

Is this ambitious enough? Too ambitious? Resonate as important and people-first?

Feedback on approach?

Is the framing intuitive? Leverage existing concepts and terminology for “extreme clarity”?

Feedback on next steps?

At a high-level, are we missing major steps for getting ready for H1 roadmaps?

[redacted] 11/19/2020

My main feedback is I'm worried that there's a metrics explosion in this space while it's unclear which ones are actually effective. I'd love for well-being to mostly follow others on metrics that have already proven effective

Feedback on ambition and examples?

Is this ambitious enough? Too ambitious? Resonate as important and people-first?

Discussion

[redacted] 11/19/2020

not all these measures are at the same stage of evaluation and understanding. I think I'm not clear on how we're thinking teams will use all these metrics? Is the idea to use a basket of metrics for problems?

s and

[redacted] 11/19/2020

Yes, this feedback seems to create inputs for step #3 on slide 20 (prioritize based on what would create impact for our teams).

At a high-level, are we missing major steps for getting ready for H1 roadmaps?

CONTEXT · **EXISTING METRICS**

Track	Metric	Question metric answers
Reach	Violating prevalence	How many impressions are seen for content violating CS policy?
	FRX	How many people report content and accounts?
	TRIPS	How many people have seen or experienced the problem in last 7d? 5-point intensity?
	USI	What's the aggregate of reporting, blocking, unfollowing, and other actions? (with aggregate weighted for correlation with CS violations)
	pTRIPS	Of the aggregate of user actions and other signals, what's our prediction for TRIPS?
Support Effectiveness	Adoption	How many people adopt the supportive product solution? Does it cannibalize?
	Adoption within problem segment	How many people with a specific problem adopt the supportive product solution?
	Supportiveness	How supportive was Instagram during this experience?
	LEGIT	Do people believe that our integrity work is effective at reducing harm and that our enforcement is defensible and fair?

For the purposes of an integrity guardrail, Peter has posted options for a single metric: https://fb.workplace.com/groups/1540209922820872/permalink/1679116712263525/

Guardrail and primary positive impact metrics will have different needs

Track	Metric	Shows progress in policy gaps?	Already proven in Nov'20?	More progress landing soon?	Works for low-reach?	Team(s) with most progress
Reach	Violating prevalence	No	**Yes**	Surface improvements landing	No (except UB Prevalence)	CI OC + IG Foundation
	FRX	**Yes**	**Yes**, for some use cases	Improving logging quality	?	CI IX, CI OC, IGWB Foundation, IG Bullying, US2020
	TRIPS	**Yes**	**Yes**, for some use cases	Major new results landing in Dec	No	FB XI, IG Bullying
	USI	No	No, not operationalized	H2 IG project in flight	?	CI IX, CI OC
	pTRIPS	**Yes**	No, not operationalized	Major new results landing in Dec	?	IG Bullying, FB XI

For the purposes of an integrity guardrail, Peter has posted options for a single metric: https://fb.workplace.com/groups/1540209922820872/permalink/1679116712263525/

Guardrail and primary positive impact metrics will have different needs

CONTEXT · **ASSESSING EXISTING METRICS**

Track	Metric	Shows progress in policy gaps?	Already proven in Nov'20?	More progress landing soon?	Works for low-reach?	Team(s) with most progress
Reach	Violating prevalence	No	**Yes**	Surface improvements landing	No (except UB Prevalence)	CI OC + IG Foundation
	FRX	**Yes**	**Yes**, for some use cases	Improving logging quality	?	CI IX, CI OC, IGWB Foundation, IG Bullying, US2020
		Yes	**Yes**, for some use cases	Major new results landing in Dec	No	FB XI, IG Bullying
		No	No, not operationalized	H2 IG project in flight	?	CI IX, CI OC
		Yes	No, not operationalized	Major new results landing in Dec	?	IG Bullying, FB XI

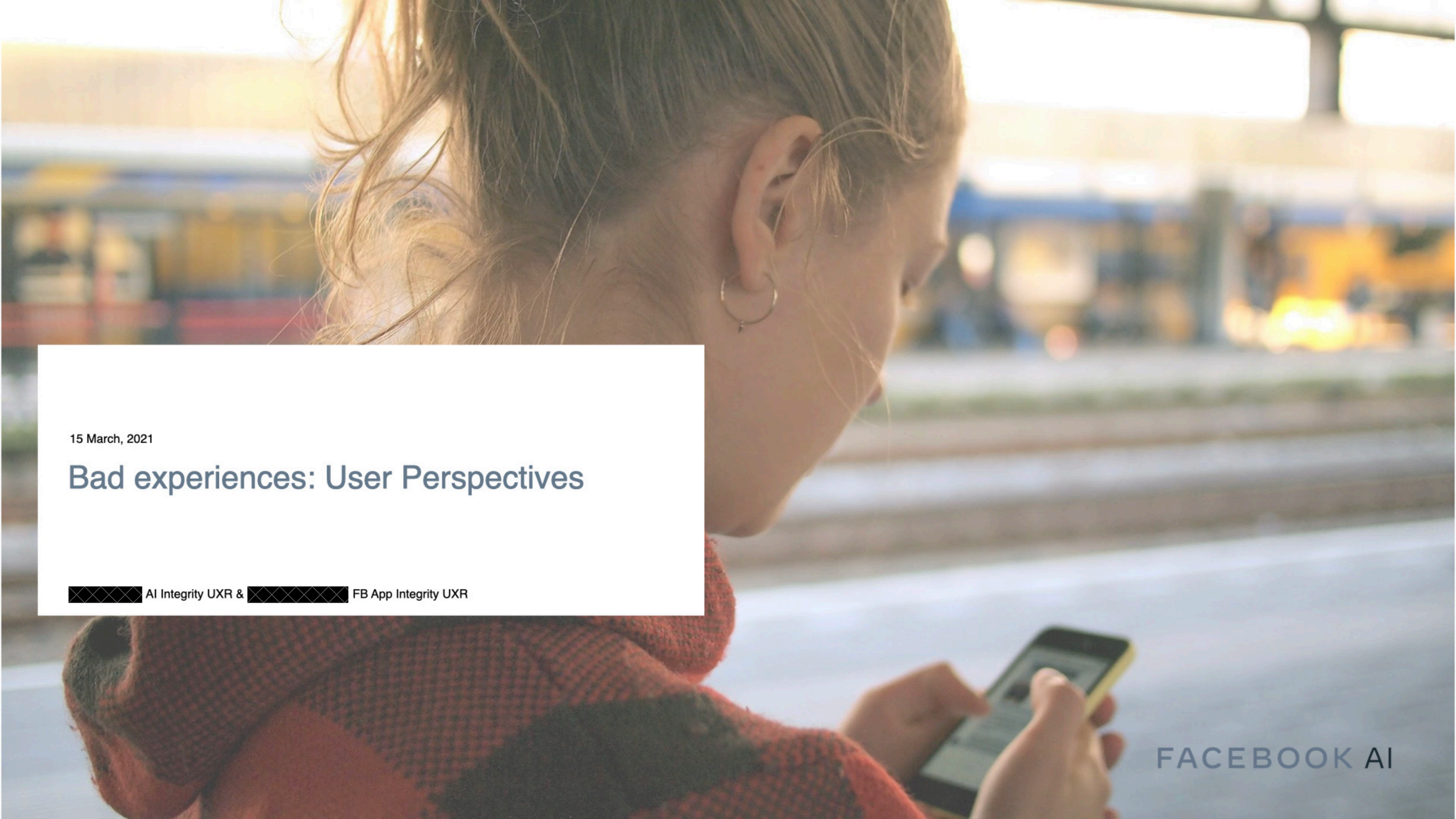
[redacted] 11/19/2020

Used as primary proxy metric for if problems were healthy during COVID surge and US2020.

For the purposes of an integrity guardrail, Peter has posted options for a single metric: https://fb.workplace.com/groups/1540209922820872/permalink/1679116712263525/

Guardrail and primary positive impact metrics will have different needs

Track	Metric	Shows progress in policy gaps?	Already proven in Nov'20?	More progress landing soon?	Works for low-reach?	Team(s) with most progress
Support Effectiveness	Adoption	?	**Yes**		**Yes**	IG Bullying
	Adoption within problem segment	**Yes**	No, initially proposed for Limited Profile		**Yes**	IG Bullying
	Supportiveness	?	Yes, but for different use case		?	CI IX
	LEGIT	?	1 early proof point for different use case		?	IGWB Foundation, CI Legitimacy



15 March, 2021

Bad experiences: User Perspectives

[redacted] AI Integrity UXR & [redacted] FB App Integrity UXR

FACEBOOK AI

TABLE OF CONTENTS

Goals for today

Develop a shared understanding of:

1. What kinds of content upsets people
2. What people think FB should be doing
3. What we've done to address user expectations
4. Motivation for personalized demotions
5. Relationship between prevalence reduction and legitimacy/user perception

Consider:

1. "Watch-outs" for implementation of systems that address bad content
2. Opportunities for AI Integrity & FB App Integrity to partner to improve legitimacy
3. Ways AI could inform user-facing solutions, and ways user-facing solutions could provide valuable signals for AI

TL; DR

	Finding	Implication
1	**Hate speech, divisive civic content, and graphic violence** are frequently and intensely experienced, and have been shown to have a negative effect on sentiment and legitimacy, **particularly with repeated exposures over time.**	• Prioritize Offensive Speech in near-term efforts to improve sentiment. • Examine ways we can identify and target high-violation ecosystems where people experience repeated exposures.
2	**Borderline content** can be seen as **equally or more harmful** than violating content and decreases sentiment and engagement. In most cases, users want Facebook to hide or remove it. 70.7% US users believe Facebook should be doing more to address harmful content	We should ensure we have an adequate understanding of which borderline content users find most offensive so that we can prioritize and refine interventions and actions.
3	Post content is not the only problem--**toxic and divisive comments** commonly appear on **benign posts.** Reshares, Links and Status Updates are more likely to be rated as a Bad Experience compared to photos and videos	Include comments as a target for classifications and actions
4	**Not every "bad experience" is unwanted.** Some respondents describe "needing to see" content they considered a bad experience, such as violence and racism.	As we design systems that classify and demote content, and make tradeoffs across user value, engagement, and legitimacy, we should be mindful that content that seems bad, upsetting or anger-inducing may be positively regarded by the viewer as meaningful and important.
5	Users want Facebook to act. They hold us responsible for negative experience, and most think Facebook should **automatically remove** severe integrity-related content and **hide** less severe content. They perceive **exposure** to integrity harms as **worse than false positive actions** on benign posts.	Continue investing in current efforts to reduce exposure to violating and borderline content.

TL; DR *cont.*

	Finding	Implication
6	**User experiences, preferences and perceptions vary.** Reaction to content varies by gender, ethnicity, culture and other factors; sentiment of Low-exposure users is more affected by integrity harms; those with low digital literacy are more likely to see violating content; some may even deliberately seek out harmful content.	• Personalization could be relevant for multiple interventions, not just soft demotion. E.g., selective display of warning screens and/or tombstones. • Offering controls can also be a way to capture relevant signals for integrity-related AI models
7	The **#1 legitimacy detractor** is the perceptions that FB is not doing enough to mitigate **bad experiences** on the platform**.** Legitimacy is also challenged by **lack of transparency & understanding of ranking & enforcement**. Content controls such as 'sensitive content preferences' serve a double role - not only do they reduce exposure, they help the user feel they understand what's under the covers.	Continue working to reduce bad experiences. .When evaluating the effectiveness of interventions, assess both impact on prevalence and impact on legitimacy.

Note: see appendix for information on observable attributes related to the likelihood a FB user has/has had a bad experience

TABLE OF CONTENTS

What is a bad content experience?

People have bad experiences on platform.

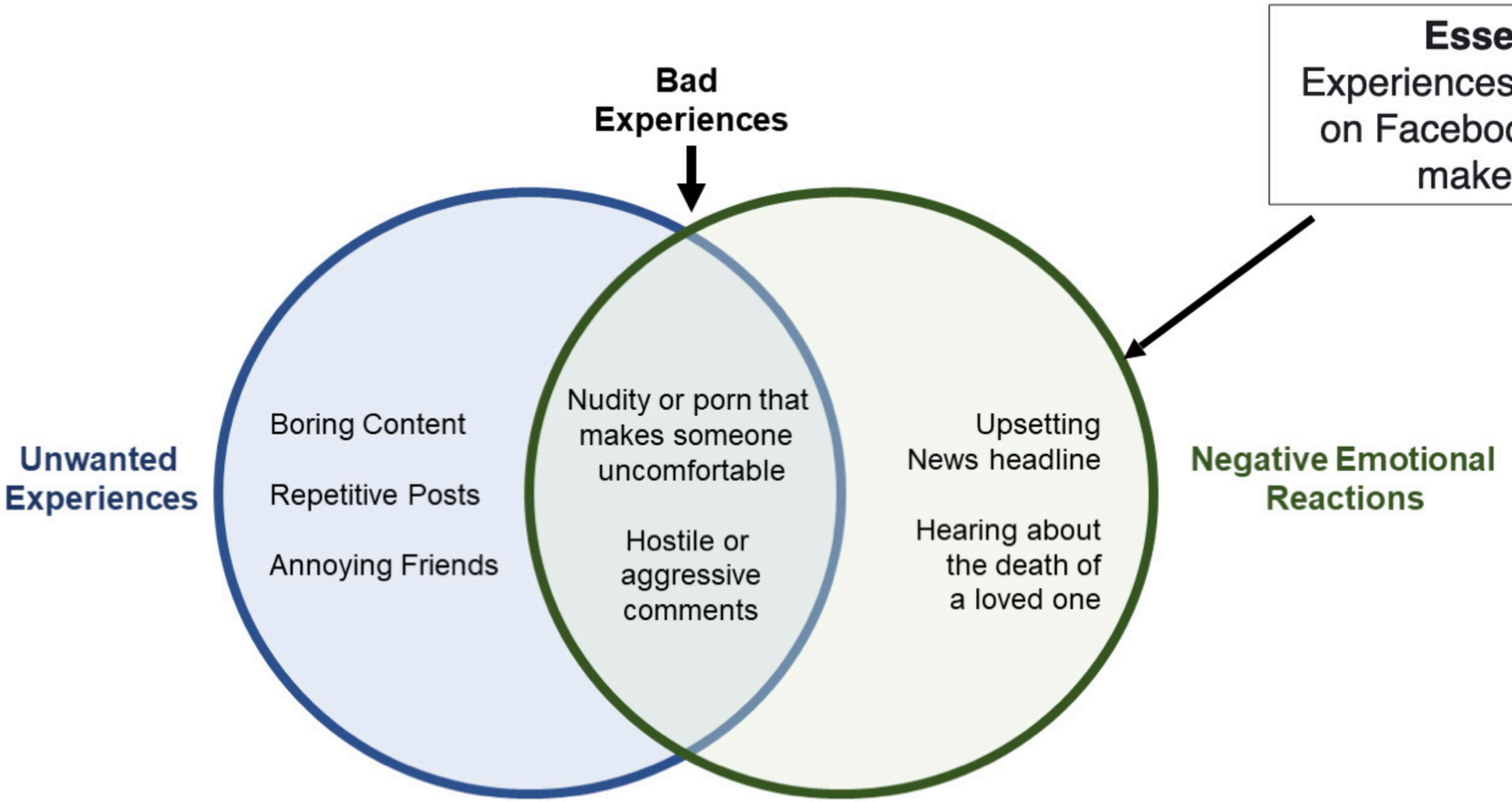
People don't know when to trust the content that they see on the platform.

People don't know what to do when they encounter bad content on the platform.

People don't trust that Facebook is actually reducing bad content on the platform.

What is a bad experience?

Bad Experiences are experiences that cause **unwanted, negative emotional reactions**.



Bad experiences are common and frequent

2 in 5 users

say they've had a bad experience while using Facebook **THIS WEEK** →

AXIS

1 minute after user opens app

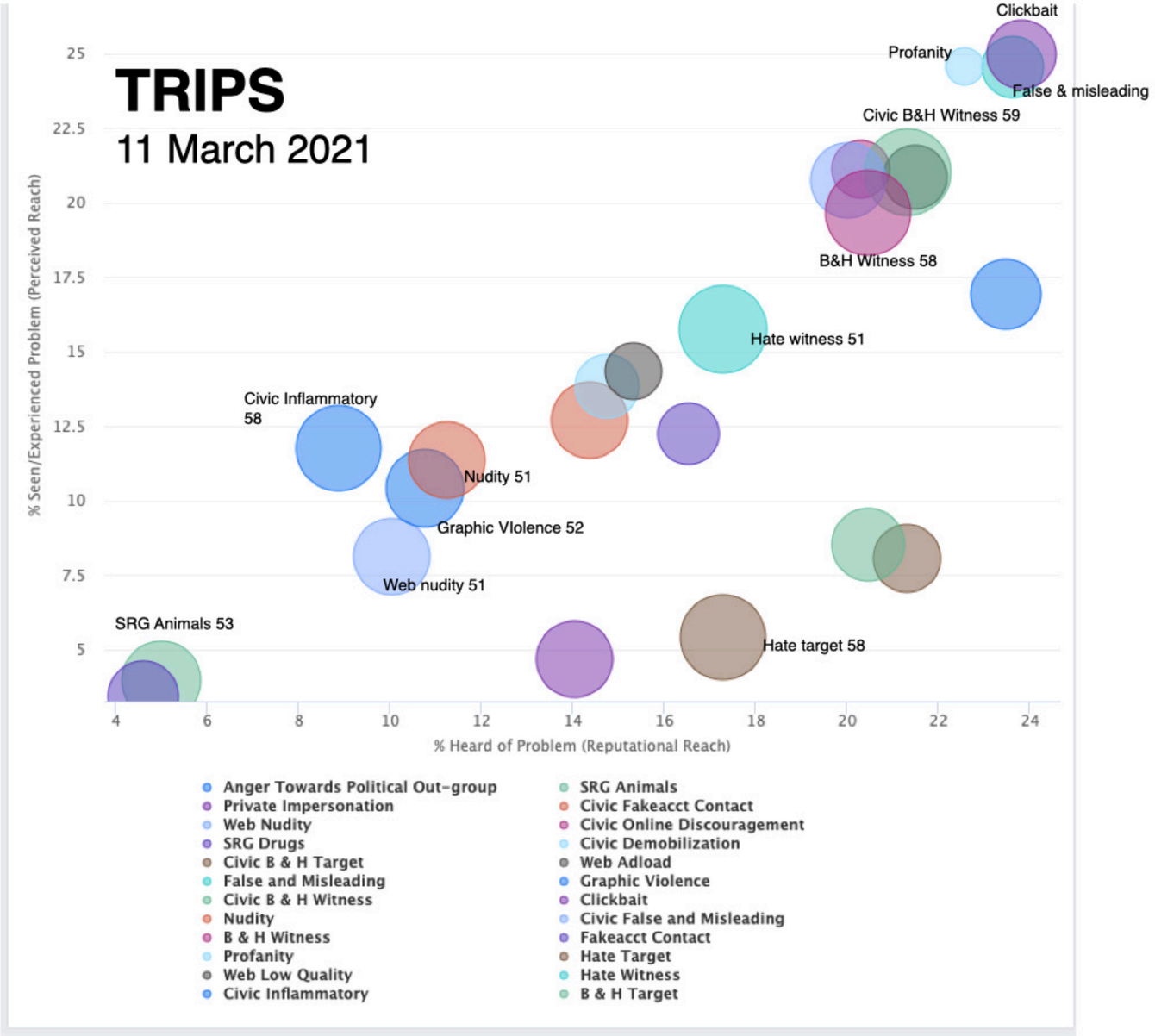
the moment **when bad experiences are likely to happen** to users →

US diary study participants

Clickbait, misinfo, and profanity are the most commonly perceived harms, but toxic, hateful content is more intense

Most <u>intense</u> experiences:

- Civic B&H Witness
- Civic inflammatory
- B&H witness
- Hate Witness
- SRG Animals
- Graphic violence
- *Nudity / Web Nudity*
  - *Note: exposure to nudity has been shown to increase FSS*



Bad experiences are diverse

Bad Experiences (as operationalized in TRIPS) can include→

- Hate speech & discrimination
- Profanity
- Graphic violence
- Private impersonation
- Nudity
- Obscene website
- Drug sales
- Bullying & harassment
- False/misleading
- Fake accounts
- Clickbait
- Low quality Link
- Animal sales
- Ads farms

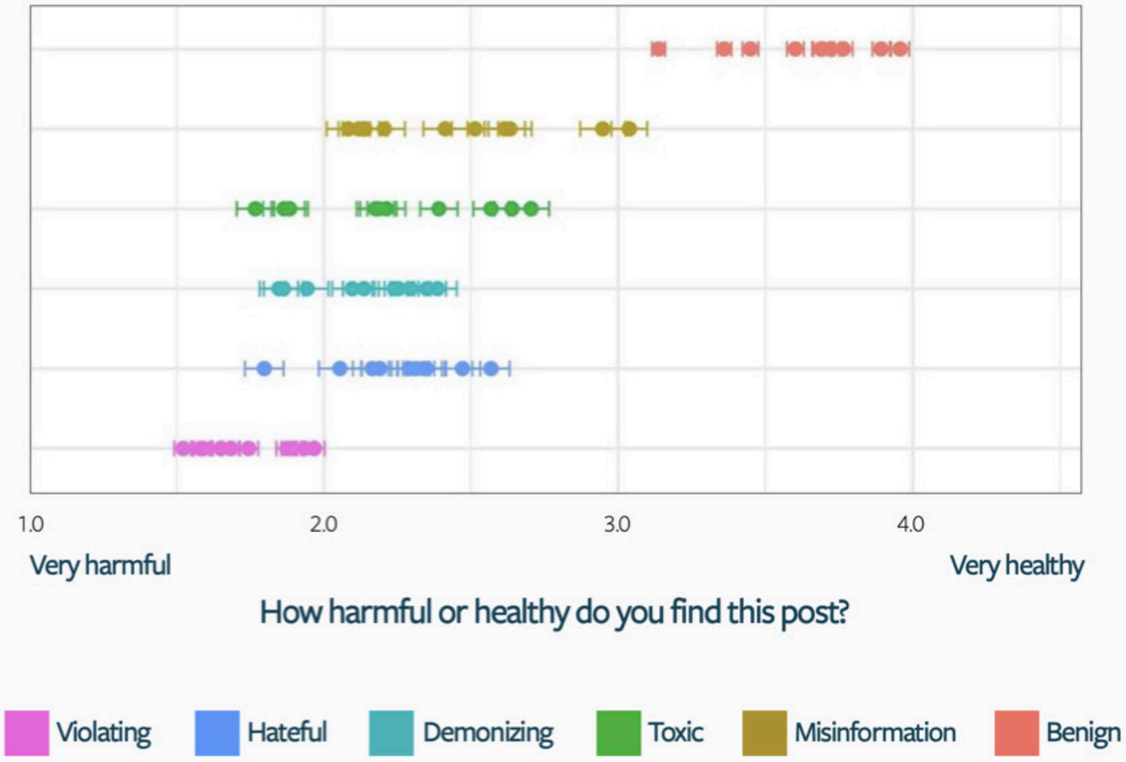
Post content is not the only problem--**toxic and divisive comments** commonly appear on **benign posts.**

And, content that leads to bad experiences does **not always violate Community Standards**. Borderline content is also a significant contributor to bad experiences.

- (What do people want us to do about Borderline Content? US survey,Jess Bodford, Eric Chen et al, Aug 2018)
- Users attitudes toward integrity enforcements Derek Powell, April 2020

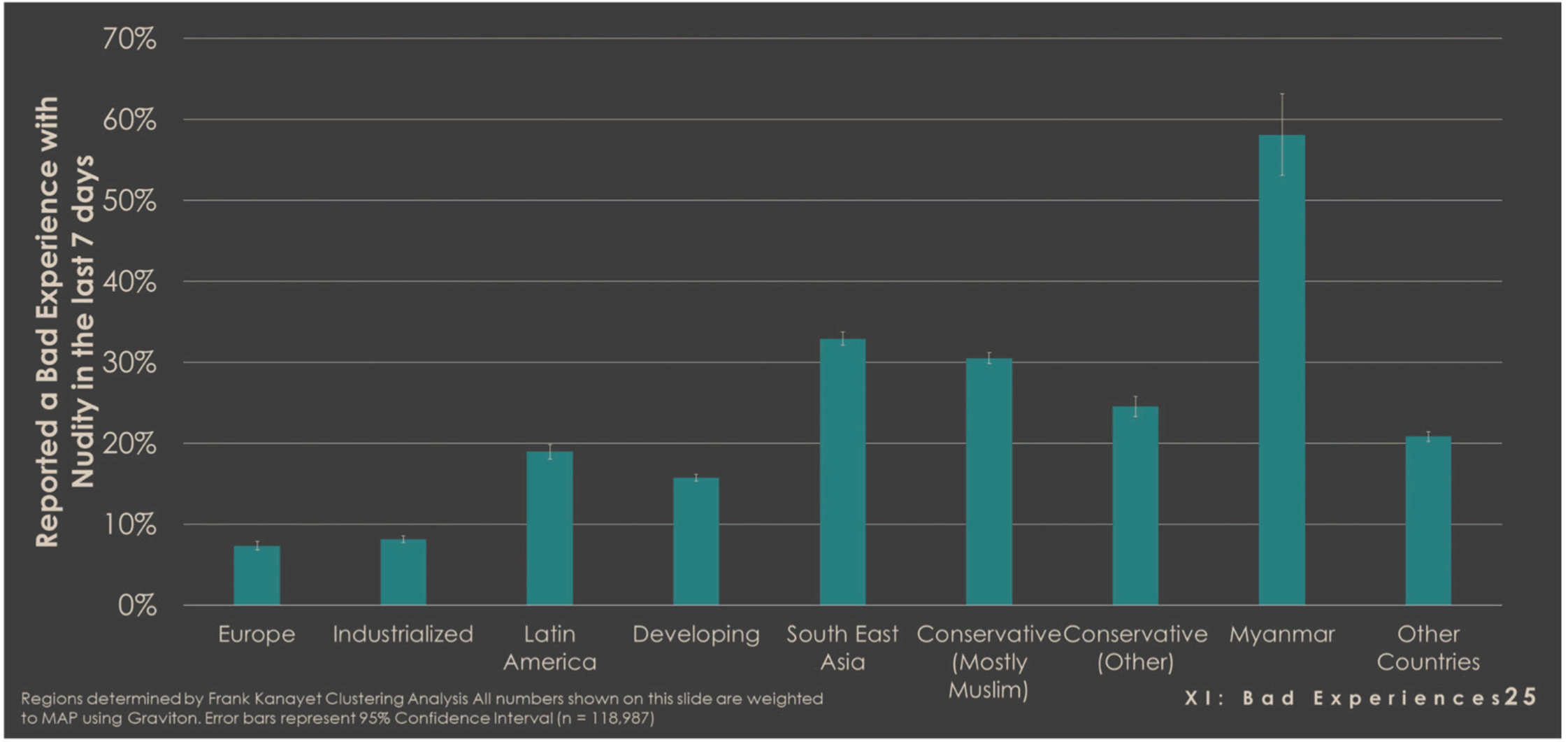
Users in this study rated borderline content* **as harmful** as violating content.

In fact, borderline content is **1.5x closer** on average to violating than benign content in perceived harmfulness.



*4 types of borderline content were tested in this study: misinformation, toxic, demonizing, and hateful content.
source: Borderline Content User Survey (Aug 2018)

We don’t all share the same set of values and beliefs. Bad experiences with harms like nudity can vary by market



Some respondents describe “needing to see” content they considered a bad experience, such as violence and racism.

- For example, they need to see content containing police brutality or military violence against civilians to better understand and contextualize the world around them.

Bad Experience

≠

"Do Not Want to See"

Why should we care about bad experiences?

Bad experiences & borderline content are related to decreased sentiment and engagement

- Increased exposure to borderline content (FUSS Red/Yellow), was **linked to a decrease in DAP** (Liu 2018).

- Exposure to borderline civic hate leads to **negative emotions, disengagement from Facebook,** and decreased engagement with offline civic and political actions (Travaglianti and Sacramone-Lutz 2019)

- Among US survey respondents, recent actual experiences with **hate speech or graphic violence** on Facebook was correlated with **lower average satisfaction** with News Feed (Powell 2020 #1, Powell 2020 #2). These users also reported that that the posts in their feed were **less likely to be worth their time**, and that they had **fewer meaningful interactions** on feed.

- **Sentiment** is most negatively associated with integrity harm exposure among **low-exposure users**
  - Possible implication: set user-level prevalence reduction goals rather than total VPV goals to prioritize these users?
  - Possible implication: Warning screens could be shown more liberally to low-exposure users?

- When shown examples of borderline content *(misinfo, toxic, hateful),* the majority of US survey respondents said: they did not want to see it, felt that Facebook should hide or remove it, and r**eported that they would spend less time** on Facebook after seeing it in their Feed (Bodford, et al 2018).

Keep in mind: some users like the bad stuff

Exposure to graphic violence has a u-shaped relationship with feed satisfaction

- For most users, graphic violence exposure is negatively associated with feed satisfaction.
- High exposure might be partially explained by users seeking out graphically violent content

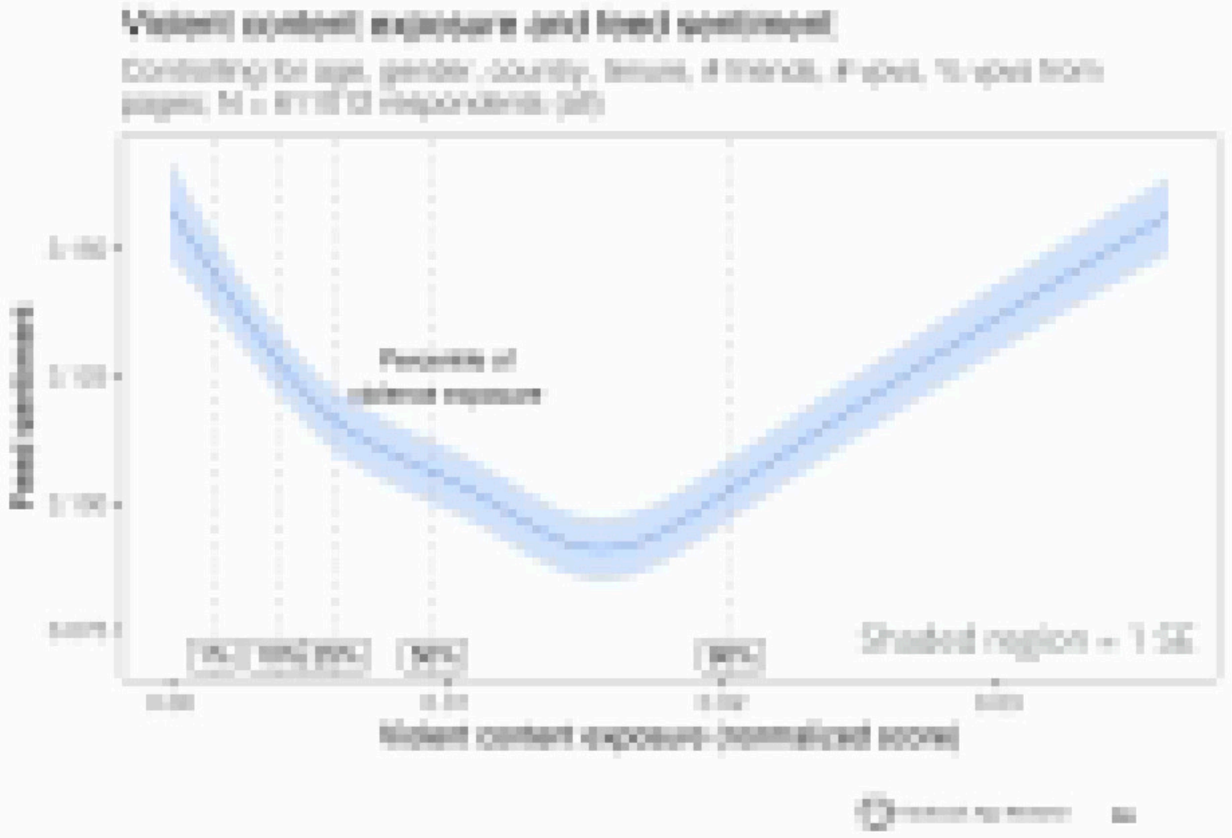
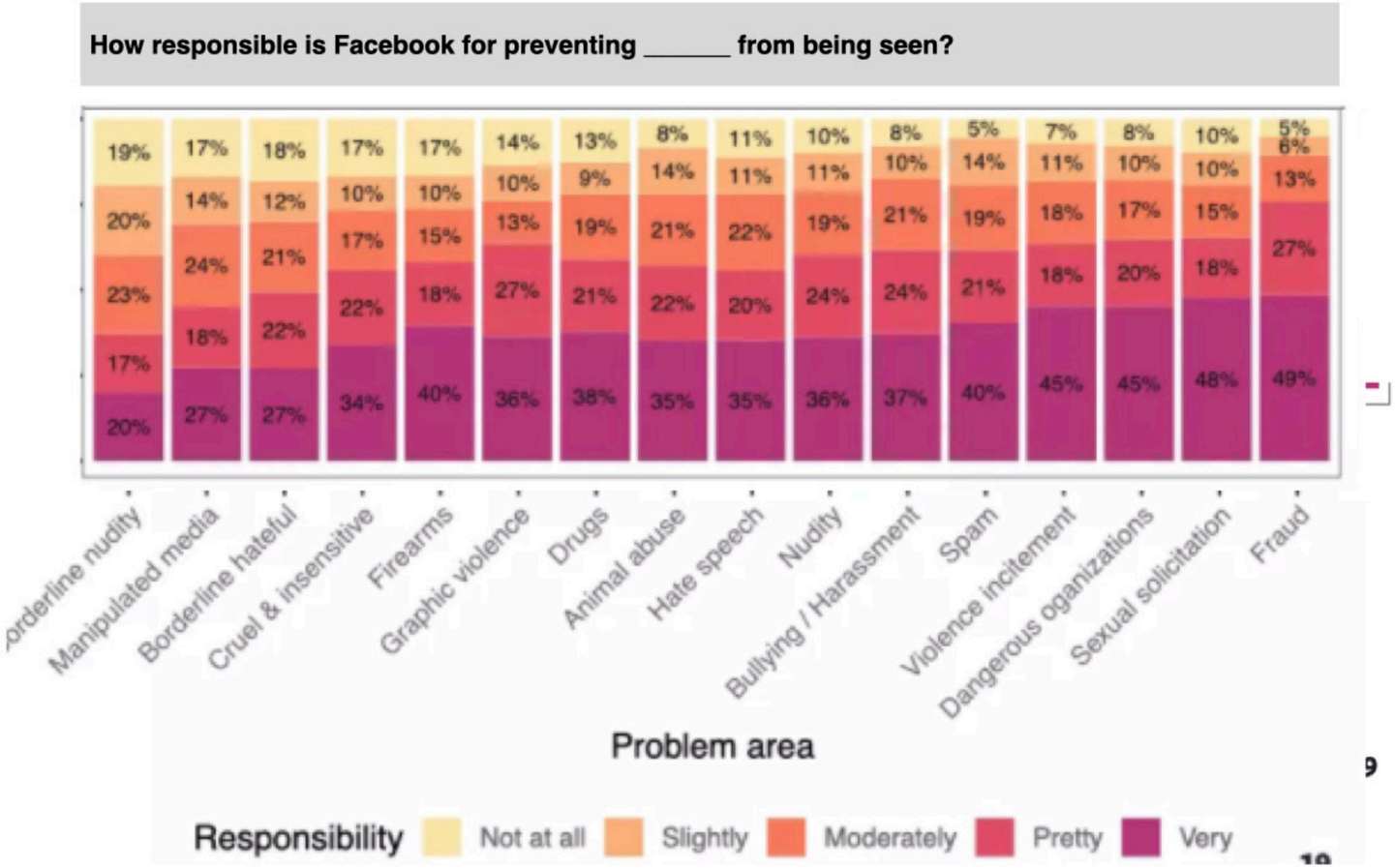


TABLE OF CONTENTS

70.7%

- What do people want us to do about borderline content?, Jess Bodford, Eric Chen at al, August 2018

Users hold Facebook responsible for addressing both violating and borderline content



Majority of respondents felt Facebook is pretty or very responsible for preventing 13 out of 16 problems (Powell 2020)

"I think that I would rather just see something where it was covered up and ask me, "Hey, are you okay with seeing this content?" And then that is my right."

- Female participant on seeing animal abuse in her Feed.

Seeing borderline content on Facebook makes participants feel like Facebook **does not care** about them. They say Facebook is….

"unsafe, a jungle" (Chile)

"dirty and negligent" (Kenya)

"irresponsible" (Bangladesh)

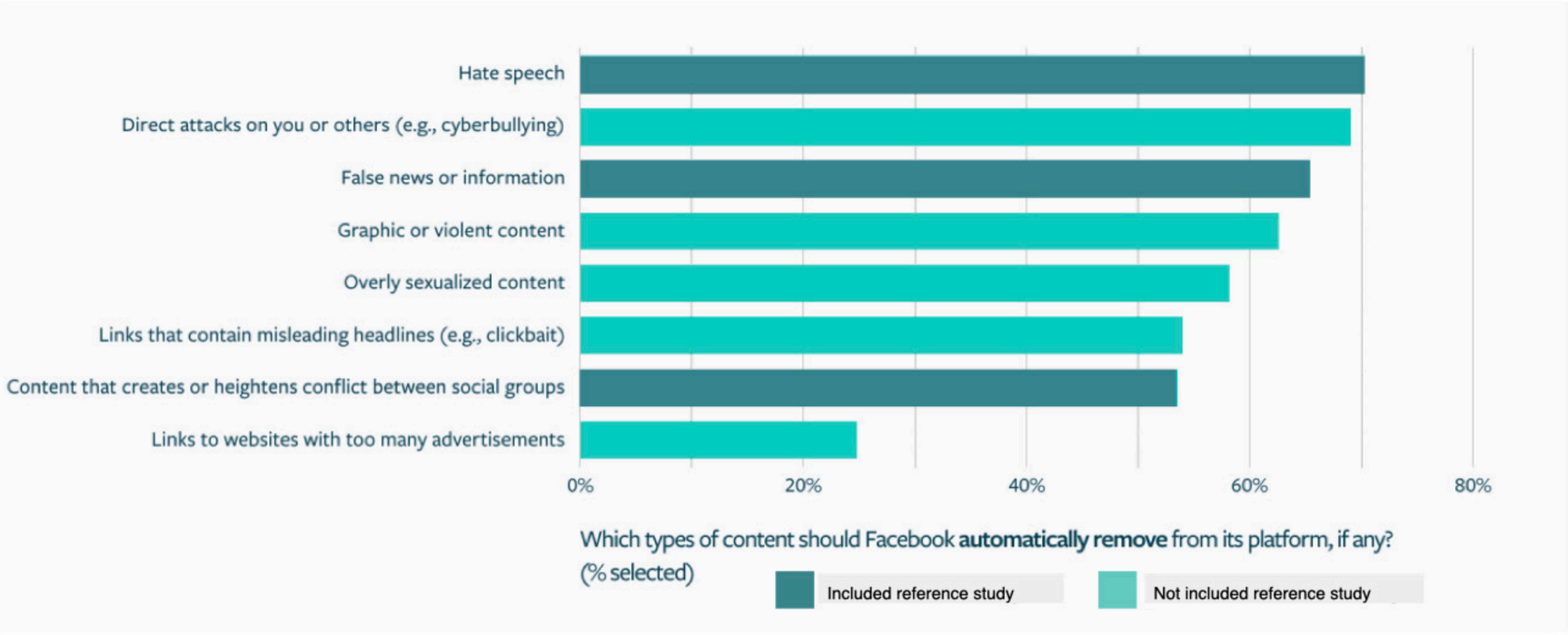
"does not care and is unreliable" (Thailand)

"unhealthy, untrustworthy, careless" (South Korea)

MINDSPARK RESEARCH INTERNATIONAL

- User Needs for User Controls over Borderline Content, Bryan Carroll Nov 2019

Most people think Facebook should automatically remove a range of integrity-related content

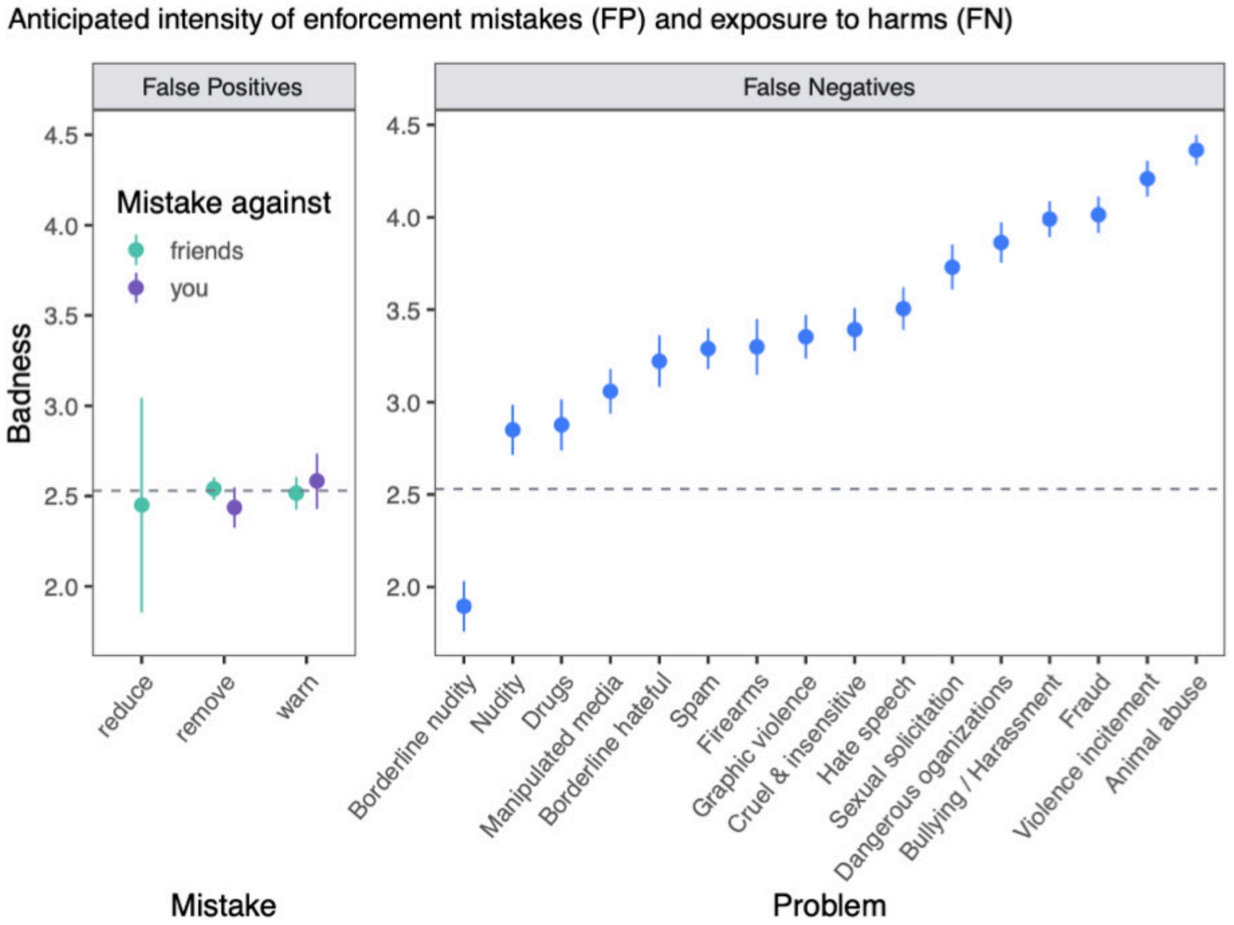


Attitudes towards Facebook's role in harmful content varies with a range of user characteristics

1. Females want Facebook to take more action than males do.
2. Latino/a and Black/African-American individuals want Facebook to take more action than White individuals do.
3. People who say they recently saw harmful content on Facebook want Facebook to take more action than those who have not.
4. Democrats want Facebook to take more action than Republicans.

- What do people want us to do about borderline content? Jess Bodford, Eric Chen at al, August 2018

Users anticipate exposure to most integrity harms would be worse than false-positive enforcements

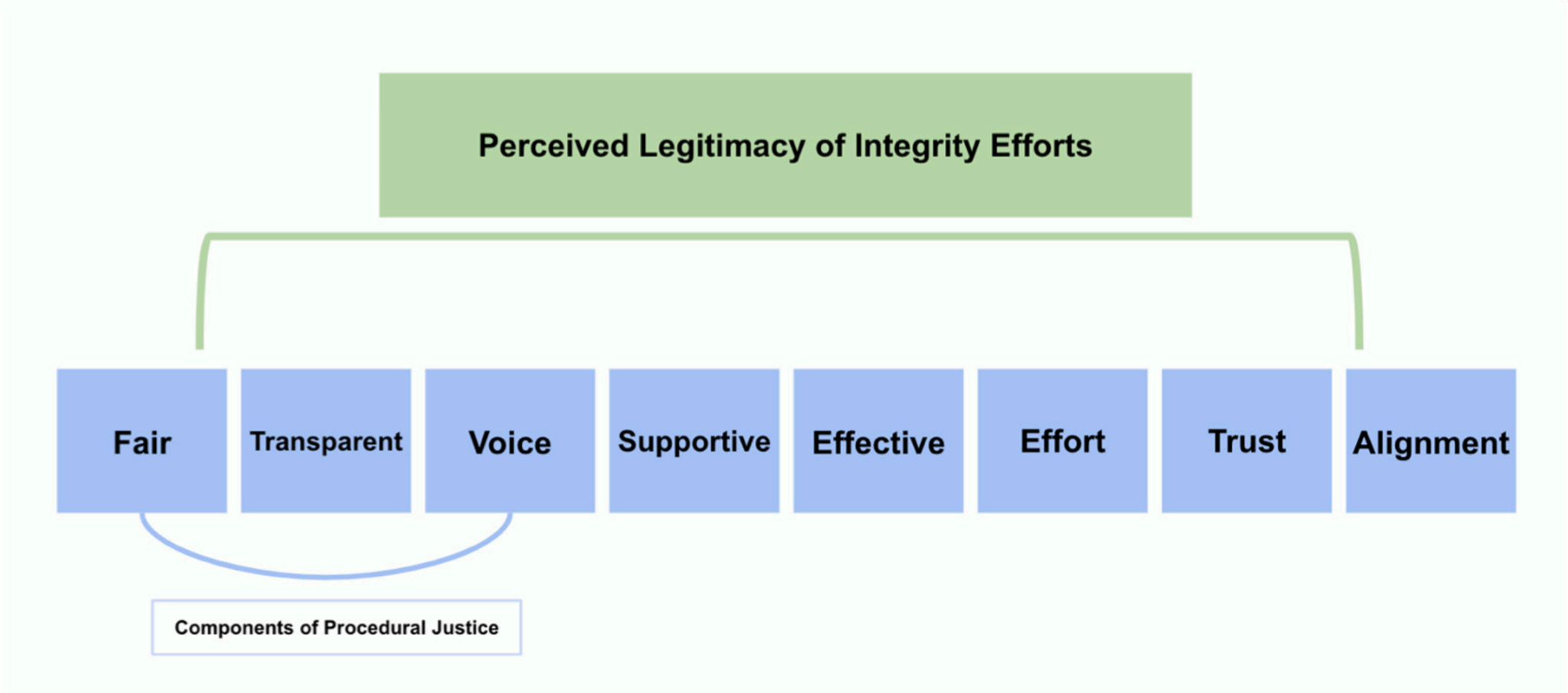


Enforcement Attitudes Survey (Powell, 2020): https://fb.workplace.com/permalink.php?story_fbid=2253290654784 50&id=100040040738159

TABLE OF CONTENTS

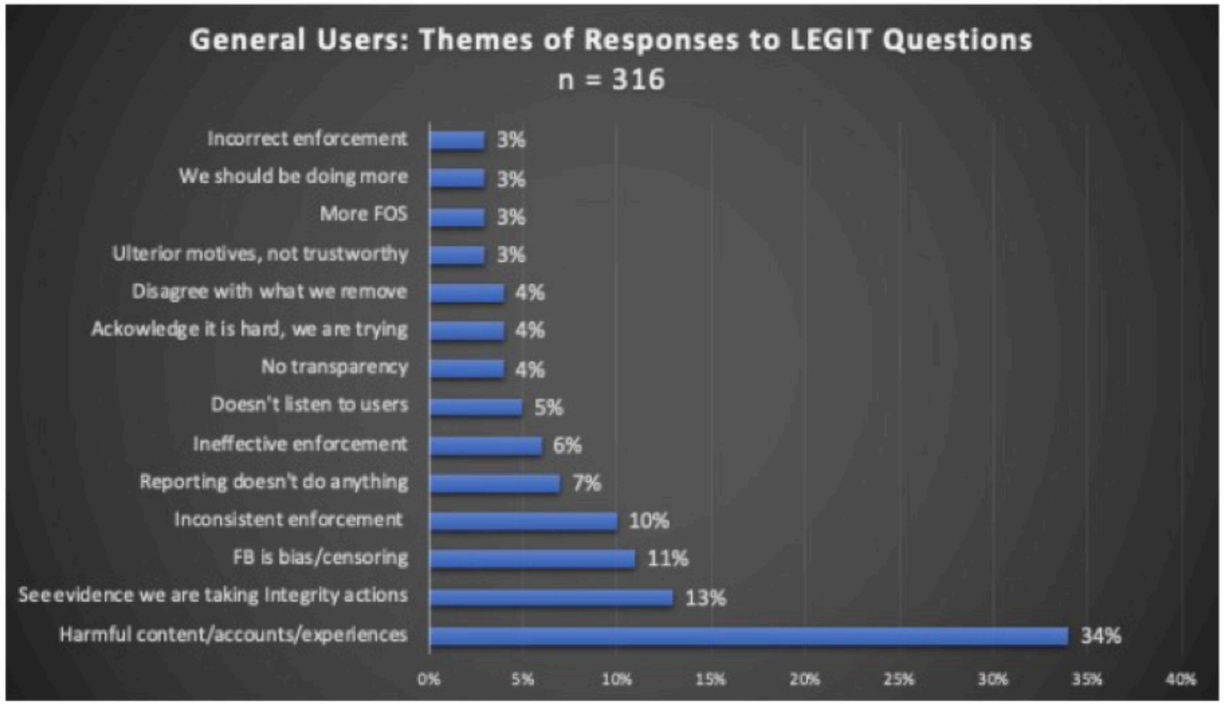
What is legitimacy and how do we measure it?

If our integrity efforts are legitimate, it means that people and external stakeholders believe that our integrity work is **effective** at reducing harm and that our enforcement is **defensible** and **fair**.



Biggest legitimacy detractor: Harmful content/accounts/experiences

- People who report seeing graphic violence or hate speech felt Facebook was significantly less effective at reducing harm on the platform
- Those who recently saw graphic violence also had more negative perceptions of how hard Facebook is trying
- We must be also careful that as we reduce prevalence, users perceive that we enforce **fairly** and **consistently**



How does seeing objectionable content relate to perceptions of Facebook's legitimacy? BRANDON CARLISLE Aug 2020

People don’t trust that Facebook is actually reducing bad content on the platform--or that we are motivated to do so

Dear Facebook, I can't take it anymore. You keep feeding me and billions of others a whole bunch of lies and crap that is designed to influence people and spread conspiracies.

*You **get paid very well to do this** and because of that have not taken the necessary measures to prevent it. **You welcome it.***

Legitimacy measurement part 1: Understanding users and stakeholders' perceptions of our integrity efforts, **Emily Wood, Sept 2020**

Lack of transparency & understanding of **ranking** and **enforcement** cause suspicion and lead to perceptions of bias

1. People don’t have a good understanding of how ranking works.

   “*I mean, I don't know if they're actively pushing stuff down based off what I don't want to see, but I do think they're pushing stuff higher on my feed based off stuff I interact with. Like I mentioned, so maybe that just ends up pushing stuff down, but I'm not a hundred percent sure.*” - Male, 18 - 24

1. People don’t have a good understanding of how Facebook enforces its rules, and enforcement can seem inconsistent and arbitrary.

   “I think that would be a good idea to put it in black and white, what they consider a hate speech and then follow their own rules. ***Don't favor one side or another. Hate is hate.”*** - Female, 55-64

1. Given the lack of understanding of our rules and ranking, the greatest value of our controls may be the transparency they provide into Facebook’s processes.

Negative experiences with our enforcement actions hurt legitimacy

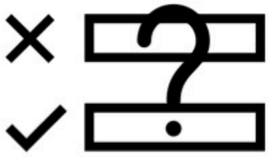
- Facebook both **over-enforced** on non-violating speech, and **under-enforced** against clear violations of policy, leaving a Black user with low trust in our systems.
  - *I have friends who, say things on Facebook, Who are saying things that aren't hate speech, but they'll get their accounts suspended. But then there are people who will say things like, Oh, these monkeys are over here protesting, or they'll say the N word or things like that. And they'll be perfectly fine never get in trouble or anything, basically let than go by. But my friends can't even say that they're tired of police shooting people and they get suspended from their Facebook.* - Female, 18 - 24
- For another, the **lack of clarity on why** a post was actioned on made her more skeptical of Facebook's systems and enforcement.
  - *"Not myself, but the magazine got banned, it said, "You cannot sell this, it's inappropriate". And they don't tell you why. And I've written them and I never heard anything."* - Female, 55-64

“Between your clearly racist and misogynistic ‘moderation’ to your laughable appeals process, there is a reason you are called RACEBOOK. Clearly I am not included in the community you claim to protect.”

Legitimacy measurement part 1: Understanding users and stakeholders’ perceptions of our integrity efforts, Emily Wood, Sept 2020

Today's realities make legitimacy challenging

low overall user digital & CS literacy makes meaningful transparency difficult →



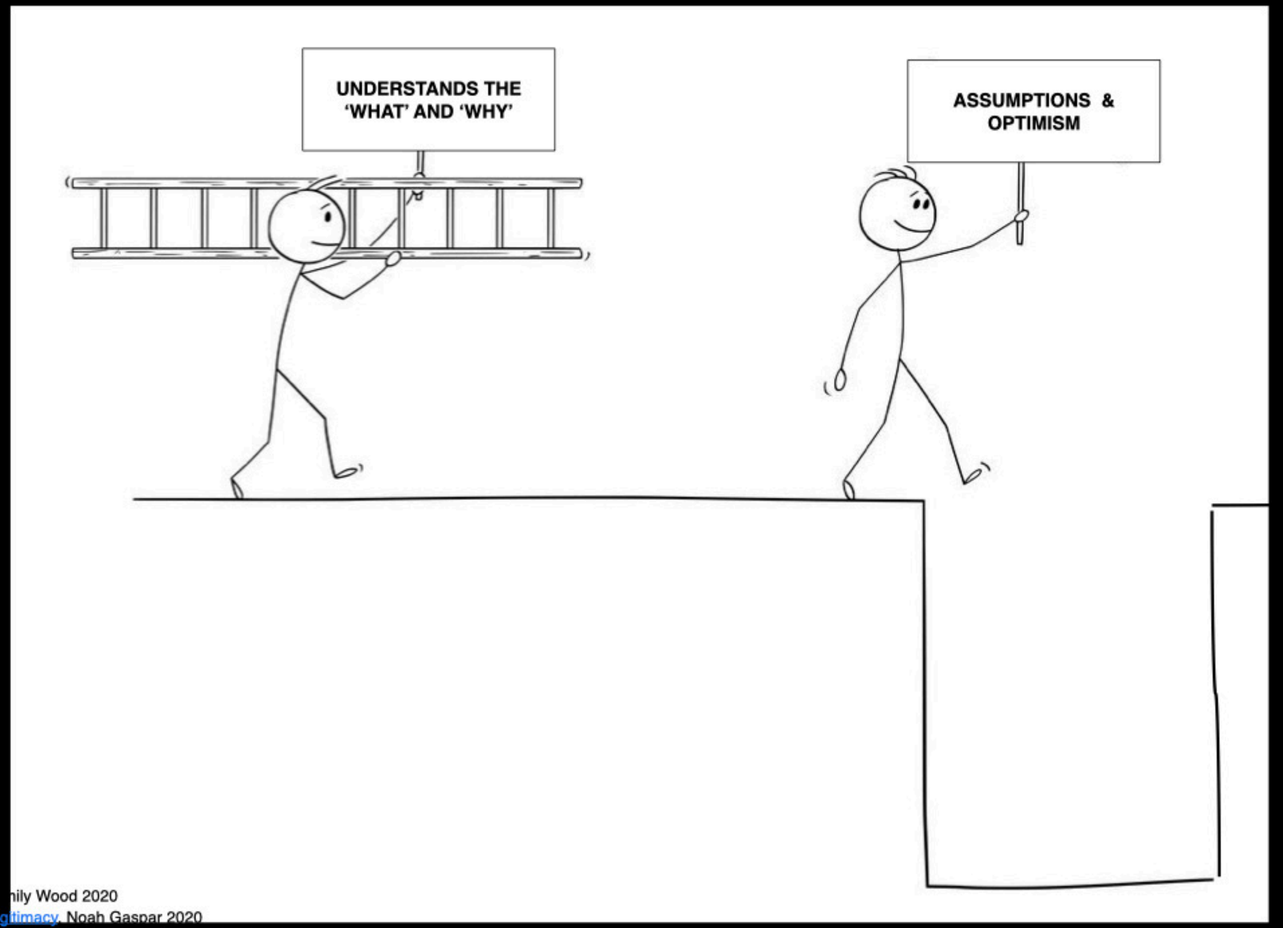
Users don't understand the rules →

Users feel enforcement is inconsistent →

Users don't know how to reform →

Some **surface-specific** examples:

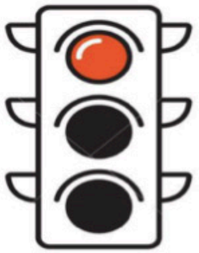
- **Marketplace sellers** don't know what our commerce policies are but are frequently flagged or banned for violating them →
- Most **Group admins** claim to be aware that Facebook has community standards, but many demonstrate a lack of understanding about implementation, & feel unsupported →



Demotions Transparency: India + Egypt Research nily Wood 2020
Demotions transparency does not a priori increase legitimacy, Noah Gaspar 2020
Legitimacy measurement part 2: Developing a framework for building legitimacy with users
Understanding How Context Can Inform, Empower, and Increase Trust in FB: Evaluating the Helpfulness of Integrity Signals in Produc

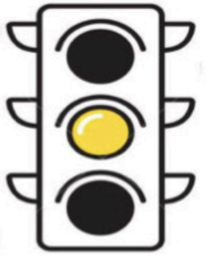
In addition, there are significant gaps in our enforcements

Quantitative analysis indicates there **are gaps in our enforcements**. →



High severity harms w/ enforcement gaps

- animal abuse
- violence incitement
- animal sales
- fraud
- bullying
- divisive content →



Low severity harms w/ enforcement gaps

- cruel/insensitive
- sexual solicitation
- firearm sales
- drug sales
- Voter misinfo
- impersonation

There are also **needs related to specific content types**, for example:

- Key user-reported pain points with **news** indicates we need to do more on: clickbait; purposefully misleading; and emotionally manipulative news →
- We over- and under-enforce in groups based on harm type →
  - For example, we underenforce V&I in complex entities, such as groups pages and events

For effective transparency, we must ensure...

Literacy: most users just don't understand our policies or process - this lack of baseline understanding will make meaningful transparency difficult

Messaging: communication of our process & policies is inconsistent - will require streamlining, unifying

Relationship: our process & policies impact FB-user relationship, positively & negatively; we can optimize

Agency: to yield benefits, we need to ensure affected user feels empowered as active participant with recourse

Confidence: user confidence that our process and policies are legitimate, fair, etc is critical to launch

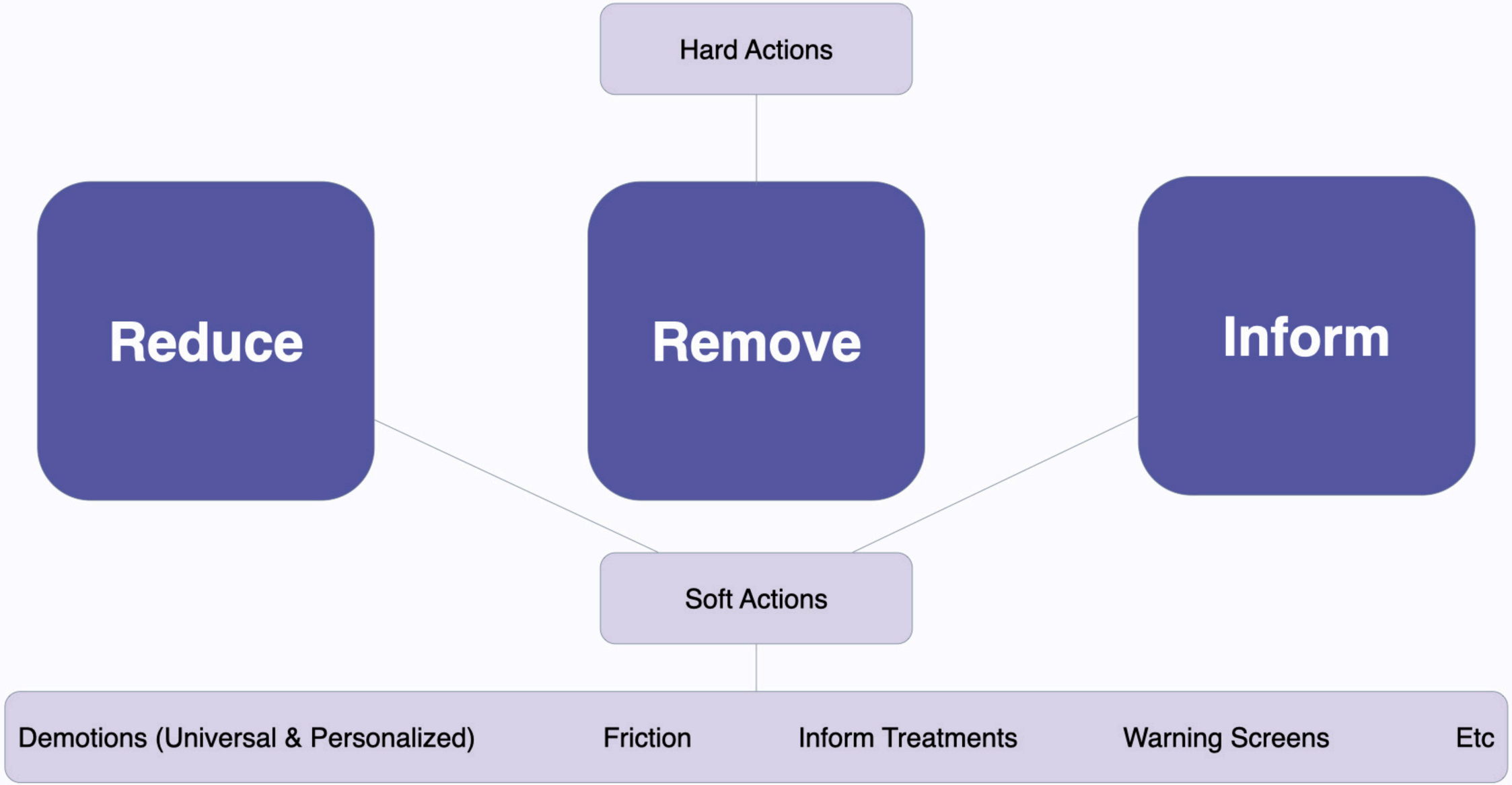
Accuracy: we must ensure user perception that our automated technology accuracy is high

Consistency: users need to feel that our process, policies are consistently applied

Flexibility: our process, policies, technology must be able to adapt to local needs, contexts, nuances

TABLE OF CONTENTS


“It must be a personalized tool based on things I keep on reporting or have reported, for them to adapt to what I consider offensive, since it is subjective.”
I.C.
23, female, Chile


Hard Actions
Reduce
Remove
Inform
Soft Actions
Demotions (Universal & Personalized)
Friction
Inform Treatments
Warning Screens
Etc

Users want more proactive interventions and autonomy.

Soft Actions - Facebook Intervention

- Users see downranking as a **valid but insufficient** as a way to protect them from harmful experiences.

- Warnings and/or additional context fills an important gap in our enforcement systems and can help users feel empowered, as well as increase trust in Facebook.

- Warning screens are the **most effective tool** in our toolkit to prevent bad experiences from borderline content:
  - Warning screens are effective, they are **viewed positively** by users, and they **drive legitimacy**.
  - Warning screen Inaccuracies are generally tolerated; False Positives in particular are seen as fairly harmless

Controls - User Autonomy

- **Users want content controls** to reduce negative experiences, make them feel empowered and let them tailor content to personal preferences
  - In a U.S. study, the majority of people want settings to control borderline content, defaulted to automatic reduction or removal. (What do people want us to do about Borderline Content? US survey,Jess Bodford, Eric Chen et al, Aug 2018)
  - Controls also provide more transparency into our content moderation efforts (Carroll 2019), with positive implications for legitimacy.
  - Controls also **saves them** from having to **unfriend/ unfollow friends or family** with different views (Leavitt 2017).

- Existing controls are underused; **discoverability** is a barrier.

Soft Actions

Outside of removing content, we have a wide-range of integrity interventions we have or are currently investing in.

Demotions | Friction | Inform | Controls

- Universal
- Personalized

We use implicit and explicit signals to build ranking models.

How are we reducing Bad Experiences?

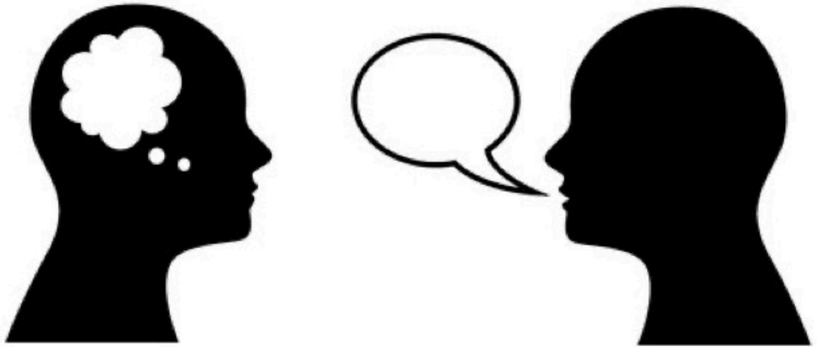
Personalized Integrity Models are designed to reduce the distribution of borderline content for only those users are most likely to feel like that content causes them a bad experience

How do we know when something causes a bad experience?

We use *explicit and implicit signals* to assess tolerances for potentially problematic content that is most likely to cause bad experiences

Implicit Signals:

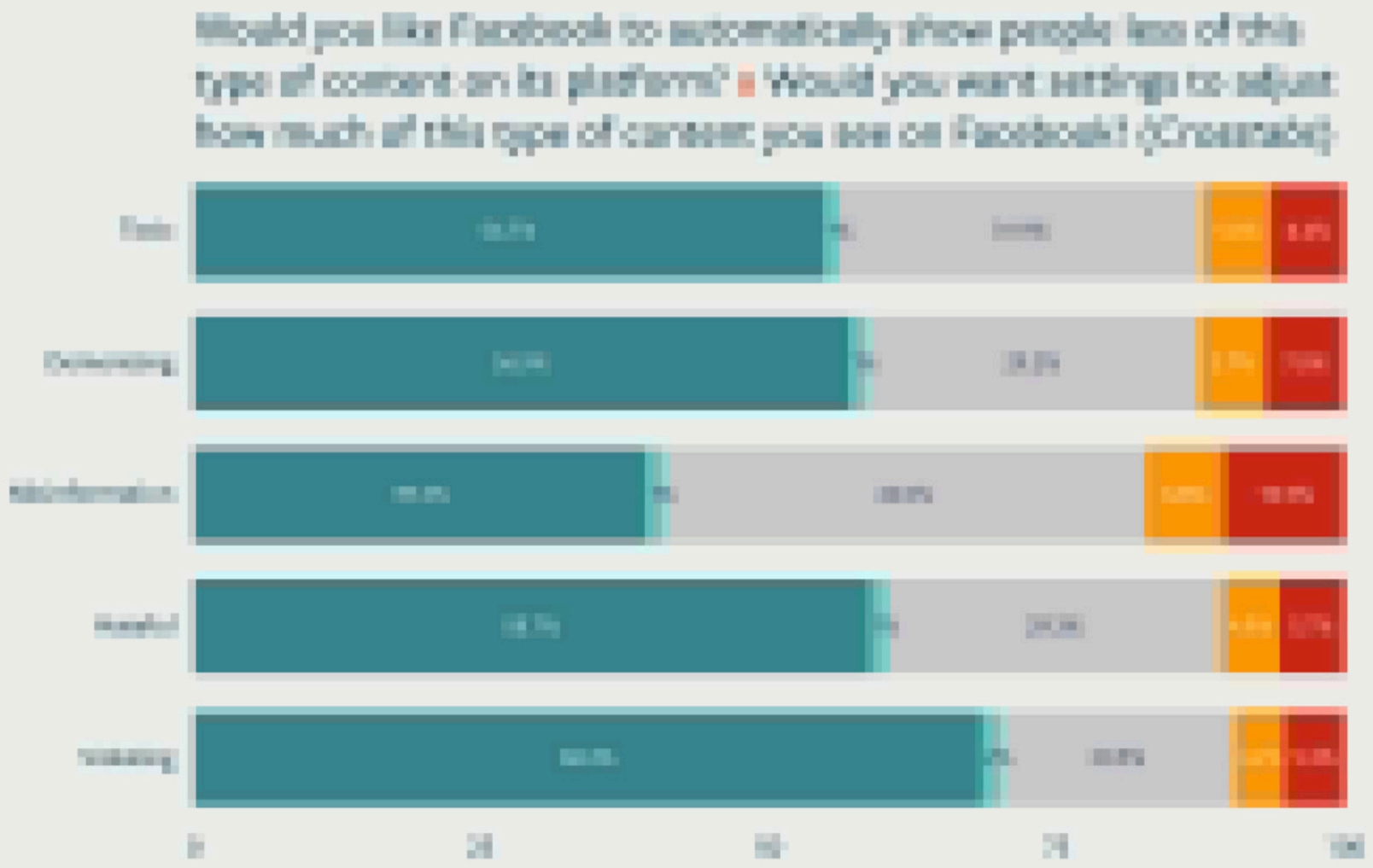
On-platform behaviors are used to as indicators of bad experiences. We implicitly assume that when users engage in any number of actions that they did so, in part, because they had a bad experience



Explicit Signals:

User-level and content-level surveys about potentially problematic content provides explicit signals about whether something caused a bad experience.

Most people want automatic reductions and settings to control harmful content exposure, or they don't care.



Grace Jackson, 2018

Soft Actions

Outside of removing content, we have a wide-range of integrity interventions we have or are currently investing in.

Demotions

- Universal
- Personalized

Friction

- Comment Friction
- Post Friction
- Reshare Friction
- Safety Notice
- Search Friction
- Group Join Friction
- Page Like Friction

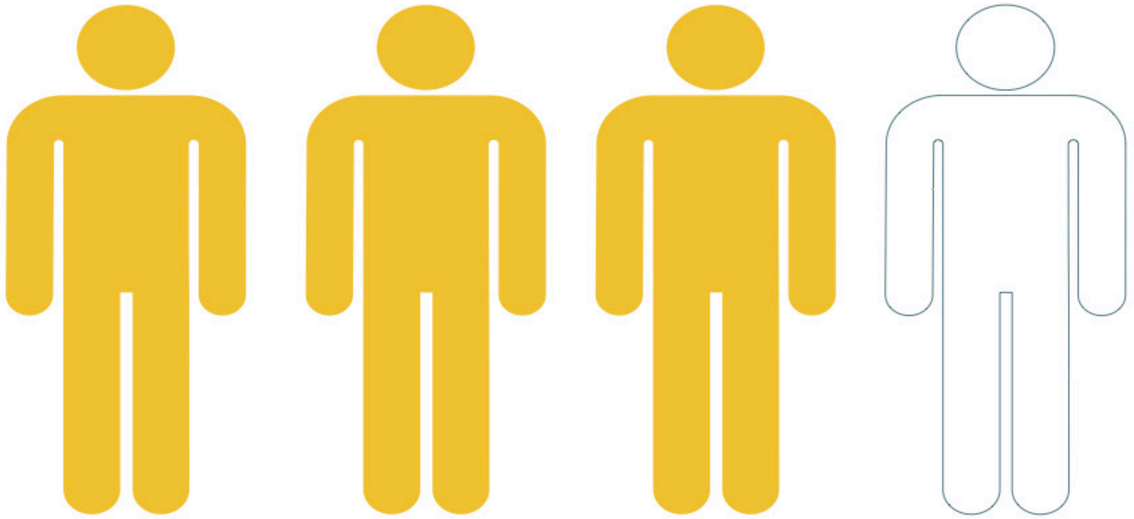
Inform

Controls

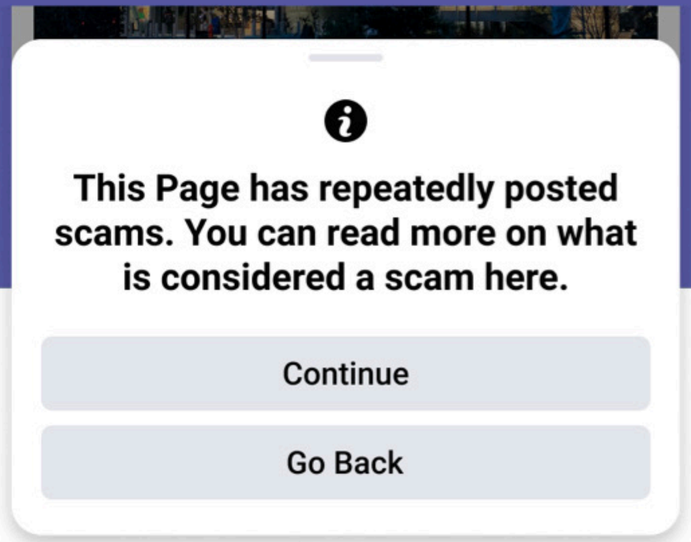
3-in-4

users want Facebook to inform them about low quality posts, even when the post doesn't violate Facebook's rules.

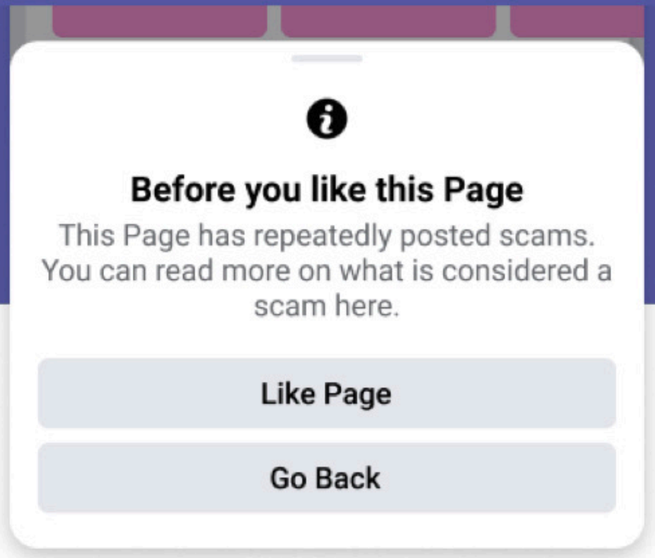
Even in the US, where more users say Facebook should do nothing, the **overwhelming majority want to be informed**, suggesting inform treatments like reshare friction and other interventions will be welcomed by users.



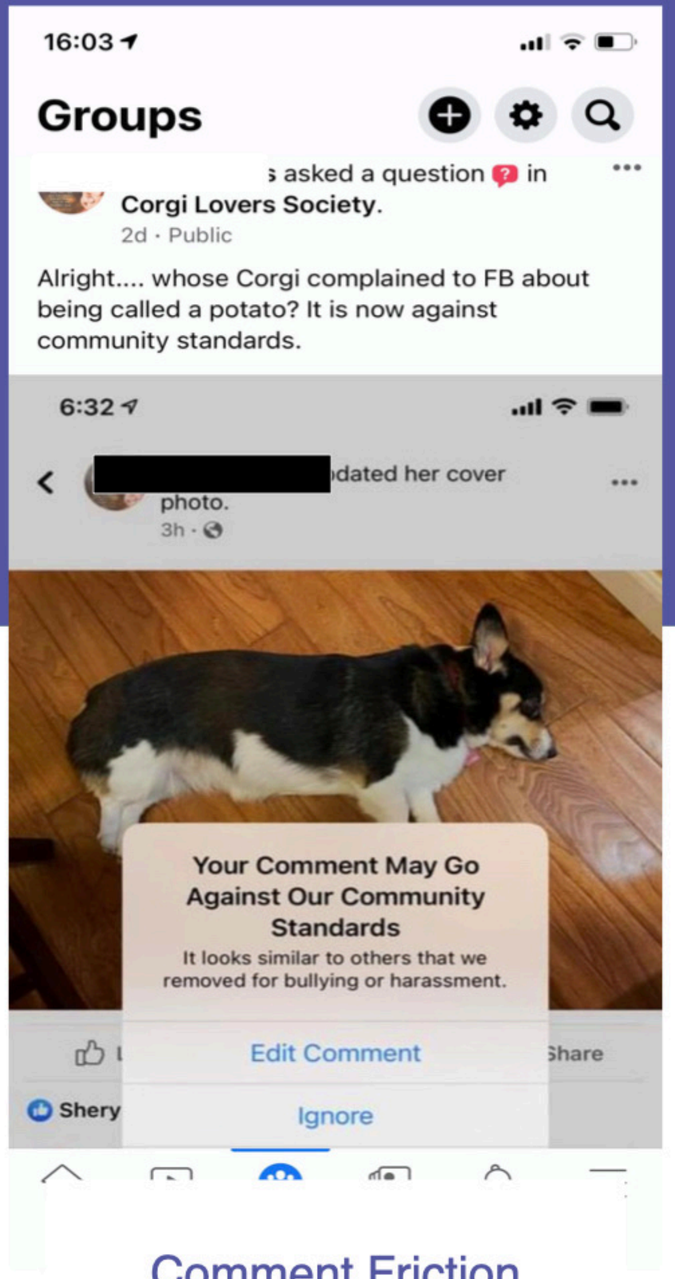
Context Entry Points



Reshare Friction



Page Like Friction



Comment Friction

Soft Actions

Outside of removing content, we have a wide-range of integrity interventions we have or are currently investing in.

Demotions

- Universal
- Personalized

Friction

- Comment Friction
- Post Friction
- Reshare Friction
- Safety Notice
- Search Friction
- Group Join Friction
- Page Like Friction

Inform

- Context Button
- Inform Labels
- Metadata
- Warning Screens
- WS Actor Experience

Controls

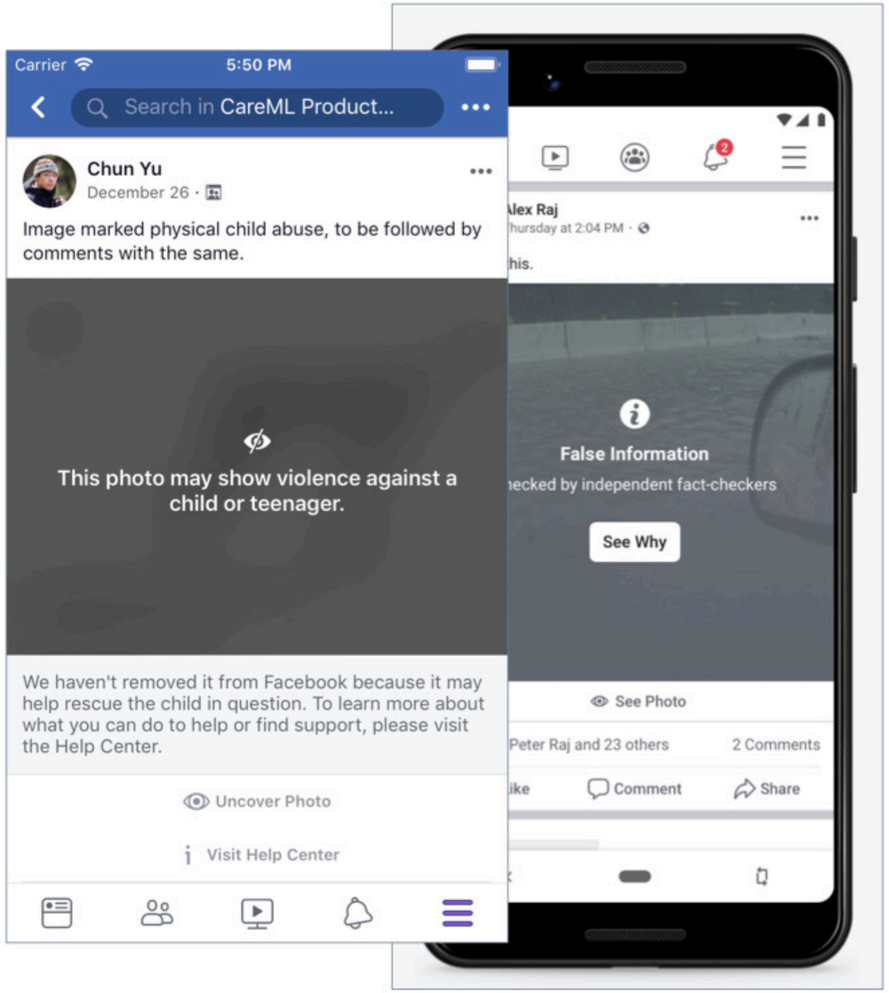
Current use of warning screens

Content warning screens are applied to content (e.g., videos and images) that is considered borderline (adjacent to Community Standard harms, but non-violating), when not every viewer may be mentally or emotionally prepared to view the content (Wiki). The intent of the content warning screen is not to punish the creator, but to protect viewers and give them control

Currently Use Cases:

- **Mark as Disturbing (MAD):** Graphic violence, physical child abuse
- **Mark as Sensitive:** Nudity in medical settings, religious animal sacrifices, certain depictions of abortion, images of SSI, and others
- **Mark as Mature:** Blocks certain videos from those under 14.
- **Mark as False or Partly False:** Third-Party Fact-Checked misinformation in photos, videos, and links.

Currently, if a screen is applied, it is applied **universally**; every instance of the content on the platform will be covered



[Exapnding Warning Screens: Integrity Research POV]

Expanding coverage of warning screens could address key user concerns and legitimacy issues...

- Warning screen use could be expanded to cover any content a viewer may not be mentally or emotionally prepared for
- This can be done on a **personalized basis** - except in the case of misinfo treatments. Reduce thresholds for applying warning screens and use personalized models to predict how low thresholds should be for a given user based on their individual tolerances
- Increase the scope of what these personalized warning screens cover; **animal abuse, hateful/toxic language** should be candidates for expansion
- Increase use of info treatments to additional kinds of content for additional categories of misinfo content, and expand pool of raters who can trigger warning screens

Improve UI to mitigate drawbacks and solicit feedback:

- Utilize straightforward messaging that explains why and how the content was covered. Utilize humility and be open that we may have gotten it wrong
- Provide opportunity for users to give feedback on appropriateness of cover, **utilize this feedback to tune algorithm**

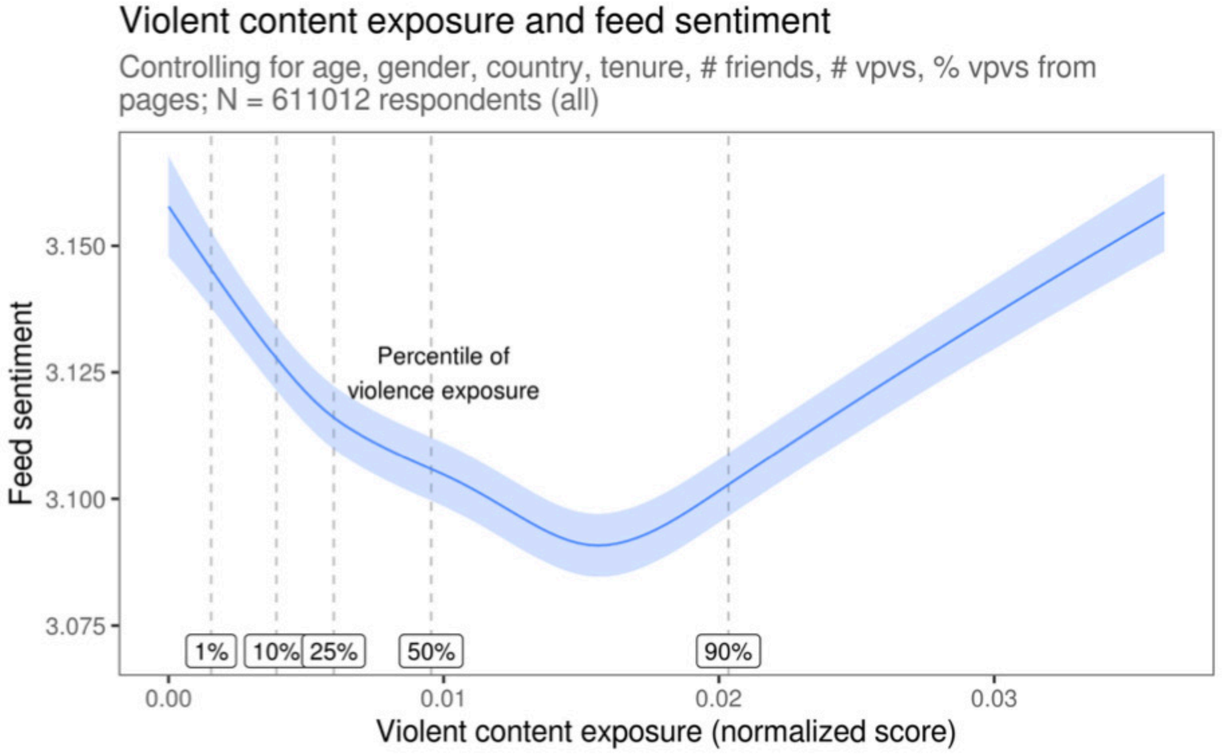
Personalize coverage to reduce cluttering

....though too many warning screens could lead excess friction and bad user experience for consumers

Warning screen applications could be **personalized per-user** to keep the frequency of their use low per user.

Exposure to graphic violence is most negatively associated with sentiment for users who rarely see violent content.

Personalizing warning screens would allow us to protect these most vulnerable users without cluttering feed for users with higher tolerances for violence.



Exposure to integrity harms and feed sentiment (Powell, 2019): https://fb.workplace.com/permalink.php?story_fbid=1779660002147 57&id=100040040738159

Soft Actions

Outside of removing content, we have a wide-range of integrity interventions we have or are currently investing in.

Demotions

- Universal
- Personalized

Friction

- Comment Friction
- Post Friction
- Reshare Friction
- Safety Notice
- Search Friction
- Group Join Friction
- Page Like Friction

Inform

- Context Button
- Inform Labels
- Metadata
- Warning Screens
- WS Actor Experience

Controls

- Topic controls
- Demotion controls
- Preference signals
- Lightweight Negative Feedback (x-out)
- Reporting
- Self-Remediation Tools
  - Hide
  - Unfollow
  - Block
  - Snooze

Why does control matter?

Currently, users feel as if they lack control over the content they see in Feed.

This sense of a lack of control, further exacerbated by the presence of unwanted content in Feed, leads to a strong want and need for user controls.

Our existing controls, broadly defined, are underused and do not properly serve users in the way we intended.

One of the greatest barriers to adoption is due to discoverability.

Providing users with greater control, either through new controls (easier ways to hide or ranking controls) or the simplification of older controls, will empower users and increase user sentiment toward Facebook.

Wrap-up

Summary of findings

1	**Hate speech, divisive civic content, and graphic violence** are frequently and intensely experienced, and have been shown to have a negative effect on sentiment and legitimacy, **particularly with repeated exposures over time.**
2	**Borderline content** can be seen as **equally or more harmful** than violating content and decreases sentiment and engagement. In most cases, users want Facebook to hide or remove it.
3	Post content is not the only problem--**toxic and divisive comments** commonly appear on **benign posts.** Reshares, Links and Status Updates are more likely to be rated as a Bad Experience compared to photos and videos
4	**Not every "bad experience" is unwanted.** Some respondents describe "needing to see" content they considered a bad experience, such as violence and racism.
5	Users want Facebook to act. They hold us responsible for negative experience, and most think Facebook should **automatically remove** severe integrity-related content and **hide** less severe content. They perceive **exposure** to integrity harms as **worse than false positive actions** on benign posts.
6	**User experiences, preferences and perceptions vary.** Reaction to content varies by gender, ethnicity, culture and other factors; sentiment of Low-exposure users is more affected by integrity harms; those with low digital literacy are more likely to see violating content; some may even deliberately seek out harmful content.
7	**Legitimacy** is challenged by **lack of transparency & understanding of ranking & enforcement**. Content controls such as 'sensitive content preferences' serve a double role - not only do they reduce exposure, they help the user feel they understand what's under the covers.

Wrap-up

For further discussion:

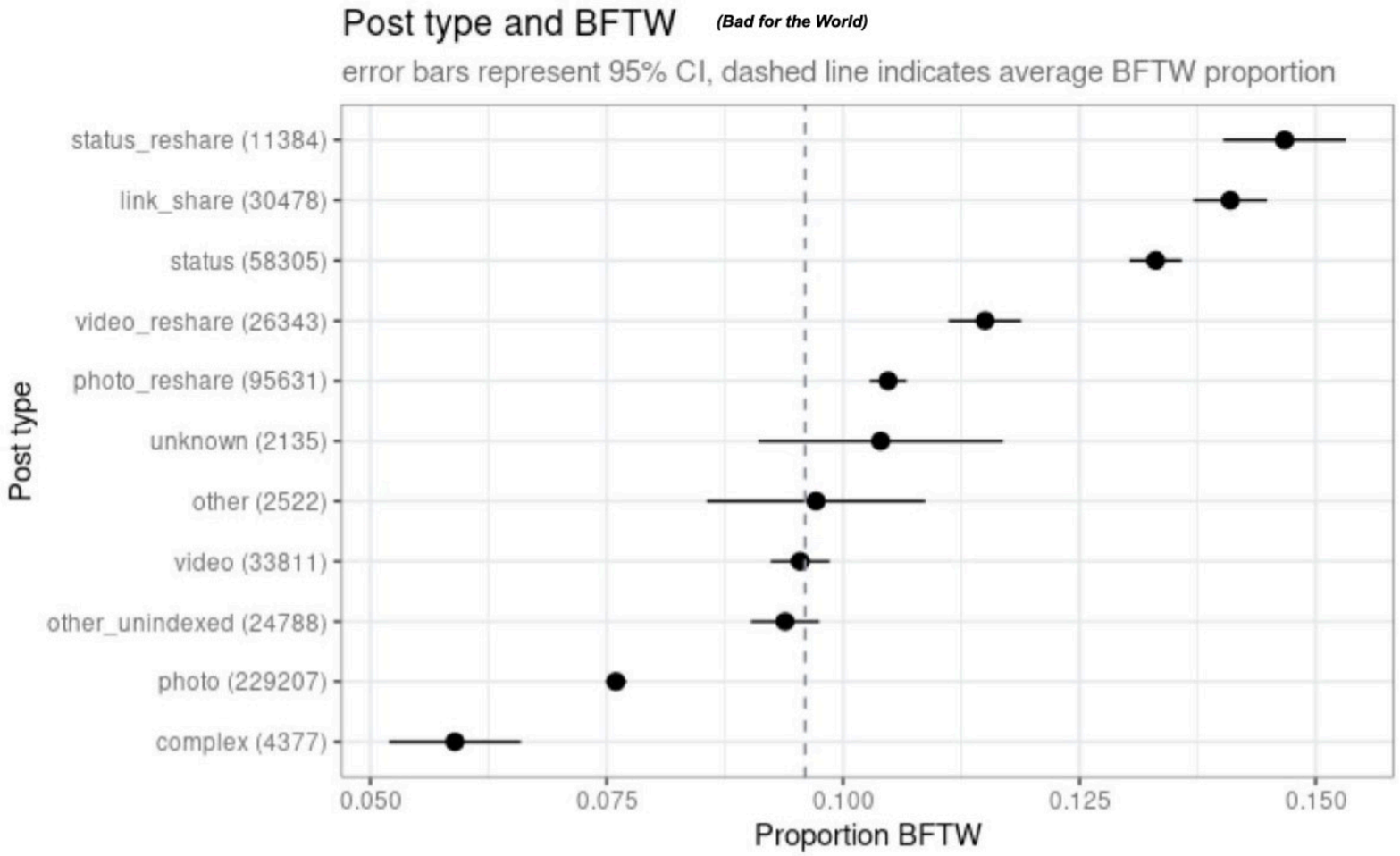
1. What are further opportunities for AI Integrity & FB App Integrity to partner to improve legitimacy?
2. How could AI could inform future iterations of Demotions, Friction, Inform, Controls?
3. Which user-facing solutions could provide valuable signals for AI and how?
4. How can we be mindful of potential ‘watch-outs’ like…?
    a. Not all bad experiences are unwanted
    b. Different groups of people may be differentially affected by our solutions and enforcements
    c. Different groups of people may have substantially different content preferences and reactions

Appendix

What observable attributes are related to the likelihood a FB user has/has had a bad experience?

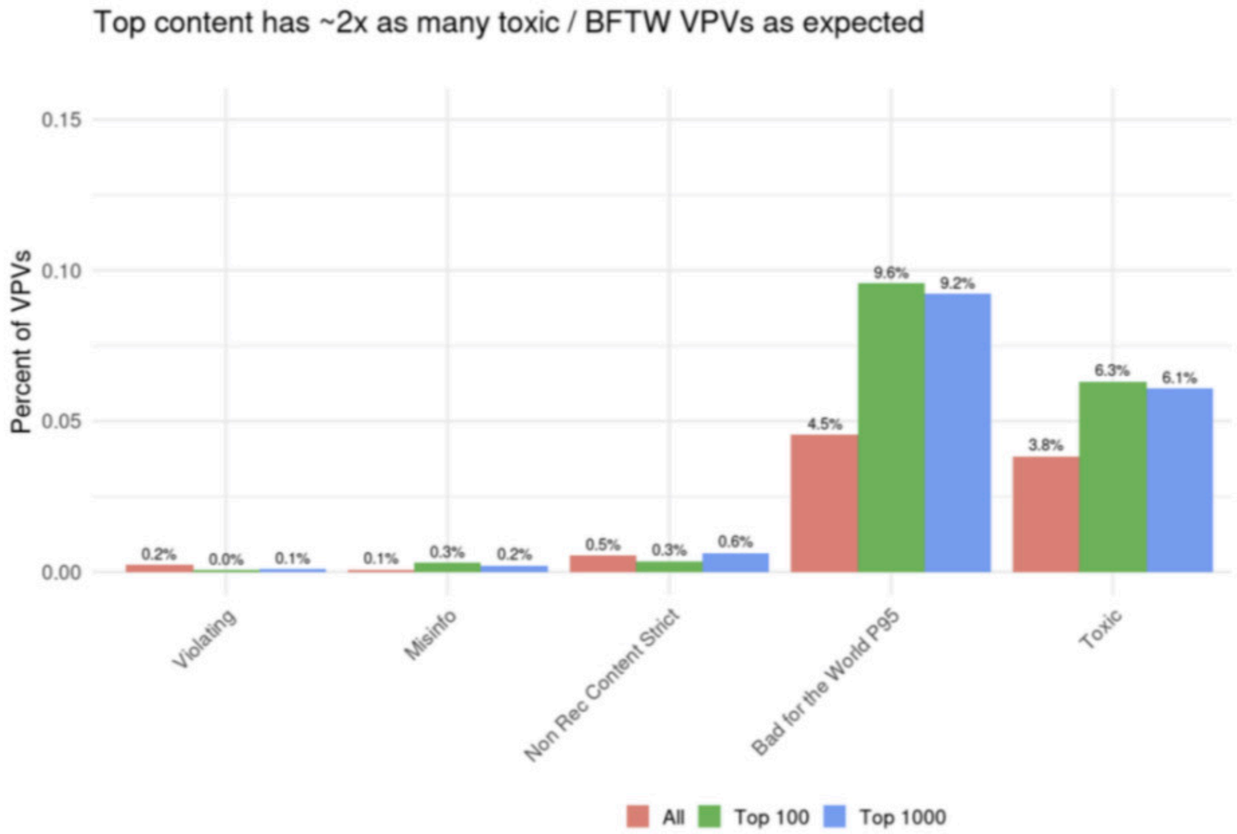
Reshares, Links and Status Updates are more likely to be rated as a Bad Experience compared to photos and videos

- **Reshares** are an affordance that disproportionately supports distribution of BFTW content
- **Status** posts have the highest prevalence of BFTW among original broadcast post types

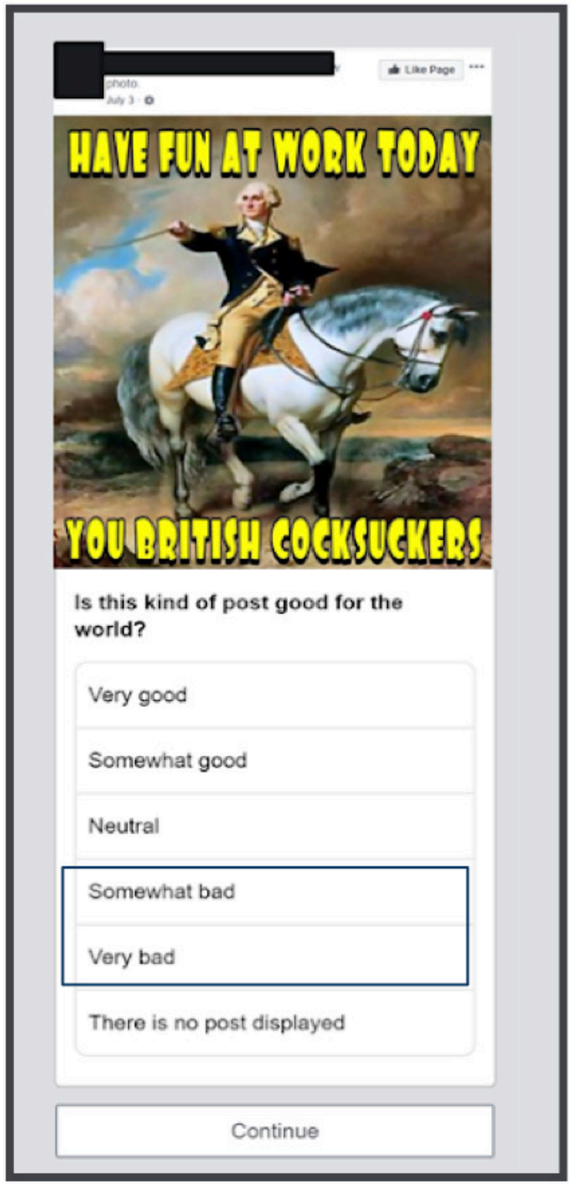


The content viewed most often is more likely to be BFTW than content overall

The more VPVs a post gets, generally the higher its toxicity and 'bad for the world' (BFTW — trained on user/content survey data) classifier scores will be, with top content (top 1K posts) having ~1.5-2x higher scores than content overall *(this pattern holds for public content, and content more broadly)*



Certain topics and behaviors are associated with Bad Experiences



- Some content topics, such as Crime & Tragedy and Civic Content are more likely to be rated as a Bad Experience by survey respondents
- Angry reacts are correlated with content rated as a Bad Experience
- Users with low digital literacy are exposed to significantly more borderline nudity and graphic violence in News Feed →
- Men and younger users are more likely to experience exposure to dense clusters of violation types

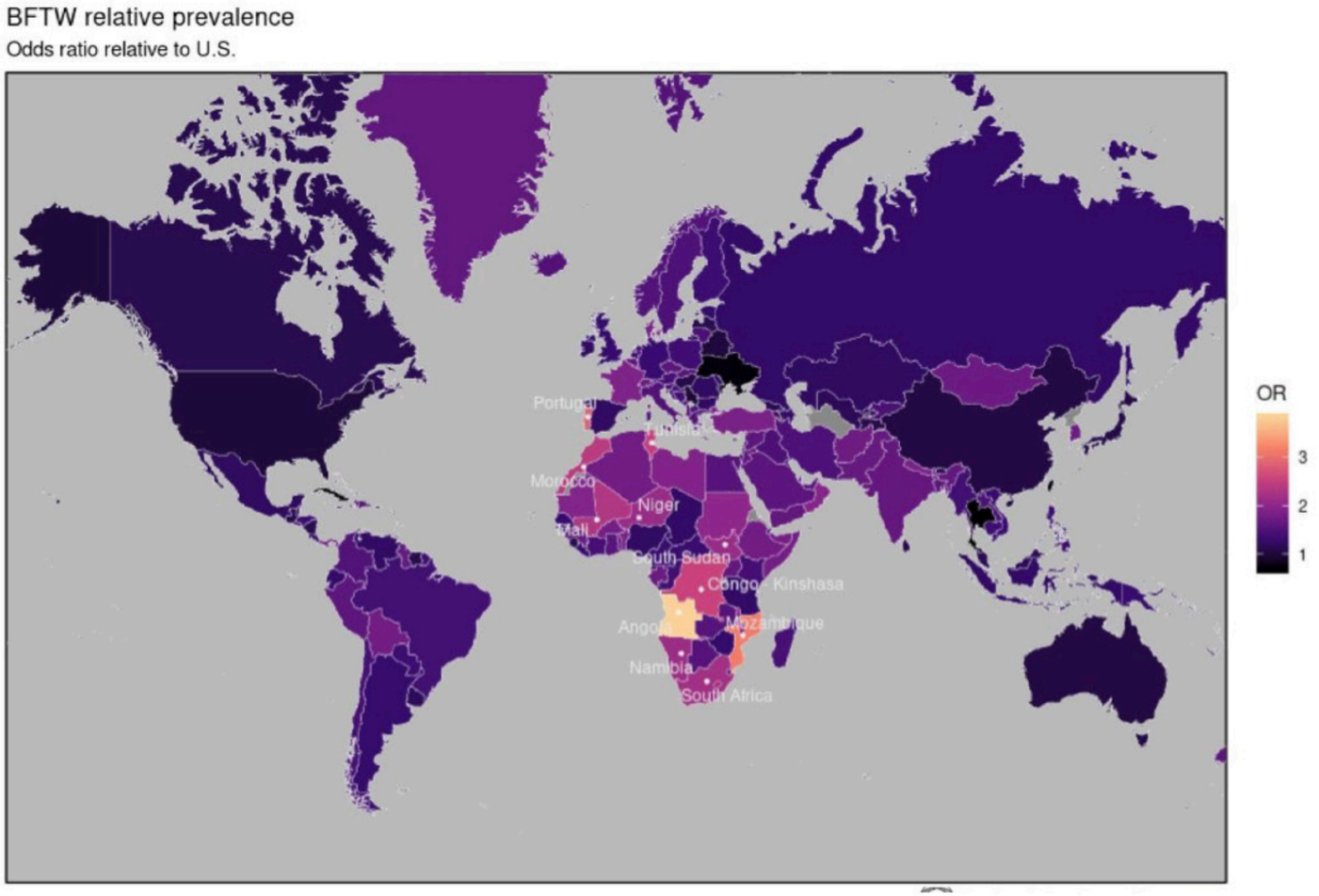
Patterns of User Exposure to Integrity Problems on Facebook, Gary Lewis Apr 2020

Some markets are disproportionately more likely to rate the content in their News Feed as bad for the world.

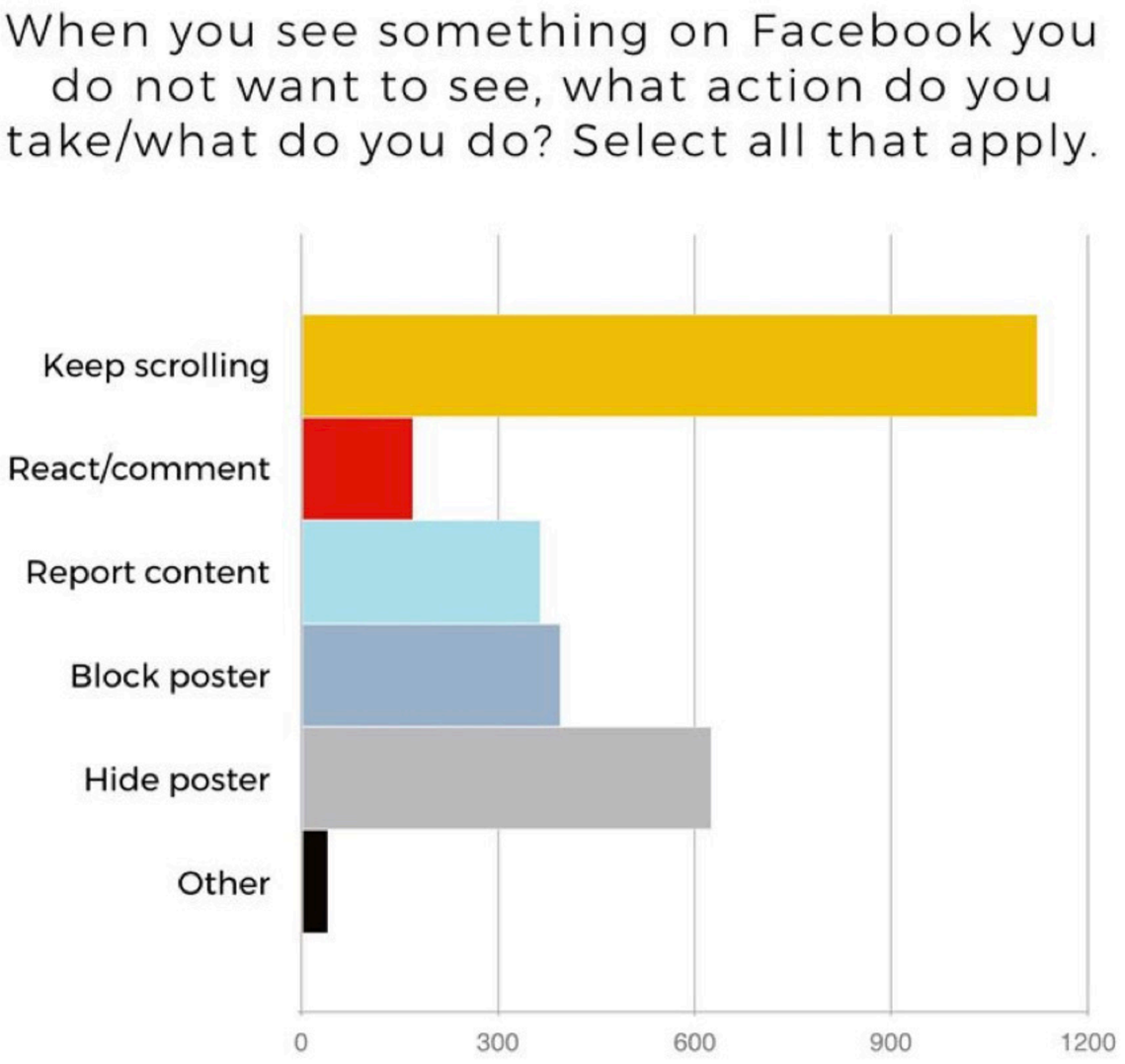
The color of each country represents the **estimated odds** that users in that country rate would rate posts in their News Feeds BFTW relative to users in the U.S. (OR = 1 means same as U.S.)

Highlighted countries are those where the odds of users rating content BFTW were > 2x as high as in the U.S.

Caveat: further investigation is needed to determine whether translation and language differences could partly explain these differences



Most users scroll past bad experiences, taking no action and providing us with little insight into why they avoided a piece of content.



LEGIT survey question wording

Fair: Facebook's rules refer to what is not allowed on Facebook. Do you think Facebook makes consistent decisions about posts that are not allowed? (fair_2)

Transparent: When it comes to removing posts that go against the rules, how transparent is Facebook? (trans_4)

Voice: How often do you think Facebook listens to what people think when deciding what isn't allowed on Facebook? (po_1)

Supportive: When a person has a negative experience on Facebook, how supportive do you think Facebook is? (supp_1)

Effective: How effective is Facebook at reducing negative experiences on Facebook? (harm_7)

Effort: How hard is Facebook trying to reduce negative experiences on Facebook? (harm_3)

Trust: When it comes to removing posts, how much do you trust Facebook to do the right thing? (trust_2)

Alignment: Think about the posts that Facebook does not allow. Are these similar to your beliefs about what shouldn't be allowed on Facebook? (align_1)

Feed Sentiment Survey - Revised (FSSR)

Tracking survey provides quantitative measure of users' sentiment toward News Feed. This study focuses on 4 main questions from the FSSR (averaged together).

Satisfaction
Overall, how satisfied are you with the posts you see in your News Feed?

WYT
In general, how many of the posts that you see in your News Feed are worth your time?

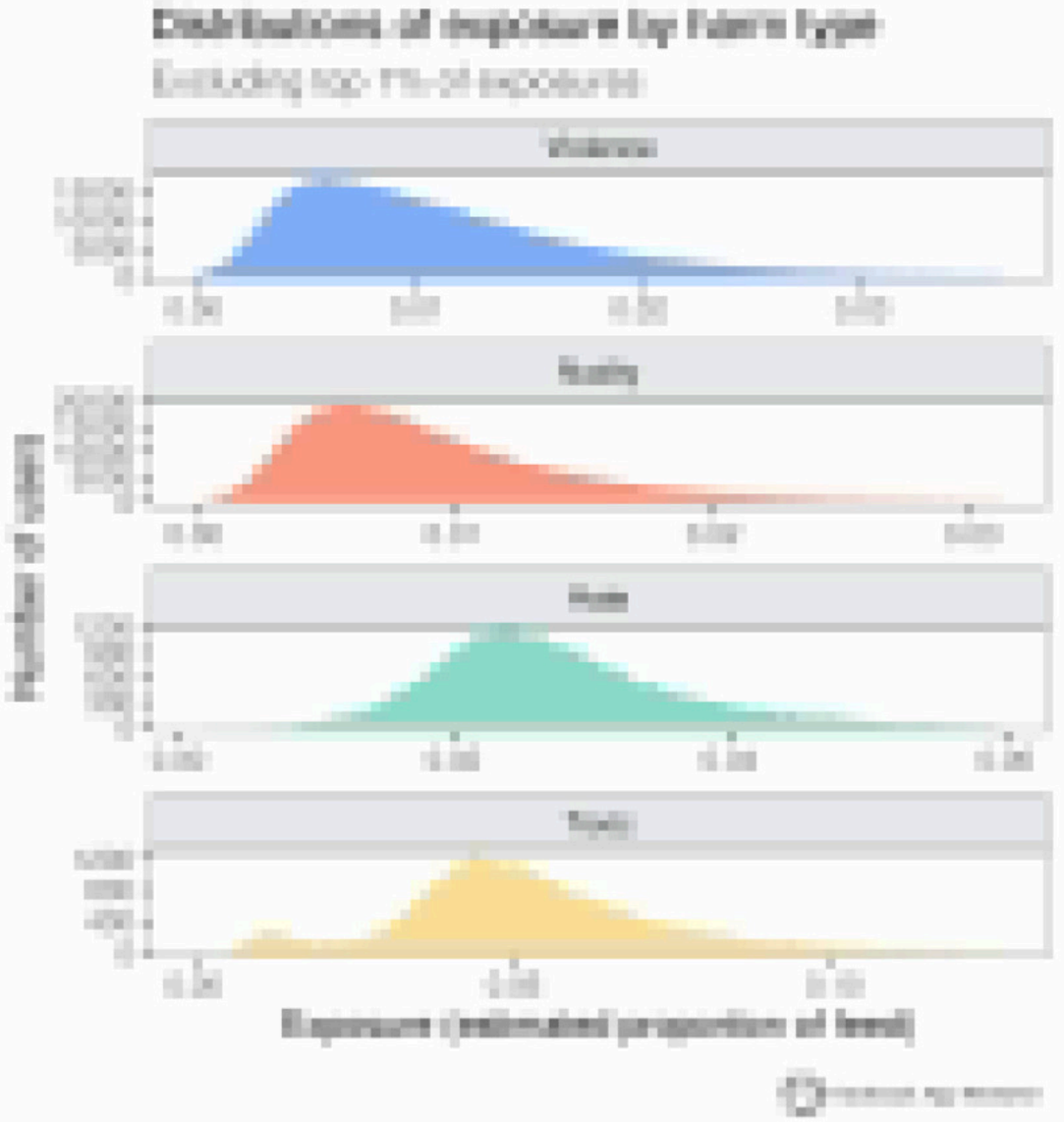
Meaningful Interactions
How meaningful are your interactions with people on News Feed?

Connected
How much do the posts you see in your News Feed make you feel connected to others?

[POC: Scott Freeman; Unidash: https://fburl.com/[illegible]]

Integrity harms and classifiers used

- Graphic violence (violating)
  - Global classifier
- Borderline nudity (3+)
  - Global classifier
- Hate speech (borderline)
  - Global classifier - analyses focus on U.S. users.
- "Toxic" content (see appendix)
  - U.S. only classifier
  - No active enforcements



https://fb.workplace.com/photo?fbid=177967676881256&set=pcb.177966000214757

Bad experiences can cause harm and detract from user value

Integrity problems can **cause harm for users via:**

An immediate negative emotional experience **or** **A loss of value**

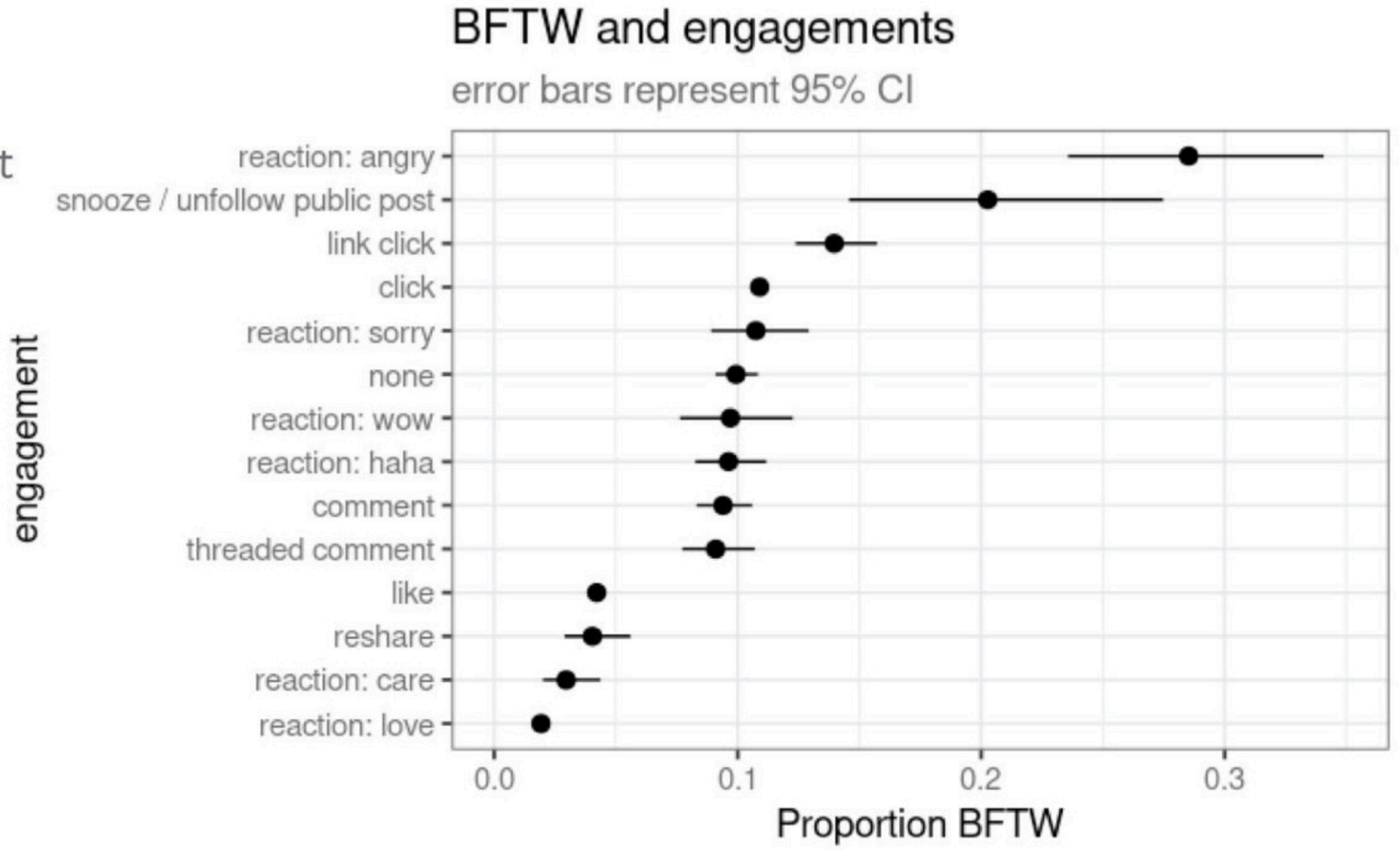
What's more, the **effects of bad experiences can compound**:

- Negative experiences lead users to **derive less value from FB *and* reduce future use** - reducing feelings of fun, happiness, connectedness →
- Repeated exposure to **divisive or depressing** content is exponentially more harmful on the user experience than the harm caused by any individual piece of content & results in **increased negative feelings toward the world & FB**. →

Some implicit signals, such as angry react, are correlated with content rated as a Bad Experience

This plot shows the proportion of respondents' ratings for content that they themselves had engaged with (n = 93,626 posts)

- **Angry** reactions were the engagement most positively associated with BFTW ratings
- **Love** reactions were the engagement most negatively associated with BFTW ratings

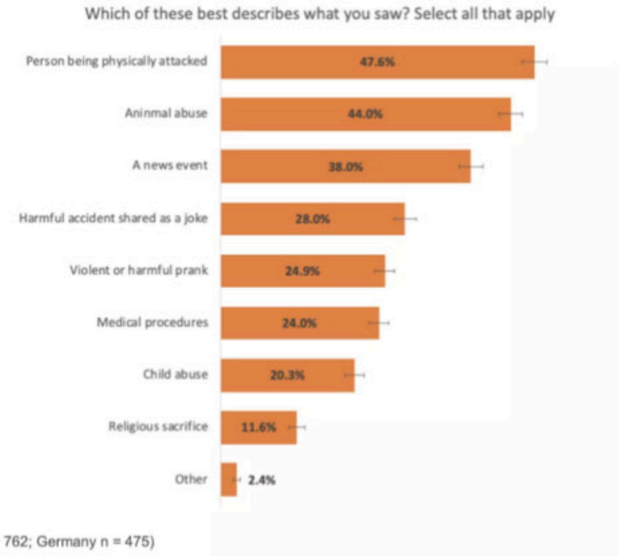


Physical attacks and animal abuse are the most commonly seen types of GV...

Among respondents who saw GV on Facebook, a person being attacked, animal abuse, and a news event were the most prevalent descriptors of what they saw.

The percentages reflect the experiences of respondents who said they saw violent, bloody, or disturbing images on Facebook that bothered them.

These experiences may have occurred in the last 7 days before taking the survey (26.8%) or more than 7 days before taking the survey (32.4%).



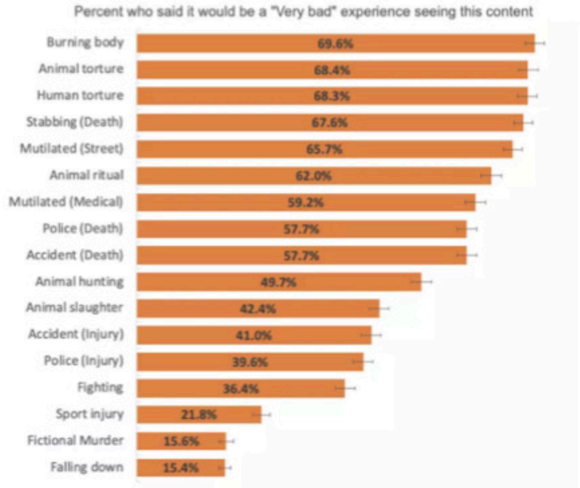
Survey question: Which of these best describes what you saw? Select all that apply.
Respondents: Total n = 2963 (USA n = 438; Brazil n = 579; Mexico n = 709; India n = 762; Germany n = 475)

Imagery of burning bodies, animal torture, and human torture were among the most severe types of GV.

Less severe content types included imagery of graphic sport injuries and fictional depictions of murder.

Overall, respondents anticipating that seeing these content types would be a bad experience is aligned with previous research on severity measurement (e.g., Powell, 2020).

Furthermore, variance in perceptions of content severity within the problem area of GV is consistent with previous research (Major, 2020) demonstrating that some encounters with GV are likely to be worse than others.



Survey question: How bad would this experience make you feel if you saw it on Facebook?
Respondents: Total n = 5004 (USA n = 1001; Brazil n = 1001; Mexico n = 1001; India n = 1000; Germany n = 1004)
See appendix for country-specific results.

- Attitudes toward violent and graphic content, Brandon Carlisle Dec 2020

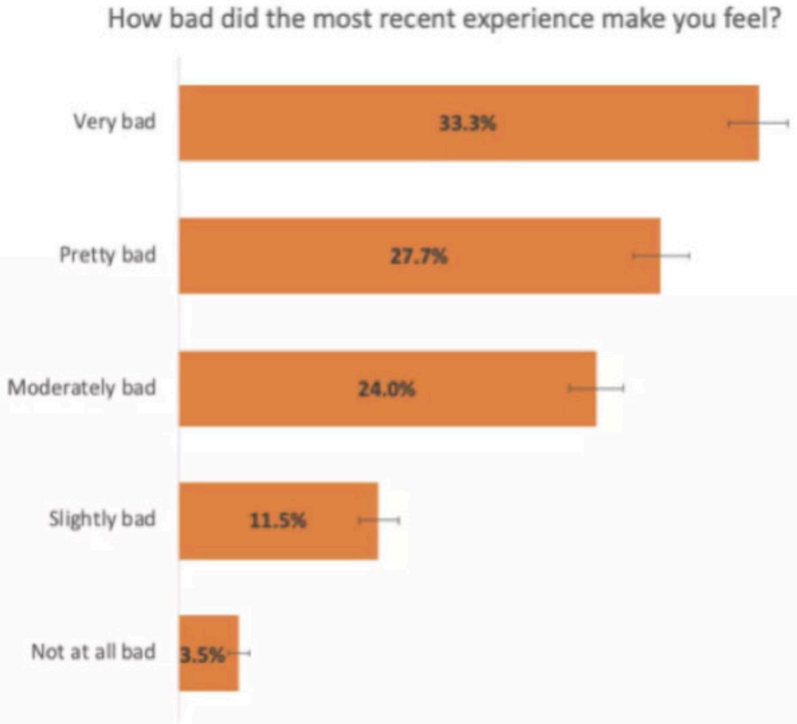
...and seeing graphic violence is very upsetting

Seeing GV can elicit bad feelings that are high in intensity and which can linger **from minutes to days**.

More than 60% of respondents indicated their most recent experience seeing GV made them feel “pretty bad” to “very bad”.

A high percentage of responses indicating that their most recent experience elicited negative feelings is consistent with past work demonstrating that seeing GV is upsetting and can result in feelings of anger or fear (Ritter, 2015).

When asked about what actions they may have taken after seeing GV in their most recent experience, the most common actions were to keep scrolling onto the next post (38.1%) followed by hiding the post (28%), and reporting the post (27%).



Survey question: How bad did this most recent experience make you feel?
Respondents: Total n = 2963 (USA n = 438; Brazil n = 579; Mexico n = 709; India n = 762; Germany n = 475)

- Attitudes toward violent and graphic content, Brandon Carlisle Dec 2020

Building a framework for well-being issues

Overview

One goal of the Signals and Insights team is to build a framework to better understand negative experiences on Instagram. This framework could then be used by teams to prioritize issues to work on, or serve as starting points for their own investigations into particular issues.

Don't we have this already?

Instagram already has many methods of collecting feedback from users. In addition to TRIPS, we have our IG sentiment tracking surveys, LORAX (intended to track the prevalence of a set of issues on FB and IG), team-specific surveys, and open-end responses from those surveys and sources monitored by Product Operations (flytrap, app store reviews, etc). But, there's no holistic monitoring/collection of these issues that would enable us to track negative experiences overall, or flag when new issues are emerging/should be added to existing tracking surveys. There also isn't an existing framework that tries to group issues with each other based on common contextual features, such as surface, frequency, intensity, and so on.

Aren't we measuring all the biggest issues already in TRIPS?

The current TRIPS survey measures many negative experiences that can occur on IG, but it's not intended to be exhaustive. Many of the questions are based on policy-violating issues, were added as a result of world events (such as COVID) or based on what different teams were working on. It also doesn't ask questions about some high severity/low prevalence issues (e.g., terrorism) because of the how people would react to seeing that in a survey and how little signal we'd get, and some low severity/high prevalence issues (e.g., spam). As a result, it's possible that there are issues on IG that cause negative reactions for people that aren't being monitored, or emerging issues that we don't notice right away.

Project plan

The framework will be developed in two phases. The first phase will involve combing through existing research and data signals to identify the most important user issues and most illuminative contextual factors, and create a draft framework to understand/visualize the full scope of negative user experiences (which we expect to include both TRIPS and non-TRIPS issues). This effort will take place in H1.

The second phase will be to add questions related to findings from phase one into the Ad Hoc TRIPS survey, which will run at the end of the half. That survey will ask about the negative experiences we identify (in addition to a set of TRIPS questions), along with contextual questions that add clarity and help in mapping these issues. The analysis portion of this won't happen to H2, so we expect to revise our initial framework then. The rest of this doc focuses on the first H1 work.

Aggregate issues and contextual dimensions

Since all these data sources are different, we won't be doing a quantitative meta-analysis. But we will do a qualitative summary, and conduct our own primary analysis (such as text analysis) for raw data sets. An example outcome is something like this spreadsheet shell, where we identify prominent issues that are or aren't in TRIPS already, and estimate the contextual factors that pertain to or help differentiate that issue.

For issues, we'll take a broad approach to what would be considered 'bad', including both high/low severity and high/low frequency. For contextual factors, we'll choose factors that help differentiate negative experiences from each other, and/or map to how we currently think about/try to tackle these issues internally at Instagram.

There are some contextual factors that we know already have been useful to teams and/or used in previous frameworks, or have been suggested as potentially helpful:

- Wanted by both sides/unwanted to one side
- Known individuals vs strangers
- Know them in real life vs just online
- Negative vs positive
- Self vs others
- Intensity (i.e., 'how did this make you feel' from TRIPS)
- Public vs private interaction
- Frequency
- Surface
- Perceived support
- Perceived resolution
- Length of emotional reaction--Brandon Carlisle done some work re: hate, animal abuse. Also could be something you think isn't bad right away, but then gets worse as you think of it. Does this still bother you after you left IG. Did you talk to anyone about this in real life
- Emotion felt (someone mentioned personas have been developed for this?)
- Action taken
  - On and off app
  - Create new account to start over
- Do they think it's our responsibility or not to fix
- Whether the bad experience was prompted by IG (maybe a notification that asks them to follow someone that is deceased, or OTD that shows a negative experience)
- Whether or not well-being concerns stop people from posting

Factors we can pull from log data:

- Age (13-15, 16-17, 18…)
- Creators
- Country

Timeline

Milestone	Due
Identify data sources, POCs	March 25
Collect most relevant data sets/existing reports	March 29
Identify relevant issues, and add to spreadsheet	April 5
Dedupe and prioritize issues, identify dimensions	April 10
Draft framework completed	April 15

Bad Experiences and Encounters Framework (BEEF) Survey

For BEEF v2, set to launch Jan 5, 2022, see here.

Research review
Survey review, holdout condition
Survey review, production condition

Goal

Similar to FB's SHIELD survey, the IG Well-Being Foundations team will be developing an ad-hoc version of TRIPS that will be delivered at the end of each half to people in prod and a well-being holdout. This will give us a high-level, overall view of how we've influenced perceptions of bad experiences through all of our well-being efforts. It will also give us an opportunity to test questions about new issues that are not policy-violating but were prominent in the Bad Experiences and Encounters Framework, and ask contextual questions related to each issue so we can build a deeper understanding of how these issues unfold for our users.

Background

How is this different from TRIPS?

TRIPS is a daily tracking survey that gives us a longitudinal picture of how our users view integrity issues on Instagram over time. It can't tell us the impact of a specific well-being intervention or set of interventions, though, because there is no control group to compare it to. This effort seeks to remedy that, by creating a new experiment at the beginning of each half: one condition will only receive our minimum well-being product experience, and the second condition will receive all of our well-being product changes. At the end of the half, we will survey users in both groups so we can see whether or not our well-being efforts are causally responsible for changes in how people perceive bad experiences on IG.

Will this be identical to AXIS/SHIELD?

It will have a similar aim as SHIELD, but will include different questions. Both the FB and IG teams will choose questions from TRIPS that best capture the issues their well-being teams are working on this half. For example, in H2 2020, AXIS included 20 out of 56 TRIPS issue questions. (The team is currently redesigning it for 2021; roadmap is here).

What else will this survey data be used for?

Another goal of the Signals and Insights team is to build additional indicators of well-being sentiment, because it's not always feasible to run well-being surveys for every experiment. These indicators could be either behavioral correlates of well-being survey responses that are worth reducing themselves (e.g., blocking another user), or they could be ML models designed to serve as proxies of the survey data itself. This effort will be led by data science, and can be seen in more detail in the Bad Experience Measurement product brief. The survey data will serve as the ground truth that the behavioral correlates and proxies will be tested against (in addition to existing TRIPS data).

In addition, the contextual data collected for each issue will be instrumental in developing principled approaches to grouping bad experiences into larger subcategories. This effort will aid in understanding/roadmapping, as well as give aggregate categories for DS to model.

Survey audience

The survey audience will be people randomly sampled from the two conditions in the well-being holdout. This is the query for the prod group, along with iData info for the custom cluster table; this is the query for the holdout group, and the iData info. Variables in each table: igid, experiment, qe_exposure_date, condition. Each table has 25M IDs.

Weighting

Sample will be weighted to be representative of the IG population using Graviton, an internal Python package that implements inverse propensity weighting, with the help of the survey infra team (need to confirm support).

Sample size

Sample size calculations and survey questions can be found here.

Sample sizes were calculated in two ways--to detect a 5% relative difference between test and control (based on prevalence estimates from past research), and to detect a 1% absolute difference between test and control. A two-tail z test for proportions was used, with 5% alpha, 80% beta (G*Power 3.1). Relative difference is the way FB's SHIELD survey is calculating power (they have been allocated 750k starts). Issues not being actively worked on by IG teams will only be asked in the test condition, not control, to cut down on sample size. I'm proposing we go with 1% absolute difference, because it reduces sample size significantly, and leads to a more balanced number for each individual question.

The ideal scenario would be to ask one issue plus follow-ups per person, but that would require a sample size of 916k(!). In survey review, we agreed to ask five issues per person, and then ask the follow-up questions about one of those issues. That leads to a total sample of 183k (238k 30% dropoff).

For respondents who answer 'no' for all five issue questions, they'll be asked a series of questions about positive experiences on IG (specifically, questions about how IG plays a role in off-line activity). This will be preliminary data to inform 'good experiences' projects next half.

		1% absolute change				
Issue	Baseline Prevalence	Detectable drop	n per group	# of groups	subtotal	Ratio
Hate Witness	19.73%	18.73%	24,381	2	48,762	5
Audience limitation	33.70%	32.70%	34,813	2	69,626	7
False or Misleading	26.68%	25.68%	30,337	2	60,674	6
Usability / action-oriented	21.30%	20.30%	25,859	2	51,718	5
B&H Target	7.80%	6.80%	10,622	2	21,244	2
B&H Witness	22.43%	21.43%	26,875	2	53,750	5
Graphic Violence	11.30%	10.30%	15,122	2	30,244	3
Nudity	13.46%	12.46%	17,707	2	35,414	4
Drugs and Related Goods	3.64%	2.64%	4,774	2	9,548	1
Over/under enforcement	30.00%	29.00%	32,647	2	65,294	7
Transparency	30.00%	29.00%	32,647	2	65,294	7
Usability / consumption-oriented	26.20%	25.20%	29,974	2	59,948	6
Negative Social Comparison	20.00%	19.00%	24,641	2	49,282	5
Data privacy	31.90%	30.90%	33,813	2	67,626	7
Protect minors / solicitation	30.00%	29.00%	32,647	2	65,294	7
Perceived control / sense of place	23.10%	22.10%	27,458	1	27,458	3
Account security (access)	20.00%	19.00%	24,641	1	24,641	2
Civic content (too much political content)	28.20%	27.20%	31,437	1	31,437	3
Commerciality	21.50%	20.50%	26,042	1	26,042	3
Impersonation (first person)	3.10%	2.10%	301	1	301	0
Self-harm (witness)	10.00%	9.00%	13,495	1	13,495	1
Spam (fake account)	50.00%	49.00%	39,240	1	39,240	4
				question n	916,332	92
				# users	305,444	31

Survey instrument

The issues asked about in the survey will mirror the focus areas of the various IG well-being teams plus the issues uncovered in the Bad Experiences and Encounters Framework. The second table below has the contextual dimensions that will be asked about for each issue: the light green rows will be asked in the survey itself, while the dark green rows will be appended from log data.

Contextual questions	Category	Relevant teams		Survey or log data
Surface	Experience	F&R Integrity		Survey
Frequency	Experience			Survey
Specific emotion felt	Experience			Survey
Length of emotional reaction	Experience			Survey
Perceived support	Aftermath	Support	Creator well-being	Survey
Action taken (both on and off app)	Aftermath			Survey
Stops posting?	Aftermath	Creator well-being		Survey
Known person, online/offline	Relationships			Survey
Age (13-15, 16-17, 18...)	Demographics impacted	Teens		Survey
Creators	Demographics impacted	Creator well-being		Log data
Country	Demographics impacted			Log data
App use frequency	Demographics impacted			Log data
Zip code: predominantly black?	Demographics impacted	Equity		Log data
gender	Demographics impacted	Equity		Log data
operating system	Demographics impacted	Equity		Log data
RAM class	Demographics impacted	Equity		Log data
Resulting from product decision?	Experience			Log data

Timeline

Date	To do	Status
April 12	Agree on which TRIPS questions to include	Complete
April 19	Use framework project results to generate proposal for additional survey constructs	Complete
April 26	Develop survey questions/ look for questions from existing surveys	Complete
May 3	Develop survey questions/ look for questions from existing surveys	Complete
May 10	Submit to research review	Complete (June 3)
~~May 24~~ June 3	Submit to survey review	Complete (June 9)
~~May 31~~ June 10	Submit to translations	Complete (June 29)
~~June 14~~ June 30	Launch survey	Complete (June 30)
~~June 21~~ July 6	Data cleaning	Complete
~~June 28~~ August 25	Data weighting/ appending	Complete (August 25)
~~July 19~~ August 26	Data cleaning	Complete (September 6)
~~July 26~~ September 7	Analysis/ reporting	In progress
October 4	Socialization	
October 11	v2 Framework	
October 18	Begin BEEF Survey v2 development	
December 16	Launch BEEF Survey v2 (code freeze Dec 16-Jan 4)	

NOTE: all analyses in this deck should be considered preliminary.

Bad Experiences and Encounters Framework (BEEF) Survey

Data collection: June 27–July 8, 2021

1

Background

2

Why are we doing this?

To develop a holistic, consistent picture of user bad experiences on Instagram that allows us to track our progress each half.

3

Survey timeline

What's next?

4

Survey questions, issues

5

Survey questions and log data, context

6

Sample size by issue

7

Results

8

Issue rates, full dataset

Improvements in perceived violence and usability action, but a decline in perceived data privacy

9

Issue rates and frequencies in last 7 days

Commercially and political posts have the highest frequencies among respondents who have experienced the issue, while fake accounts and impersonation have the lowest.

10

Issue rates by surface

DM/chat has the two highest rates of issues across surfaces: 73.7% of the time for fake account contact, and 68.6% of the time for unwanted sexual advances.

11

Relationship between respondent and antagonist

Issues are largely caused by strangers, with negative social comparison (15.5%) and bullying target (14.6%) having the highest rates of known instigators.

12

Relationship with antagonist by creator status

Low response rates among creators prevent strong conclusions

13

Actions taken as a result of experiencing an issue

Blocking or muting the account is the most frequent action taken as a result of an issue experience, with closing Instagram as the second most frequent.

14

Impact of issue experience on creation behavior

All issues cause some reduction in self-reported creation behavior, with impersonation hindering creation the most.

15

Results by age and gender

16

Respondents experiencing one or more issues

A little over half of respondents (51.6%) experienced at least one issue in the past 7 days

17

Issues by most and least impacted age groups

The 13-15 age group has the highest rates for 8/22 issues, while the 45+ group has the lowest rates of 21/22 issues

18

Issues by all age groups

Younger people report higher rates of every issue, with some issues felt more universally than others

19

Issue rates by age group and gender*

Males and females experience issues at significantly different rates, and those patterns remain consistent across age groups.

20

Bullying and known antagonists, by gender

Females respondents are more likely than males to know the antagonist offline when the target of bullying, but less likely to know the antagonist offline when witnessing bullying.

21

Results by creator segmentation

22

Creator segmentation

5.41% of survey respondents are classified as creators, based on ig_creator_segment.

23

Issues by creator segmentation

Most issues are experienced more often by larger creators, but the gap varies by issue

24

25

From: Arturo [redacted]

Subject: Gap in our understanding of harm and bad experiences

Date: October 5, 2021 at 9:37:59 PM PDT

To: Mark Zuckerberg [redacted]

Cc: Sheryl Sandberg [redacted], Chris Cox [redacted], Adam Mosseri [redacted], Mark Zuckerberg [redacted]

Dear Mark,

I saw the note you shared today after the testimony, and I wanted to bring to your attention what I believe is a critical gap in how we as a company approach harm, and how the people we serve experience it. I've raised this to Chris, Sheryl, and Adam in the last couple of weeks.

I want to start by saying that my personal experience, and what I believe, is that you and m-team care deeply about everyone we serve, and my goal in sending this is to be of service to that. It's been 2 years since I've been back part-time.

51% of Instagram users say 'yes' to having had a bad or harmful experience in the last 7 days. Out of those 1% of report and of those 2% have the content taken down (i.e. 0.02%). The numbers are probably similar on Facebook.

Two weeks ago my daughter [redacted], 16, and an experimenting creator on Instagram, made a post about cars, and someone commented 'Get back to the kitchen.' It was deeply upsetting to her. At the same time the comment is far from being policy violating, and our tools of blocking or deleting mean that this person will go to other profiles and continue to spread misogyny. I don't think policy/reporting or having more content review are the solutions.

There is detailed data about what people experience in TRIPS, a statistically significant survey. We ran a more detailed survey, I've attached the full age breakdown below, but here are some key numbers (these questions are in the last 7 days):

21.8% of 13-15 year olds said they were the target of bullying.
39.4% of 13-15 year olds said they experienced negative comparison.
24.4% of 13-15 year old responded said they received unwanted advances.

Why does someone think it is ok to post 'get back to the kitchen' or harass someone? I believe it is because it doesn't violate policy, and other than deleting or blocking, there is no feature that helps people know that kind of behavior is not ok. Another example, is unsolicited penis pictures. [redacted] has received those from boys too since the age of 14, and her tool is to block them. I asked her why boys keep doing that? She said if the only thing that happens is they get blocked, why wouldn't they?

Why the gap between Prevalence and TRIPS? Today we don't don't know what % of content people experience as misinformation, harassment, or racism is policy violating. We have done great work in driving down prevalence, and there will always be more to do, but what if policy based solutions only cover a single digit percentage of what is harming people?

Policy is necessary when the content is unambiguously inappropriate, yet it has many limitations. It trails behavior, the interventions are heavy and risk over-enforcement and getting the border line right is extraordinarily difficult. Policy enforcement is analogous to the police, it is necessary to prevent crime, but it is not what makes a space feel safe.

What makes a workplace, or a school, or a dinner table feel safe is social norms.

If someone goes around telling women to 'get back to the kitchen', and the only thing that happens is their content is deleted or they get blocked, don't we run the risk of normalizing bad behavior? How are people to learn to be members of a safe and supportive community without visible interventions that help set the social norms for the environment? I believe social norms also protect speech.

At dinner tonight [redacted] said: my car videos are getting 100,000 views, it's natural that I'm going to get a lot of hate with that. Is it? Why is it acceptable for someone to harass someone on their surface? The most powerful solution for the integrity and safety space is to affect the supply of bad experiences via the actors creating them.

I might be wrong about my assessment, and welcome feedback about any effort or data that I'm missing. I believe that it is important to get the following efforts well-funded and prioritized:

- What is the content that is causing bad experiences for our users? How intense is the experience?
- What % of that content is policy violating? (i.e. how much of TRIPS is driven by content other than what drives Prevalence?)
- What are visible product solutions that make the community better over time? e.g. actor feedback, comment covers, pinned comments, etc.

The solutions we create I believe should have the following properties:

- The person who has the negative experience should feel heard, you don't 'perceive' racism or harassment, you experience it, and you are the source of truth for that. The feedback flow should not be just about filing a report, but about understanding the experience the person is having so we can give them the right solution.
- We should empower creators, communities, and Instagram, in setting the social norms for the spaces the are a part of.
- Where appropriate we should give feedback to actors, in the belief that they are acting with good intention and might have caused unintentional harm. There can be a range of interventions that start with 'nudges' that assume positive intention. This will allow us to separate the people who would behave differently given feedback, from the ones who are intentionally causing harm. We can then approach people who are intentionally malicious with the integrity tools.

If you would like I can give more details or specifics on this. I am appealing to you because I believe that working this way will require a culture shift. I know that everyone in m-team team deeply cares about the people we serve, and the communities we are trying to nurture, and I believe that this work will be of service to that.

Arturo

		Overall rank	Overall %	13-15	16-17	18-21	22-26	27-34	35-44	45+
	[illegible]			Column N %	Column N %	Column N %	Column N %	Column N %	Column N %	Column N %
commerciality	Yes, last 7 days	1	48.2%	59.20%	63.20%	69.70%	74.70%	77.80%	76.20%	80.20%
usability passive	Yes, last 7 days	5	25.5%	51.60%	49.30%	48.50%	44.20%	41.20%	38.30%	31.90%
audience limitation	Yes, last 7 days	4	26.8%	50.70%	48.80%	47.50%	46.40%	47.30%	44.20%	46.90%
bully witness	Yes, last 7 days	3	28.3%	48.70%	50.00%	51.10%	53.90%	51.90%	45.40%	35.10%
hate witness	Yes, last 7 days	6	25.3%	46.10%	47.50%	47.70%	48.40%	43.50%	36.40%	29.20%
negative comparison	Yes, last 7 days	10	19.2%	39.40%	35.30%	35.70%	35.00%	34.80%	31.00%	23.10%
nudity	Yes, last 7 days	13	16.3%	36.70%	32.10%	32.00%	30.70%	26.70%	26.00%	20.30%
perceived control	Yes, last 7 days	8	23.9%	33.40%	34.70%	43.00%	46.40%	46.30%	44.30%	42.90%
data privacy	Yes, last 7 days	7	24.4%	32.70%	36.70%	40.40%	46.10%	47.80%	44.20%	45.00%
fake acct 1st	Yes, last 7 days	9	21.8%	29.20%	36.60%	42.80%	40.10%	42.50%	38.00%	39.30%
transparency	Yes, last 7 days	11	17.5%	29.20%	33.80%	34.00%	32.20%	30.50%	28.50%	28.00%
misinfo	Yes, last 7 days	2	40.1%	27.90%	24.00%	23.10%	19.90%	20.80%	24.40%	31.40%
over enforcement	Yes, last 7 days	14	14.8%	26.50%	29.70%	29.80%	27.30%	26.60%	22.20%	20.60%
political posts	Yes, last 7 days	12	17.0%	25.80%	28.70%	30.00%	32.30%	29.00%	27.00%	27.90%
violence	Yes, last 7 days	15	12.8%	24.40%	25.60%	23.80%	24.70%	23.60%	19.80%	17.80%
unwanted advances	Yes, last 7 days	16	11.9%	24.40%	25.40%	26.10%	20.80%	17.30%	18.60%	23.60%
usability action	Yes, last 7 days	17	10.7%	22.00%	20.20%	20.30%	19.60%	20.00%	20.70%	21.30%
bully target	Yes, last 7 days	18	8.1%	21.80%	18.90%	15.70%	14.40%	14.90%	12.60%	12.40%
self harm	Yes, last 7 days	19	6.7%	16.90%	12.90%	13.80%	10.80%	7.20%	6.30%	6.90%
impersonation 1st	Yes, last 7 days	22	3.7%	11.60%	6.70%	9.90%	6.10%	2.50%	1.80%	1.20%
acct security	Yes, last 7 days	20	3.9%	9.70%	6.00%	7.70%	6.00%	4.30%	3.90%	4.10%
drugs	Yes, last 7 days	21	3.9%	7.10%	6.70%	7.30%	5.40%	8.00%	6.30%	6.80%

Message

From: Arturo Bejar [redacted]
Sent: 10/14/2021 11:56:05 PM
To: [redacted]
Subject: Fwd: Pre-read for our conversation tomorrow

Hi [redacted]

Sharing with you the pre-read of my conversation with Adam tomorrow, I will keep you posted.

Arturo

Begin forwarded message:

From: Arturo Bejar [redacted]
Subject: Pre-read for our conversation tomorrow
Date: October 14, 2021 at 4:40:38 PM PDT
To: Adam Mosseri [redacted]

Hi Adam,

In order to make the best use of our time tomorrow I put together a short pre-read that I've vetted with the team in well-being.

Data points (last 7 days/more than 7 days)

Have you ever received unwanted sexual advances on Instagram?

- 13-15 year olds: 13%/27%

Have you ever seen anyone discriminating against people on Instagram because their gender, religion, race, sexual orientation, or another part of their identity?

- 13-15 year olds: 26%/31%

Has anyone done any of these things to you on Instagram? Insulted or disrespected you, contacted you in an in appropriate way, damaged your reputation, threatened you, excluded you or left you out.

- 13-15 year olds: 11%/25%

Have you ever felt worse about yourself because of other peoples' posts on Instagram?

- 13-15 year olds: 21%/23%

Questions:

- What should be the goal/number of 13-15 year olds on each of the BEEF categories?

- Are users able to express these experiences to us in the product? (e.g. for unwanted sexual advances, or negative comparison you can't)
- What would we build if >90% of the content which drives these experiences is not policy violating or borderline?

If you'd like to look at the data directly, here is the data by age:
https://docs.google.com/spreadsheets/d/10rR5hbK4v1W-2QmUUMSLGiUFfljgm9ngf-[redacted]

I also find it helpful to put the data in the context of the questions (which convey better than our labels what people are experiencing):
https://docs.google.com/spreadsheets/d/1gOGpXK7UkC_Z7_C41Z8SLqb4Puom4vv4KfX0Zs6-[redacted]

Recommendations

1. For Instagram to set goals based on TRIPS/BEEF, use people's experience as the north star for the work:
    1. What would you build if the goal was to get to 1% unwanted sexual advances? Or 3% witnessed hate? Or 2% target of bullying?
    2. Change the use of the word 'perceived' to 'experienced' - people don't perceive being harassed.
2. Provide features that help us understand the issues and content that people are experiencing so that we may develop interventions/features that help them and improve the community over time.
    1. Secondary actions to block/delete where we get user experience data. This has been difficult to date because the team has been running into XFN limitations on understand efforts.
    2. Make the reporting flow, or add experiences at the beginning to make people feel heard and supported with what they are experiencing, as well as generate insights on the issues they are having.
3. Invest in features that help us learn how to develop and maintain social norms, and actor feedback.

Can we shift the conversation into one of hope and leadership?

- Everyone in the industry has the same problems right now.
- Prevalence-based measures, while necessary, don't create a safe and supportive community, you're always behind the latest harmful thing.
- We have few visible features that help create a safe and supportive environment for everyone.
- There is a great product opportunity in figuring out the features that make a community feel safe and supportive.
- It is possible and important to work these issues in partnership with other industry leaders and academics. We have much learn about each of these issues. I believe is possible to help create public conversation on these topics for good.

A point which might be good for you to know (which I did not put in the document reviewed by the team) is that many employees I've spoken who are doing this work (and are of different levels) are distraught about how the last few weeks have unfolded. These are people who love FB/IG, and are heart/mission driven to the work.

internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/

Wiki

Search TRIPS (Tracking Reach of Integrity Problems Survey)

+ New Page | My User Page | My Activity

Central Integrity Research

- TRIPS (Tracking Reach of Integrity ...
  - Survey Methodology
  - Show Me the Data!
  - Experimenting with TRIPS
  - So You Want to Goal on TRIPS
  - Frequently Asked Questions
  - C-TRIPS (Comparative TRIPS)
  - Contact the TRIPS Team
  - What does Onboarding to TRIPS Lo...

DOCUMENT INFO

Last Update: about a month ago

Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Brendan Read, Jess Bodford, Brett Major, Joe Tullio, Frank Kanayet, Umer Farooq, Zachary Bodnar

Viewers: 217 in past month

This page is considered up-to-date

Wiki > Central Integrity Research > TRIPS (Tracking Reach of Integrity Problems Survey)

Edit Page

TRIPS (Tracking Reach of Integrity Problems Survey)

What is TRIPS?

The "Tracking Reach of Integrity Problems Survey" (TRIPS), also known as the Perception Framework, systematically measures global perceptions of Integrity problems on Facebook, Instagram, and Messenger (in progress). TRIPS tracks integrity problems relevant to Central Integrity, FB App Integrity, Instagram Well-Being, and Civic Integrity teams. Each survey item has undergone rounds of cognitive testing and statistical validation, ensuring a scalable, reliable way to compare user perceptions across problems, estimating:

1. Perceived reach of a given integrity problem
2. Reputational reach (i.e., for those who have heard of others experiencing this problem)
3. Perceived intensity of experiencing this problem

How can I stay updated about TRIPS?

- Please join the TRIPS FYI and Updates group
- H2 2020 Roadmap
- Want to see the data? Please see Show me the data!

Why should we care about perception?

Sometimes we define Integrity problems differently than our users. The Fact Framework leverages problem team definitions to measure *prevalence* in VPVs. TRIPS, on the other hand, measures person-level *reach* according to user definitions. This prevalence-reach distinction is important because we don't always define Integrity problems in the same way as our users; for example, what is nudity for some may simply be liberal attire for others. Similarly, our measurement systems may miss events that users define as intense, violating, or problematic, such as signals of impersonation or "friendly hacking" between contacts. Although we may have ways to capture badness in our own definitions, decades of research suggest that perceptions of harm are subjective—and that subjectivity is tied to willing use, enjoyment, and feelings of safety on our platforms.

Is this page useful?

To measure intensity, we must rely on user perception. Furthermore, the Fact Framework has no way of m
perceived intensity without relying on logged behaviors, which are noisy and highly variable depending on c
context. By gathering user perceptions of intensity, we can better establish personalized demotions based on user
thresholds for what constitutes a bad experience; these data also inform parallel measurement systems including the

Chrome File Edit View History Bookmarks People Tab Window Help

internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/

Wiki

Search TRIPS (Tracking Reach of Integrity Problems Survey)

+ New Page My User Page My Activity

Central Integrity Research

- TRIPS (Tracking Reach of Integrity ...
  - Survey Methodology
  - Show Me the Data!
  - Experimenting with TRIPS
  - So You Want to Goal on TRIPS
  - Frequently Asked Questions
  - C-TRIPS (Comparative TRIPS)
  - Contact the TRIPS Team
  - What does Onboarding to TRIPS Lo...

DOCUMENT INFO

Last Update: about a month ago

Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Brendan Read, Jess Bodford, Brett Major, Joe Tullio, Frank Kanayet, Umer Farooq, Zachary Bodnar

Viewers: 217 in past month

This page is considered up-to-date

Wiki > Central Integrity Research > TRIPS (Tracking Reach of Integrity Problems Survey)

Edit Page

To measure intensity, we must rely on user perception. Furthermore, the Fact Framework has no way of measuring perceived intensity without relying on logged behaviors, which are noisy and highly variable depending on country and context. By gathering user perceptions of intensity, we can better establish personalized demotions based on user thresholds for what constitutes a bad experience; these data also inform parallel measurement systems including the Severity Framework.

Perception allows us to verify whether product interventions actually work. Integrity teams were founded to identify and minimize harm on our platforms; accordingly, our product teams frequently launch in-product interventions to test whether our solutions are indeed reducing the spread of harm on Facebook and other surfaces. What is more important, however, is ensuring that our users actually perceive a reduction in harm in their day-to-day experience with our products. TRIPS enables product teams to validate experiments and in-product solutions against user perceptions in real-time, ensuring that our Integrity efforts yield noticeably better experiences for the people we serve.

What is TRIPS' mission?

Our mission: Enable the systematic, rigorous, and reliable measurement of people's perceptions of bad experiences on FACEBOOK, tracking integrity problems across the family of apps.

TRIPS was founded as a way to listen to users' voices in real time in the immediate wake of an Integrity experience on our platforms. The data we collect capture the extent to which group-level preferences for content change over time, vary by country or group, and differ from our policy classifications of harm. We aim to empower product teams to monitor and proactively detect sudden changes in perceived reach or intensity, guiding product solutions to minimize exposure to harm on a global scale.

We combine rigorous methods with subject matter expertise. The TRIPS team comprises survey methodologists, social and cognitive psychologists, and quantitative specialists with deep-rooted expertise in the Integrity space. We rigorously develop and test questions to ensure ease of comprehension across 58 locales, as well as ease of comparability across problem types.

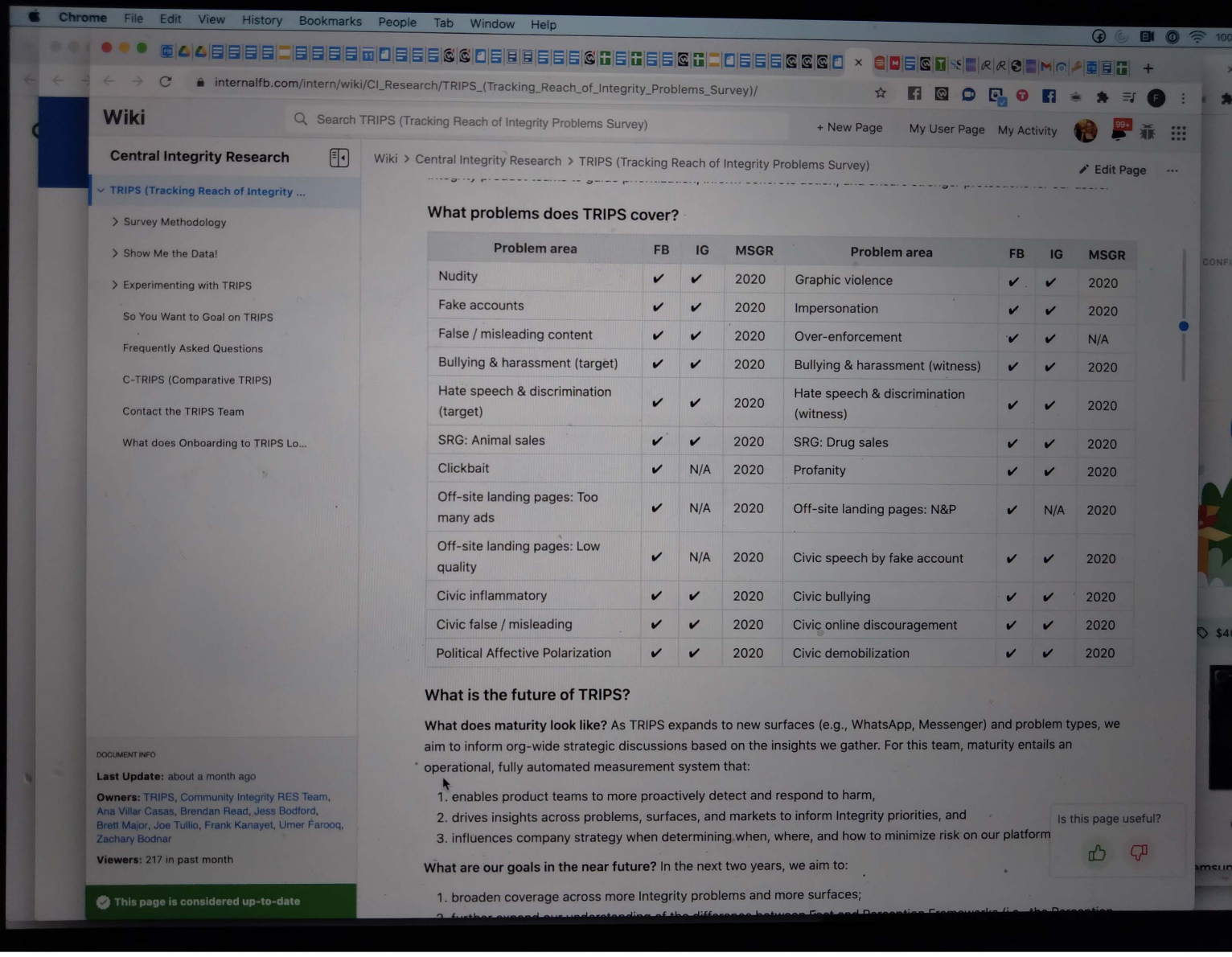
Broadly, we serve as a mouthpiece for the people who have experienced harm on our platforms; by proxy, we echo these voices by surfacing patterns, inconsistencies, and unexpected changes to company leadership, Policy teams, and Integrity product teams to guide prioritization, inform concrete action, and ensure stronger protections for our users.

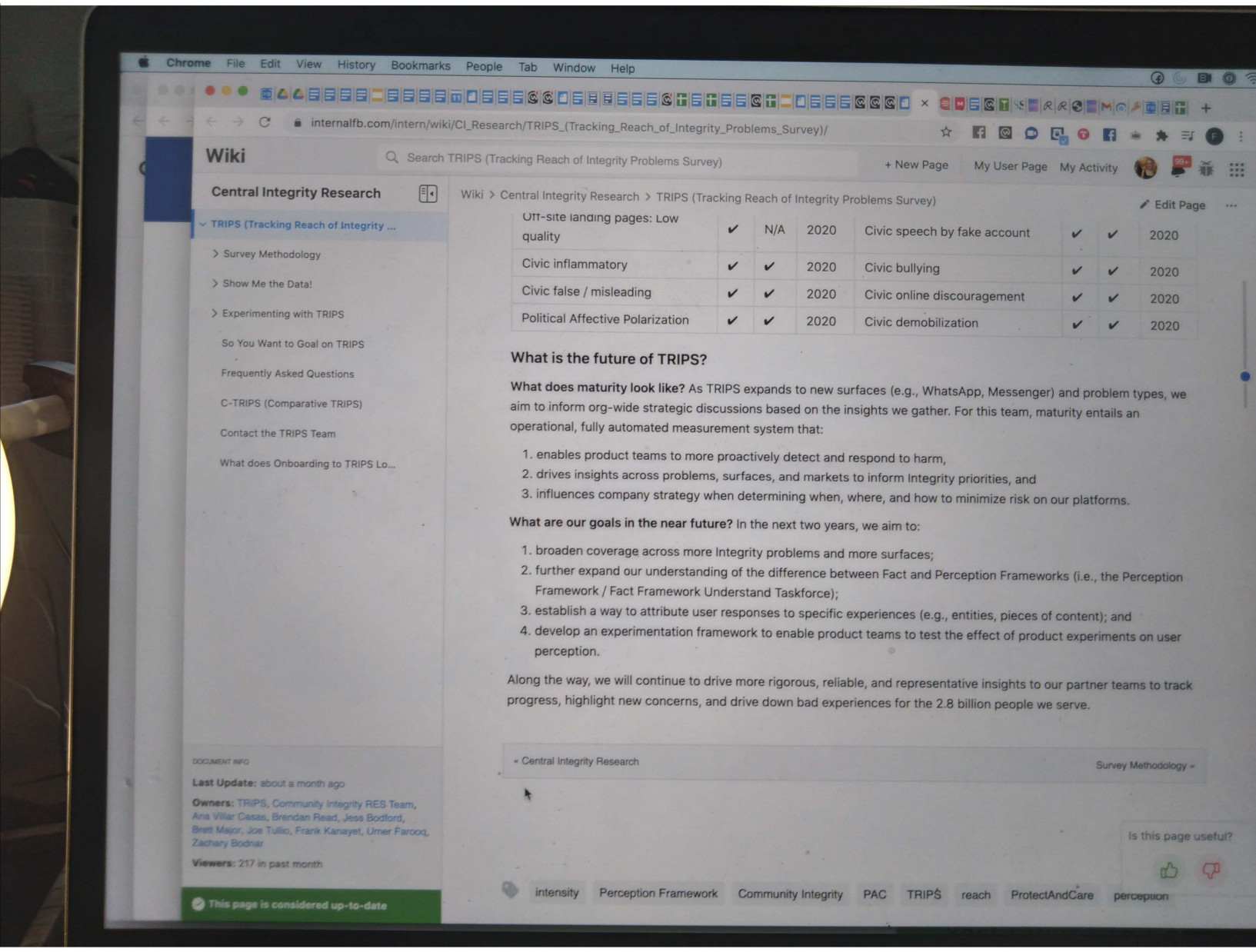
What problems does TRIPS cover?

Is this page useful?

Problem area	FB	IG	MSGR	Problem area	FB	I
Nudity	✔	✔	2020	Graphic violence	✔	✔
Fake accounts	✔	✔	2020	Impersonation	✔	✔

2020


Chrome File Edit View History Bookmarks People Tab Window Help
internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/
Wiki
Search TRIPS (Tracking Reach of Integrity Problems Survey)
+ New Page
My User Page
My Activity
Central Integrity Research
TRIPS (Tracking Reach of Integrity ...
Survey Methodology
Show Me the Data!
Experimenting with TRIPS
So You Want to Goal on TRIPS
Frequently Asked Questions
C-TRIPS (Comparative TRIPS)
Contact the TRIPS Team
What does Onboarding to TRIPS Lo...
Wiki > Central Integrity Research > TRIPS (Tracking Reach of Integrity Problems Survey)
Edit Page
What problems does TRIPS cover?
Problem area | FB | IG | MSGR | Problem area | FB | IG | MSGR
Nudity | ✔ | ✔ | 2020 | Graphic violence | ✔ | ✔ | 2020
Fake accounts | ✔ | ✔ | 2020 | Impersonation | ✔ | ✔ | 2020
False / misleading content | ✔ | ✔ | 2020 | Over-enforcement | ✔ | ✔ | N/A
Bullying & harassment (target) | ✔ | ✔ | 2020 | Bullying & harassment (witness) | ✔ | ✔ | 2020
Hate speech & discrimination (target) | ✔ | ✔ | 2020 | Hate speech & discrimination (witness) | ✔ | ✔ | 2020
SRG: Animal sales | ✔ | ✔ | 2020 | SRG: Drug sales | ✔ | ✔ | 2020
Clickbait | ✔ | N/A | 2020 | Profanity | ✔ | ✔ | 2020
Off-site landing pages: Too many ads | ✔ | N/A | 2020 | Off-site landing pages: N&P | ✔ | N/A | 2020
Off-site landing pages: Low quality | ✔ | N/A | 2020 | Civic speech by fake account | ✔ | ✔ | 2020
Civic inflammatory | ✔ | ✔ | 2020 | Civic bullying | ✔ | ✔ | 2020
Civic false / misleading | ✔ | ✔ | 2020 | Civic online discouragement | ✔ | ✔ | 2020
Political Affective Polarization | ✔ | ✔ | 2020 | Civic demobilization | ✔ | ✔ | 2020
What is the future of TRIPS?
What does maturity look like? As TRIPS expands to new surfaces (e.g., WhatsApp, Messenger) and problem types, we aim to inform org-wide strategic discussions based on the insights we gather. For this team, maturity entails an operational, fully automated measurement system that:
1. enables product teams to more proactively detect and respond to harm,
2. drives insights across problems, surfaces, and markets to inform Integrity priorities, and
3. influences company strategy when determining when, where, and how to minimize risk on our platform
What are our goals in the near future? In the next two years, we aim to:
1. broaden coverage across more Integrity problems and more surfaces;
DOCUMENT INFO
Last Update: about a month ago
Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Brendan Read, Jess Bodford, Brett Major, Joe Tullio, Frank Kanayet, Umer Farooq, Zachary Bodnar
Viewers: 217 in past month
This page is considered up-to-date
Is this page useful?


Chrome File Edit View History Bookmarks People Tab Window Help
internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/
Wiki
Search TRIPS (Tracking Reach of Integrity Problems Survey)
+ New Page
My User Page
My Activity
Central Integrity Research
TRIPS (Tracking Reach of Integrity ...
Survey Methodology
Show Me the Data!
Experimenting with TRIPS
So You Want to Goal on TRIPS
Frequently Asked Questions
C-TRIPS (Comparative TRIPS)
Contact the TRIPS Team
What does Onboarding to TRIPS Lo...
Wiki > Central Integrity Research > TRIPS (Tracking Reach of Integrity Problems Survey)
Edit Page
Off-site landing pages: Low quality ✔ N/A 2020
Civic speech by fake account ✔ ✔ 2020
Civic inflammatory ✔ ✔ 2020
Civic bullying ✔ ✔ 2020
Civic false / misleading ✔ ✔ 2020
Civic online discouragement ✔ ✔ 2020
Political Affective Polarization ✔ ✔ 2020
Civic demobilization ✔ ✔ 2020
What is the future of TRIPS?
What does maturity look like? As TRIPS expands to new surfaces (e.g., WhatsApp, Messenger) and problem types, we aim to inform org-wide strategic discussions based on the insights we gather. For this team, maturity entails an operational, fully automated measurement system that:
1. enables product teams to more proactively detect and respond to harm,
2. drives insights across problems, surfaces, and markets to inform Integrity priorities, and
3. influences company strategy when determining when, where, and how to minimize risk on our platforms.
What are our goals in the near future? In the next two years, we aim to:
1. broaden coverage across more Integrity problems and more surfaces;
2. further expand our understanding of the difference between Fact and Perception Frameworks (i.e., the Perception Framework / Fact Framework Understand Taskforce);
3. establish a way to attribute user responses to specific experiences (e.g., entities, pieces of content); and
4. develop an experimentation framework to enable product teams to test the effect of product experiments on user perception.
Along the way, we will continue to drive more rigorous, reliable, and representative insights to our partner teams to track progress, highlight new concerns, and drive down bad experiences for the 2.8 billion people we serve.
« Central Integrity Research
Survey Methodology »
DOCUMENT INFO
Last Update: about a month ago
Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Brendan Read, Jess Bodford, Brett Major, Joe Tullio, Frank Kanayet, Umer Farooq, Zachary Bodnar
Viewers: 217 in past month
This page is considered up-to-date
Is this page useful?
intensity
Perception Framework
Community Integrity
PAC
TRIPS
reach
ProtectAndCare
perception

Show Me the Data!

Where to find the data

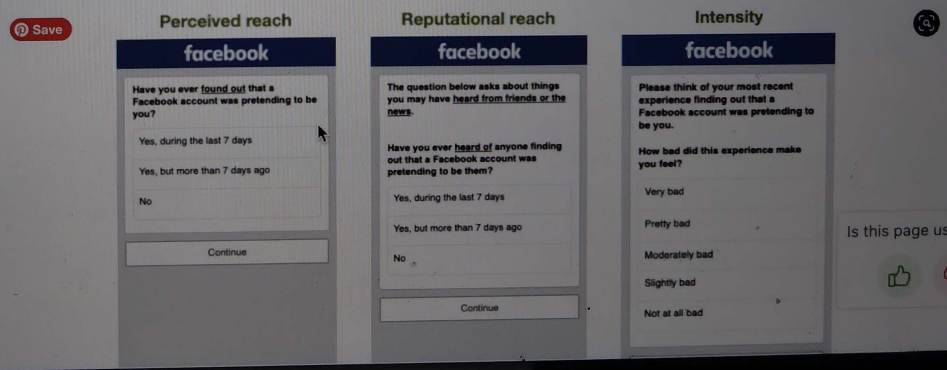
There are two dashboards through which to explore TRIPS data: TRIPS team dashboard and SWAMP.

HOW TO INTERPRET THE VARIABLES

- **Perceived reach:** % of people who have seen/experienced the integrity problem during the last 7 days.
- **Reputational reach:** % of people who heard about the integrity problem during the last 7 days. In other words, it's a measure of reputational hearsay — something heard from friends or in the news but not necessarily something someone saw personally on our platform.
- **Intensity:** % of people who said seeing/experiencing the integrity problem was "pretty bad" or "very bad". The intensity question is only administered to people who said yes to the perceived reach question. The intensity question is not administered as a follow-up to the reputational reach question.

Note: We only report reach estimates for "Yes, during the last 7 days" on the TRIPS team dashboard. In the screenshots below, you may notice users can also answer "Yes, but more than 7 days ago" to the reach questions. We included this response option to reduce over-reporting: if we asked users only about the last 7 days, some were motivated to tell us about bad experiences that happened more than 7 days ago, and thus add error to our metric.

You can find the full verbatim wording of all questions here.



DOCUMENT INFO

Last Update: 3 months ago

Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Jess Bodford, Brett Major, Frank Kanayet

Viewers: 64 in past month

This page is considered up-to-date

internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/Show_me_the_d...

Wiki | Search Show Me the Data! | + New Page | My User Page | My Activity

Central Integrity Research

- TRIPS (Tracking Reach of Integrity ...
  - Survey Methodology
  - Show Me the Data!
    - TRIPS Team Dashboard
    - TRIPS SWAMP Dashboard
    - TRIPS Data Sources
  - Experimenting with TRIPS
  - So You Want to Goal on TRIPS
  - Frequently Asked Questions
  - C-TRIPS (Comparative TRIPS)
  - Contact the TRIPS Team
  - What does Onboarding to TRIPS Lo...

DOCUMENT INFO

Last Update: 3 months ago

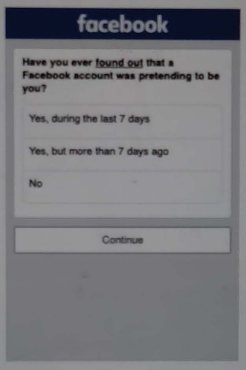
Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Jess Bodford, Brett Major, Frank Kanayet

Viewers: 64 in past month

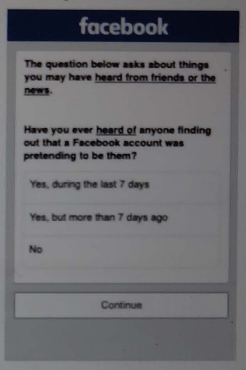
Edit Pa

You can find the full verbatim wording of all questions here.

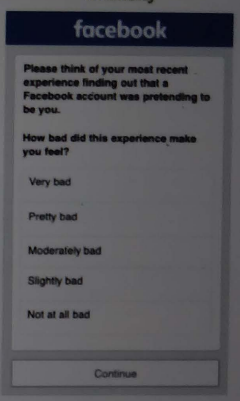
Perceived reach



Reputational reach



Intensity



ACCESSING (AND EXPERIMENTING WITH) TRIPS DATA

1. TRIPS Team Dashboard
2. TRIPS SWAMP Dashboard
3. TRIPS Data Sources
4. Experimenting with TRIPS

INTERESTED IN GOALING ON TRIPS?

Please see So You Want to Goal on TRIPS.

Is this page use

internalfb.com/intern/wiki/CI_Research/TRIPS_(Tracking_Reach_of_Integrity_Problems_Survey)/Show_me_the_d...

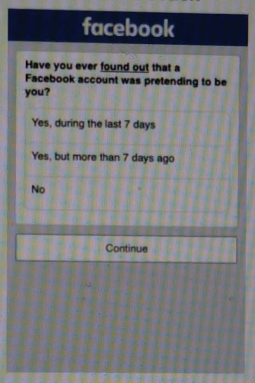
Wiki

Search Show Me the Data!

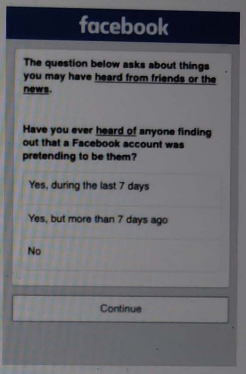
+ New Page | My User Page | My Activity

You can find the full verbatim wording of all questions here.

Perceived reach



Reputational reach



Intensity

ACCESSING (AND EXPERIMENTING WITH) TRIPS DATA

1. TRIPS Team Dashboard
2. TRIPS SWAMP Dashboard
3. TRIPS Data Sources
4. Experimenting with TRIPS

INTERESTED IN GOALING ON TRIPS?

Please see So You Want to Goal on TRIPS.

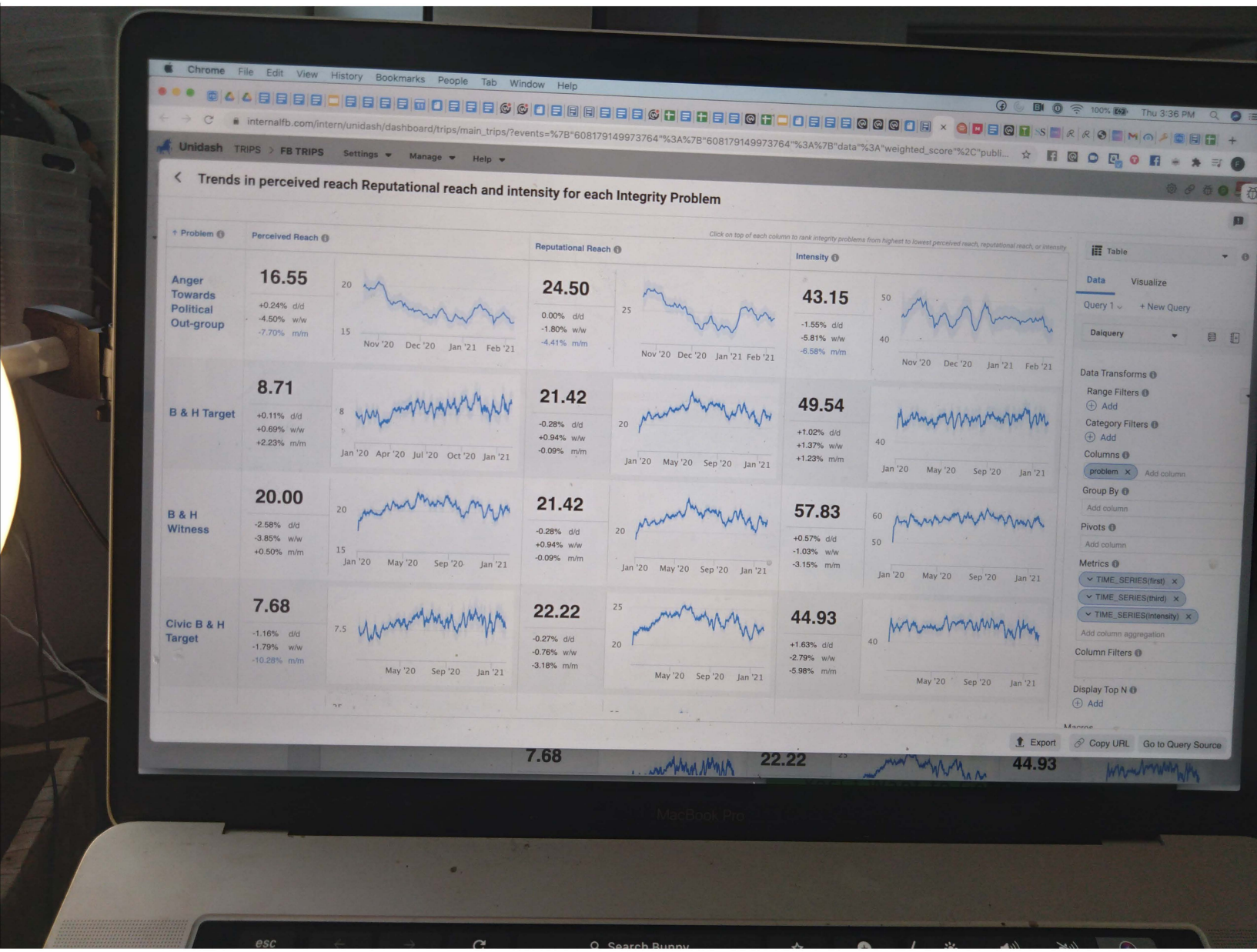
DOCUMENT INFO

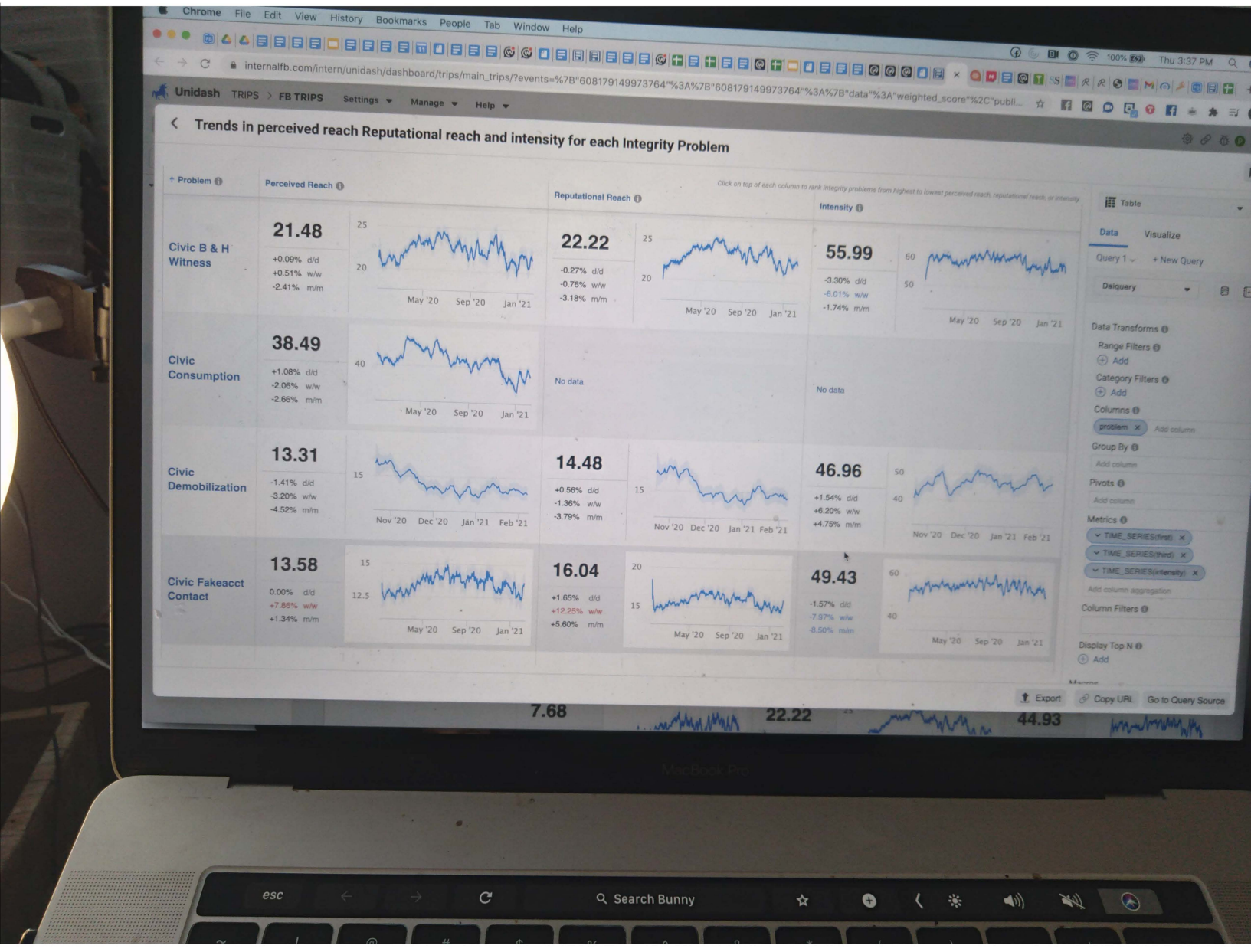
Last Update: 3 months ago

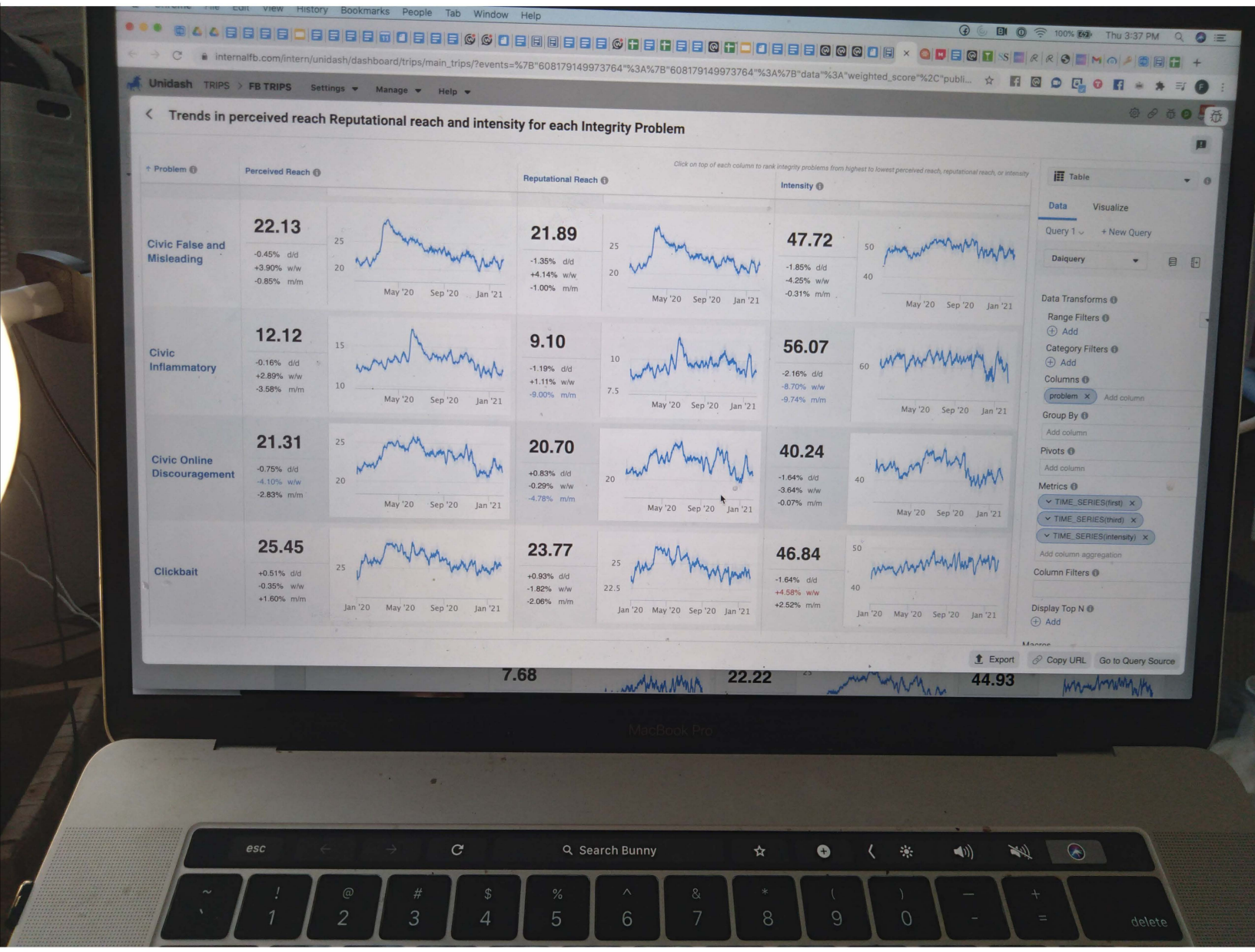
Owners: TRIPS, Community Integrity RES Team, Ana Villar Casas, Jess Bodford, Brett Major, Frank Kanayet

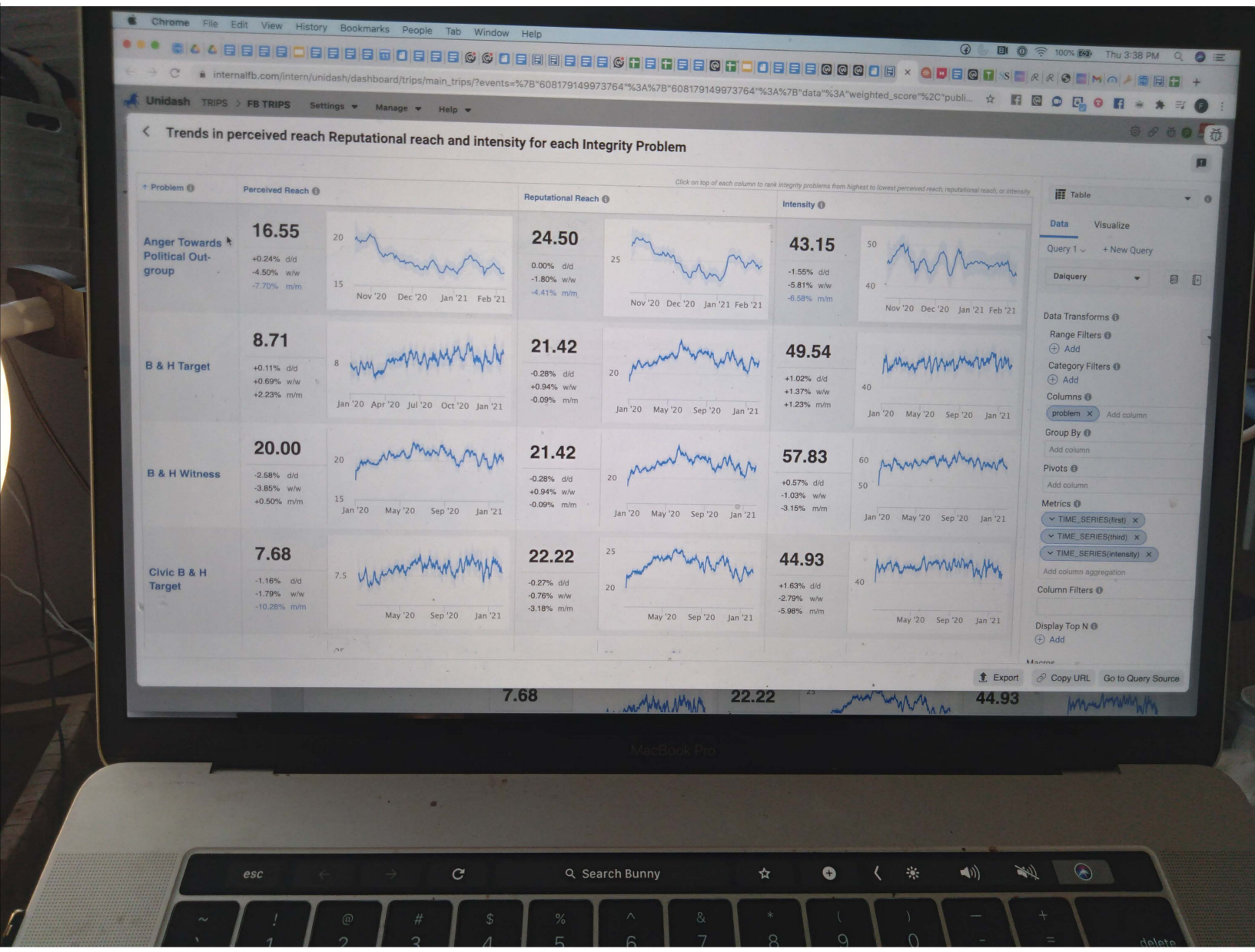
Viewers: 64 in past month

Is this page useful

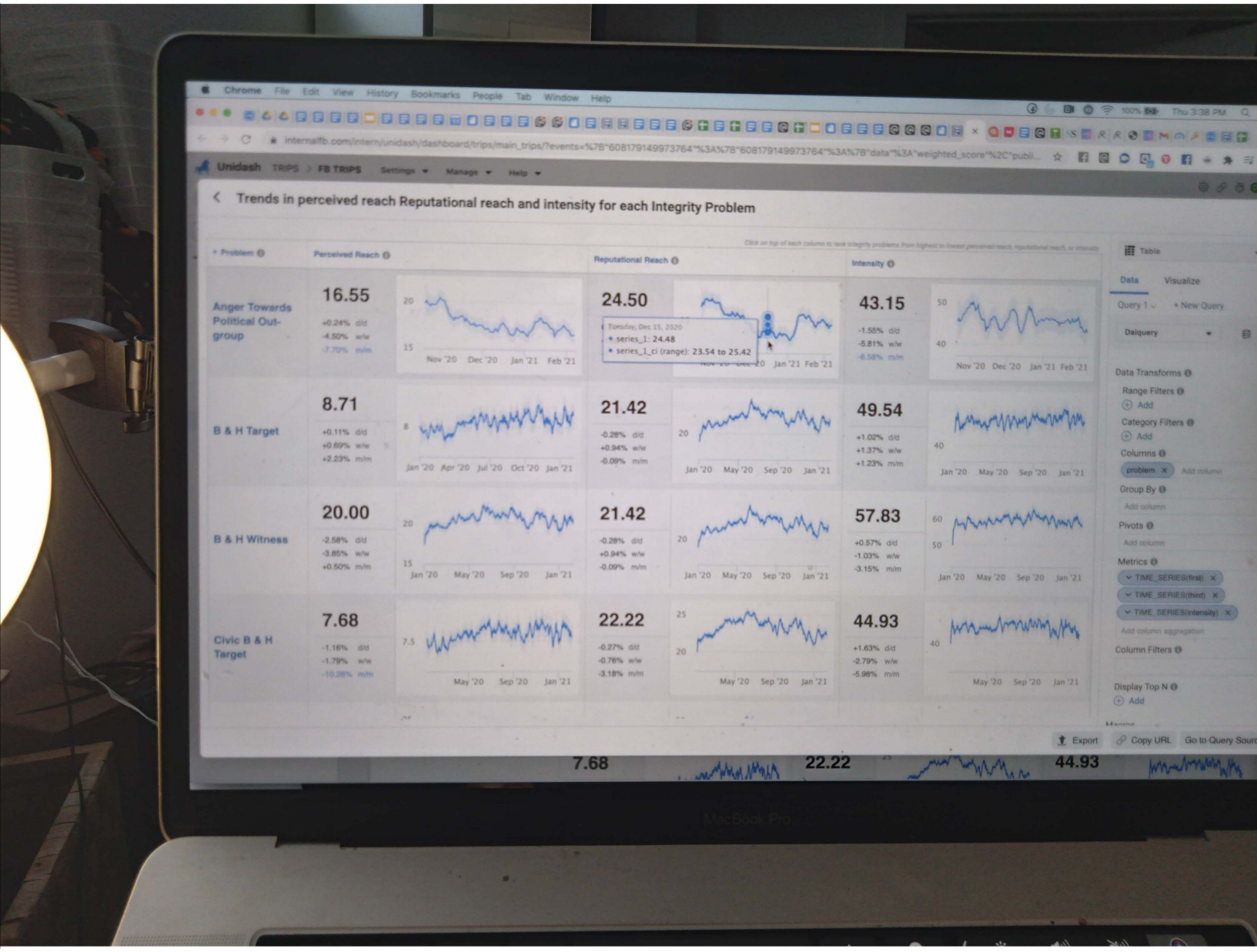

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:36 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Anger Towards Political Out-group
16.55
+0.24% d/d
-4.50% w/w
-7.70% m/m
Nov '20 Dec '20 Jan '21 Feb '21
24.50
0.00% d/d
-1.80% w/w
-4.41% m/m
43.15
-1.55% d/d
-5.81% w/w
-6.58% m/m
B & H Target
8.71
+0.11% d/d
+0.69% w/w
+2.23% m/m
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
Jan '20 May '20 Sep '20 Jan '21
49.54
+1.02% d/d
+1.37% w/w
+1.23% m/m
B & H Witness
20.00
-2.58% d/d
-3.85% w/w
+0.50% m/m
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
57.83
+0.57% d/d
-1.03% w/w
-3.15% m/m
Civic B & H Target
7.68
-1.16% d/d
-1.79% w/w
-10.28% m/m
May '20 Sep '20 Jan '21
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
44.93
+1.63% d/d
-2.79% w/w
-5.98% m/m
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
MacBook Pro
esc
Search Bunny

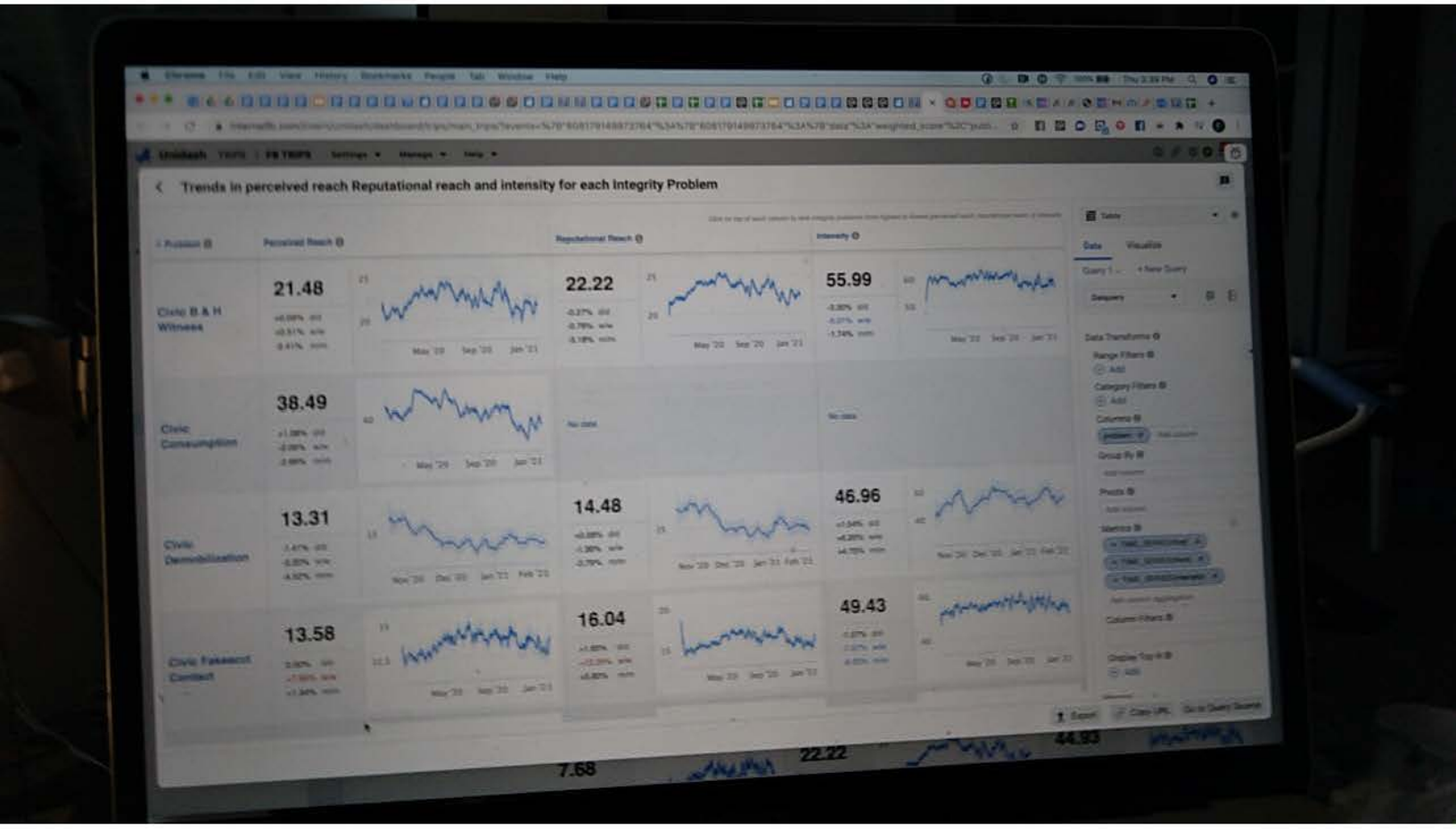

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:37 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
↑ Problem
Perceived Reach
Reputational Reach
Intensity
Civic B & H Witness
21.48
+0.09% d/d
+0.51% w/w
-2.41% m/m
May '20 Sep '20 Jan '21
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
55.99
-3.30% d/d
-6.01% w/w
-1.74% m/m
Civic Consumption
38.49
+1.08% d/d
-2.06% w/w
-2.66% m/m
No data
No data
Civic Demobilization
13.31
-1.41% d/d
-3.20% w/w
-4.52% m/m
Nov '20 Dec '20 Jan '21 Feb '21
14.48
+0.56% d/d
-1.36% w/w
-3.79% m/m
46.96
+1.54% d/d
+6.20% w/w
+4.75% m/m
Civic Fakeacct Contact
13.58
0.00% d/d
+7.86% w/w
+1.34% m/m
16.04
+1.65% d/d
+12.25% w/w
+5.60% m/m
49.43
-1.57% d/d
-7.97% w/w
-8.50% m/m
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Group By
Pivots
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
esc
Search Bunny

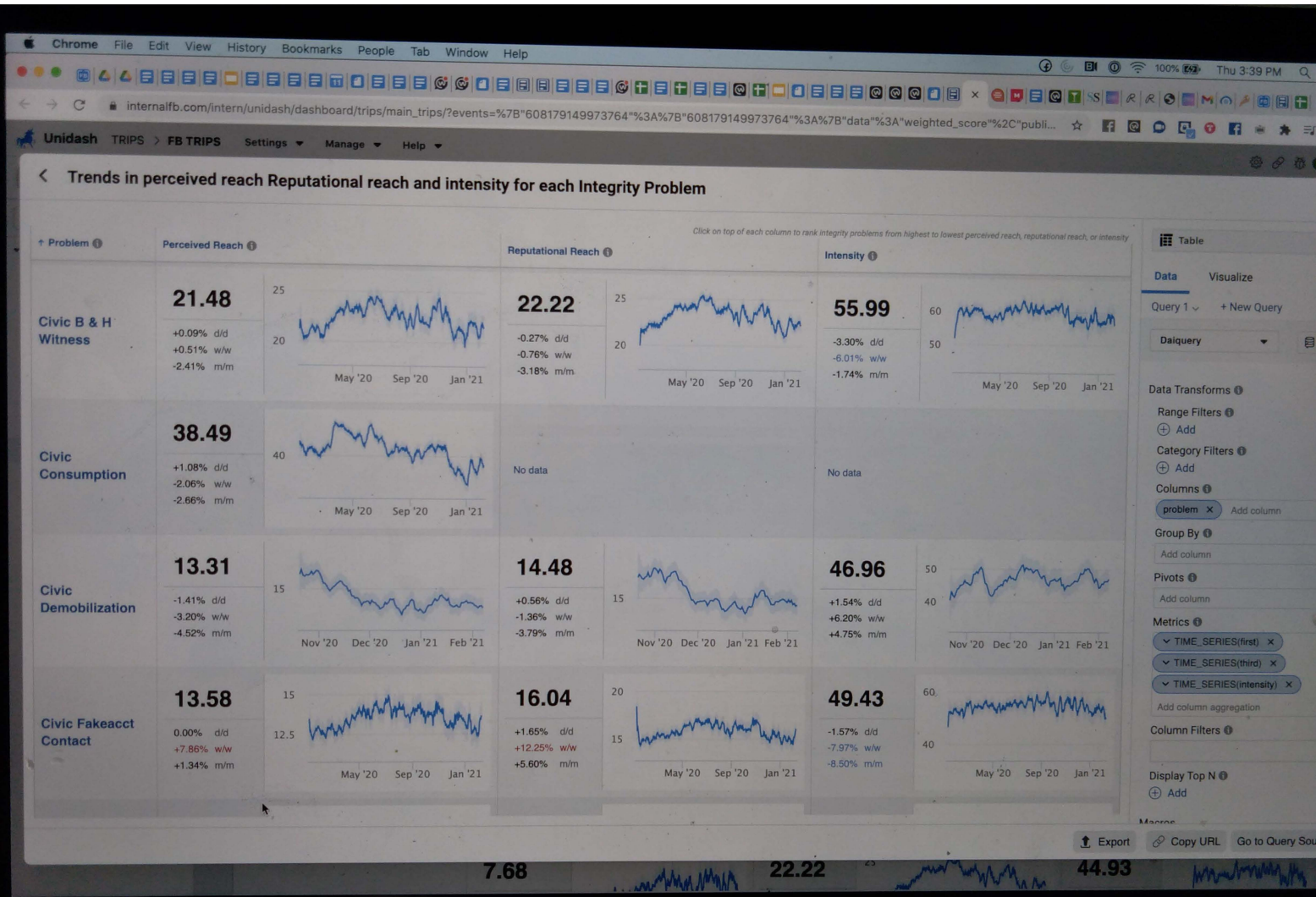

Unidash TRIPS FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Civic False and Misleading
22.13
-0.45% d/d
+3.90% w/w
-0.85% m/m
21.89
-1.35% d/d
+4.14% w/w
-1.00% m/m
47.72
-1.85% d/d
-4.25% w/w
-0.31% m/m
Civic Inflammatory
12.12
-0.16% d/d
+2.89% w/w
-3.58% m/m
9.10
-1.19% d/d
+1.11% w/w
-9.00% m/m
56.07
-2.16% d/d
-8.70% w/w
-9.74% m/m
Civic Online Discouragement
21.31
-0.75% d/d
-4.10% w/w
-2.83% m/m
20.70
+0.83% d/d
-0.29% w/w
-4.78% m/m
40.24
-1.64% d/d
-3.64% w/w
-0.07% m/m
Clickbait
25.45
+0.51% d/d
-0.35% w/w
+1.60% m/m
23.77
+0.93% d/d
-1.82% w/w
-2.06% m/m
46.84
-1.64% d/d
+4.58% w/w
+2.52% m/m
May '20 Sep '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
MacBook Pro
esc
Search Bunny
delete

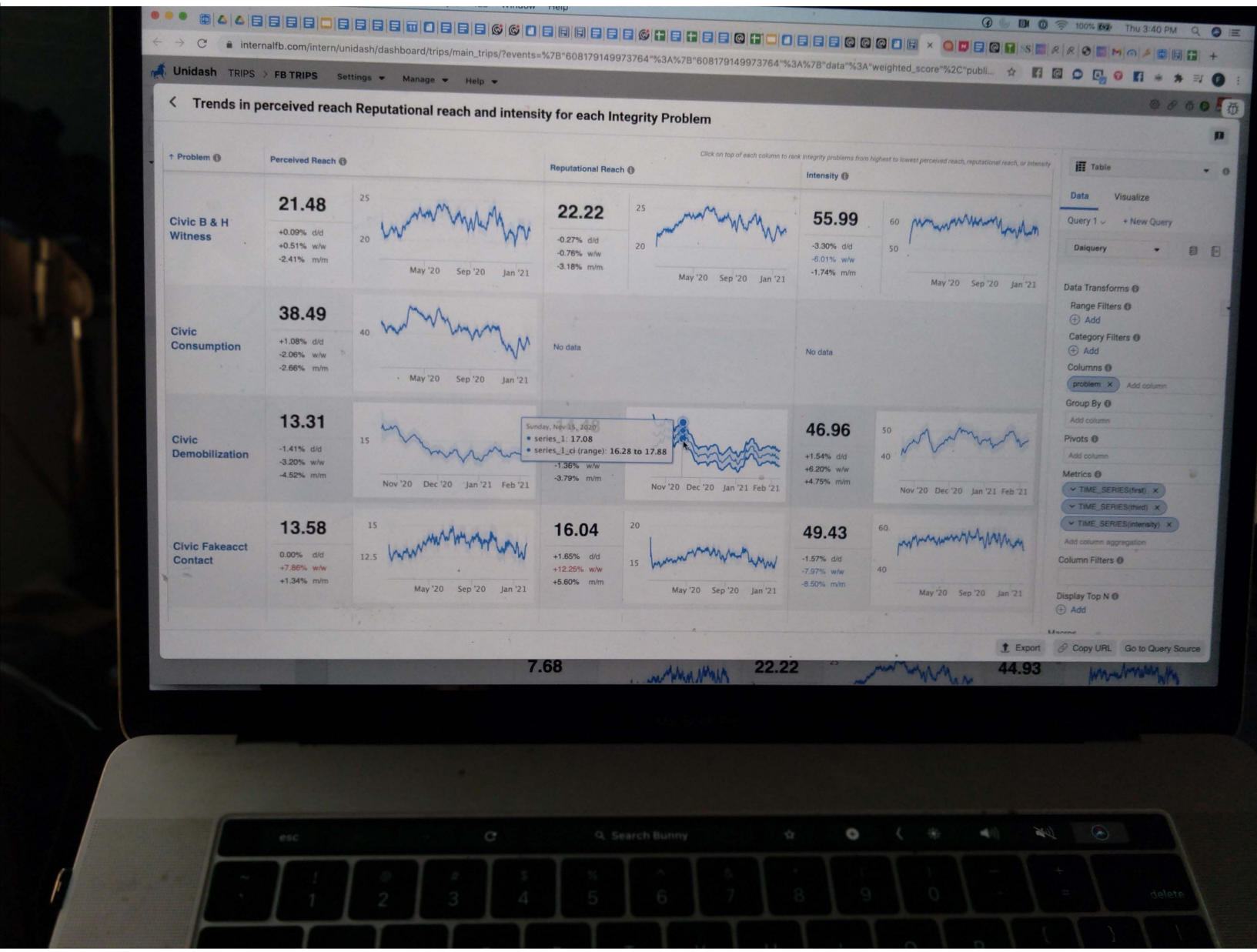

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:38 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Anger Towards Political Out-group
16.55
+0.24% d/d
-4.50% w/w
-7.70% m/m
24.50
0.00% d/d
-1.80% w/w
-4.41% m/m
43.15
-1.55% d/d
-5.81% w/w
-6.58% m/m
B & H Target
8.71
+0.11% d/d
+0.69% w/w
+2.23% m/m
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
49.54
+1.02% d/d
+1.37% w/w
+1.23% m/m
B & H Witness
20.00
-2.58% d/d
-3.85% w/w
+0.50% m/m
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
57.83
+0.57% d/d
-1.03% w/w
-3.15% m/m
Civic B & H Target
7.68
-1.16% d/d
-1.79% w/w
-10.28% m/m
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
44.93
+1.63% d/d
-2.79% w/w
-5.98% m/m
Nov '20 Dec '20 Jan '21 Feb '21
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
May '20 Sep '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
MacBook Pro
esc
Search Bunny

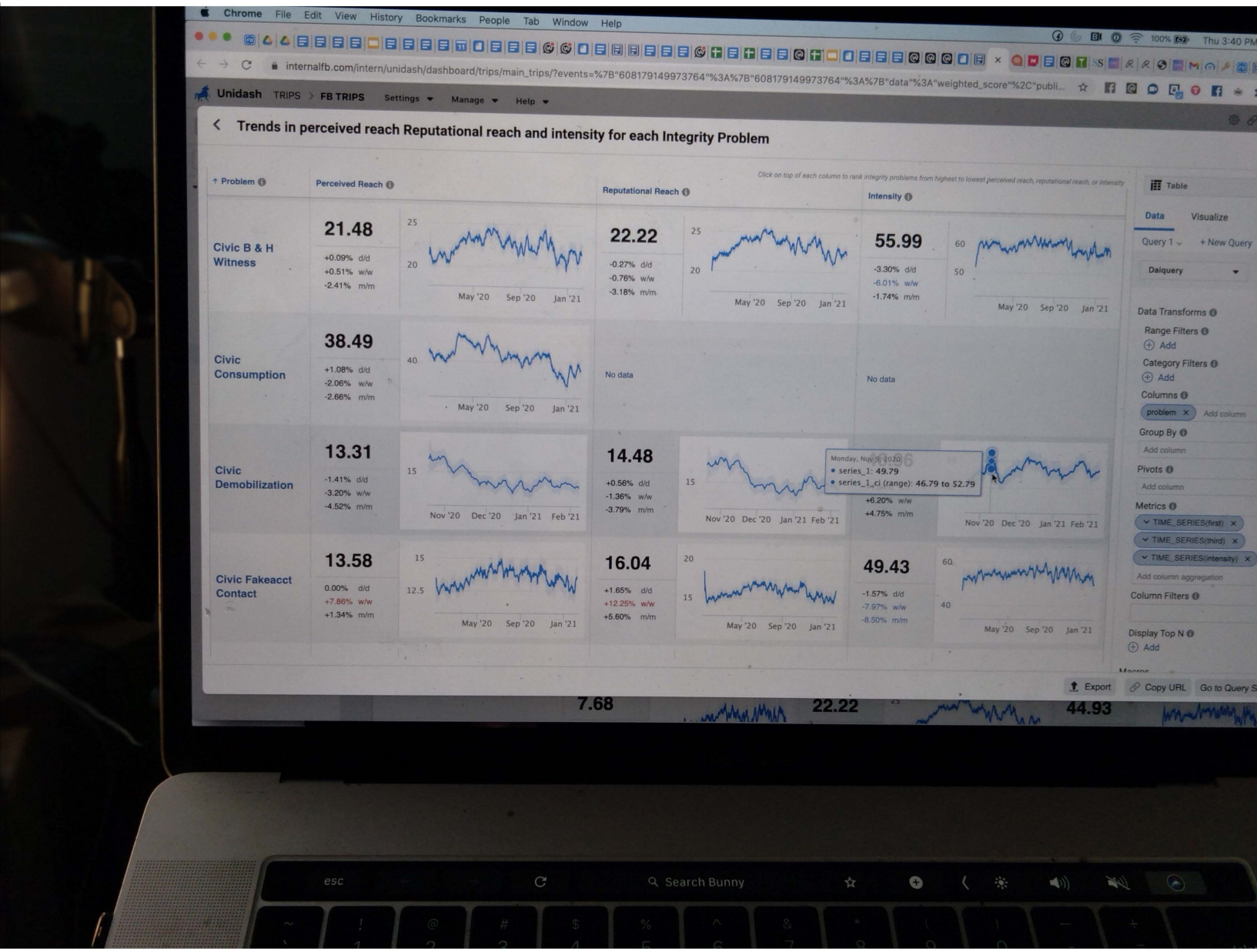

Chrome File Edit View History Bookmarks People Tab Window Help
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Anger Towards Political Out-group
16.55
+0.24% d/d
-4.50% w/w
-7.70% m/m
Nov '20 Dec '20 Jan '21 Feb '21
24.50
0.00%
-1.80%
-4.41%
Sunday, Jan 3, 2021
series_1: 22.64
series_1_ci (range): 21.73 to 23.55
43.15
-1.55% d/d
-5.81% w/w
-6.58% m/m
B & H Target
8.71
+0.11% d/d
+0.69% w/w
+2.23% m/m
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
Jan '20 May '20 Sep '20 Jan '21
49.54
+1.02% d/d
+1.37% w/w
+1.23% m/m
B & H Witness
20.00
-2.58% d/d
-3.85% w/w
+0.50% m/m
Jan '20 May '20 Sep '20 Jan '21
21.42
-0.28% d/d
+0.94% w/w
-0.09% m/m
57.83
+0.57% d/d
-1.03% w/w
-3.15% m/m
Civic B & H Target
7.68
-1.16% d/d
-1.79% w/w
-10.28% m/m
May '20 Sep '20 Jan '21
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
44.93
+1.63% d/d
-2.79% w/w
-5.98% m/m
Table
Data
Query 1
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Group By
Add column
Pivots
Add column
Metrics
Column Filters
Display Top N
Add
Export
Copy URL
7.68
22.22
44.93
MacBook Pro
esc
Search Bunny

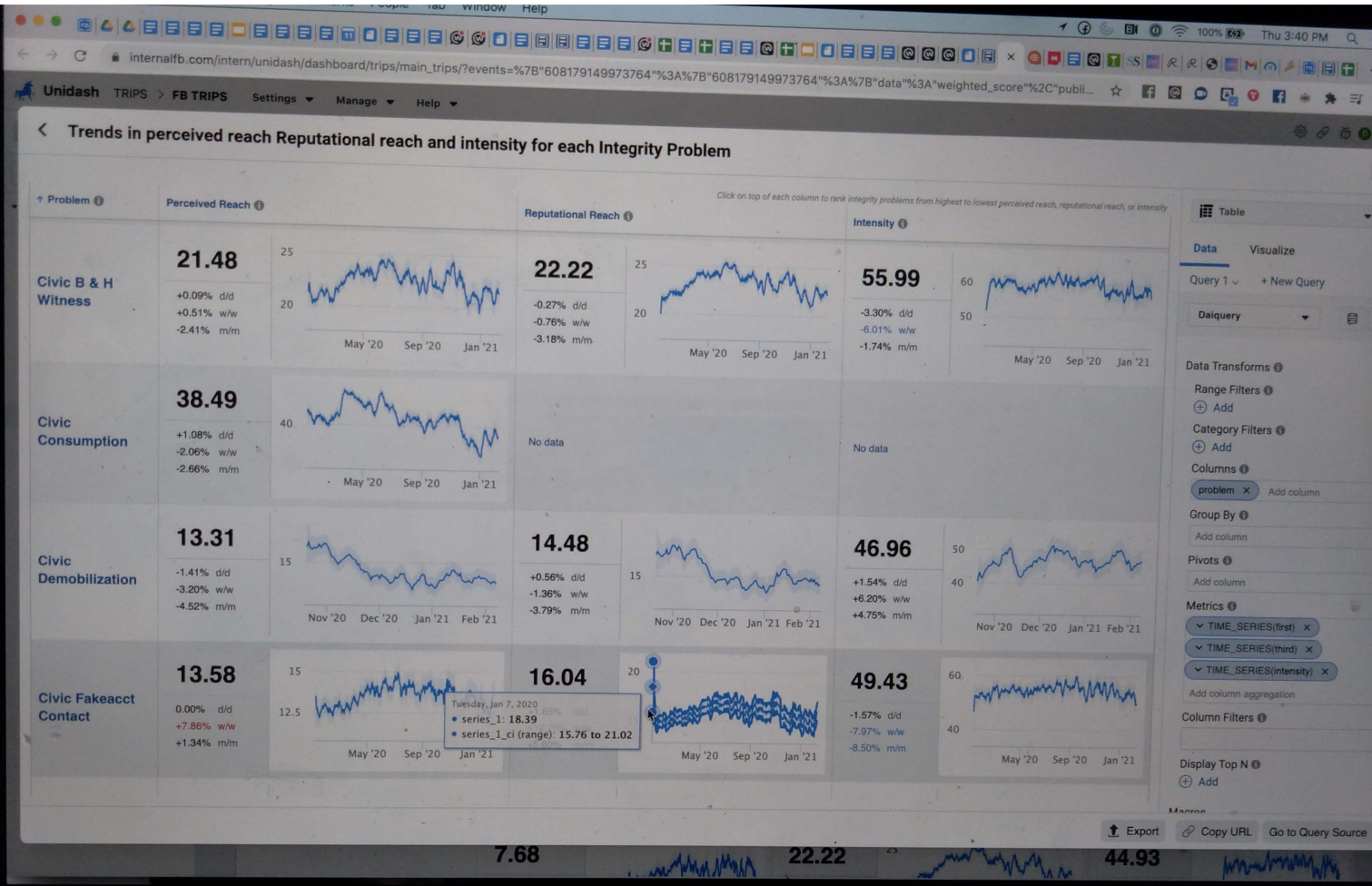

Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Anger Towards Political Out-group
16.55
24.50
43.15
B & H Target
8.71
21.42
49.54
B & H Witness
20.00
21.42
57.83
Civic B & H Target
7.68
22.22
44.93
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Category Filters
Columns
Group By
Pivots
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Column Filters
Display Top N
Export
Copy URL

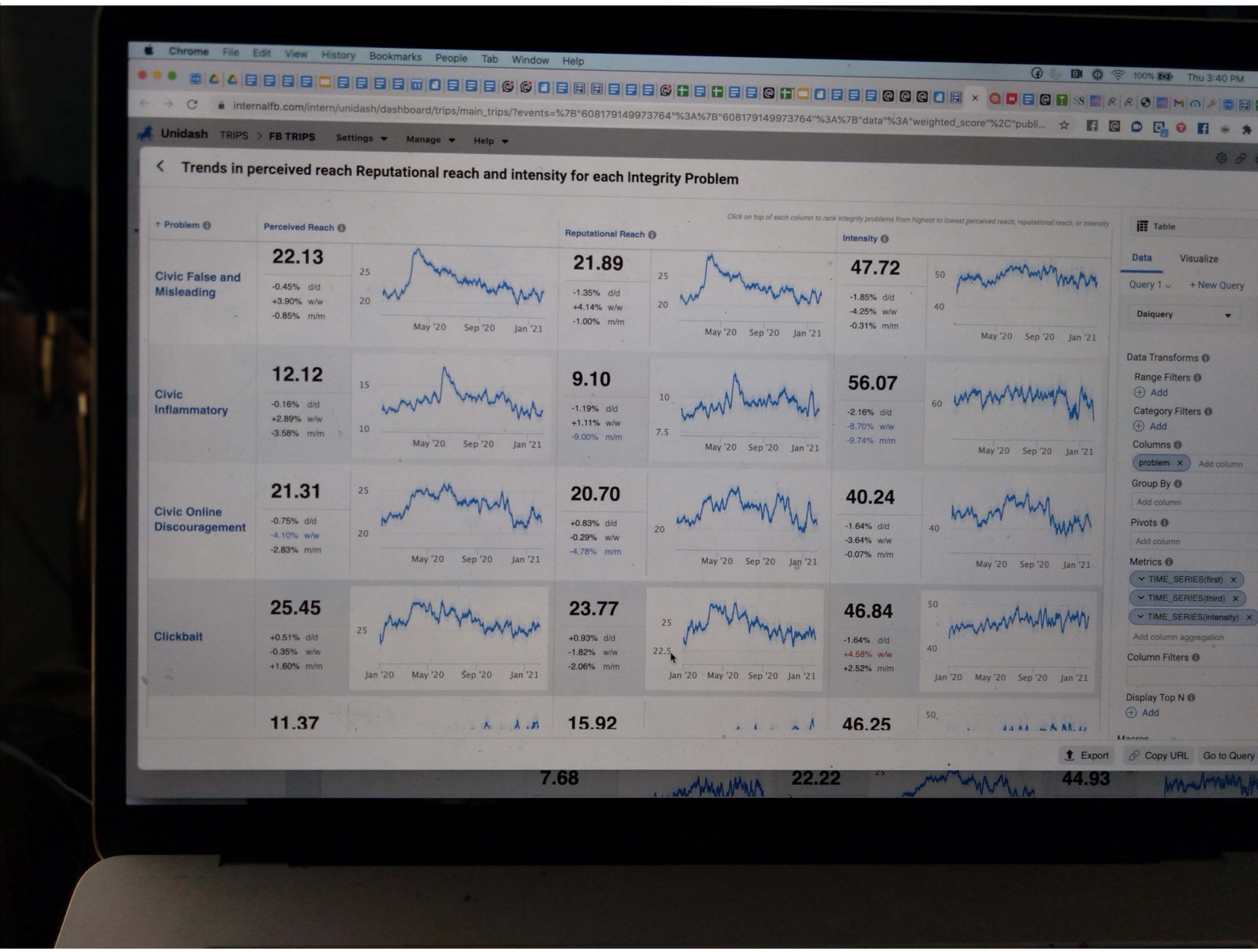

Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Perceived Reach
Reputational Reach
Intensity
Civic B & H Witness
21.48
22.22
55.99
Civic Consumption
38.49
Civic Demobilization
13.31
14.48
46.96
Civic Fakeacct Contact
13.58
16.04
49.43
7.68
22.22
44.93

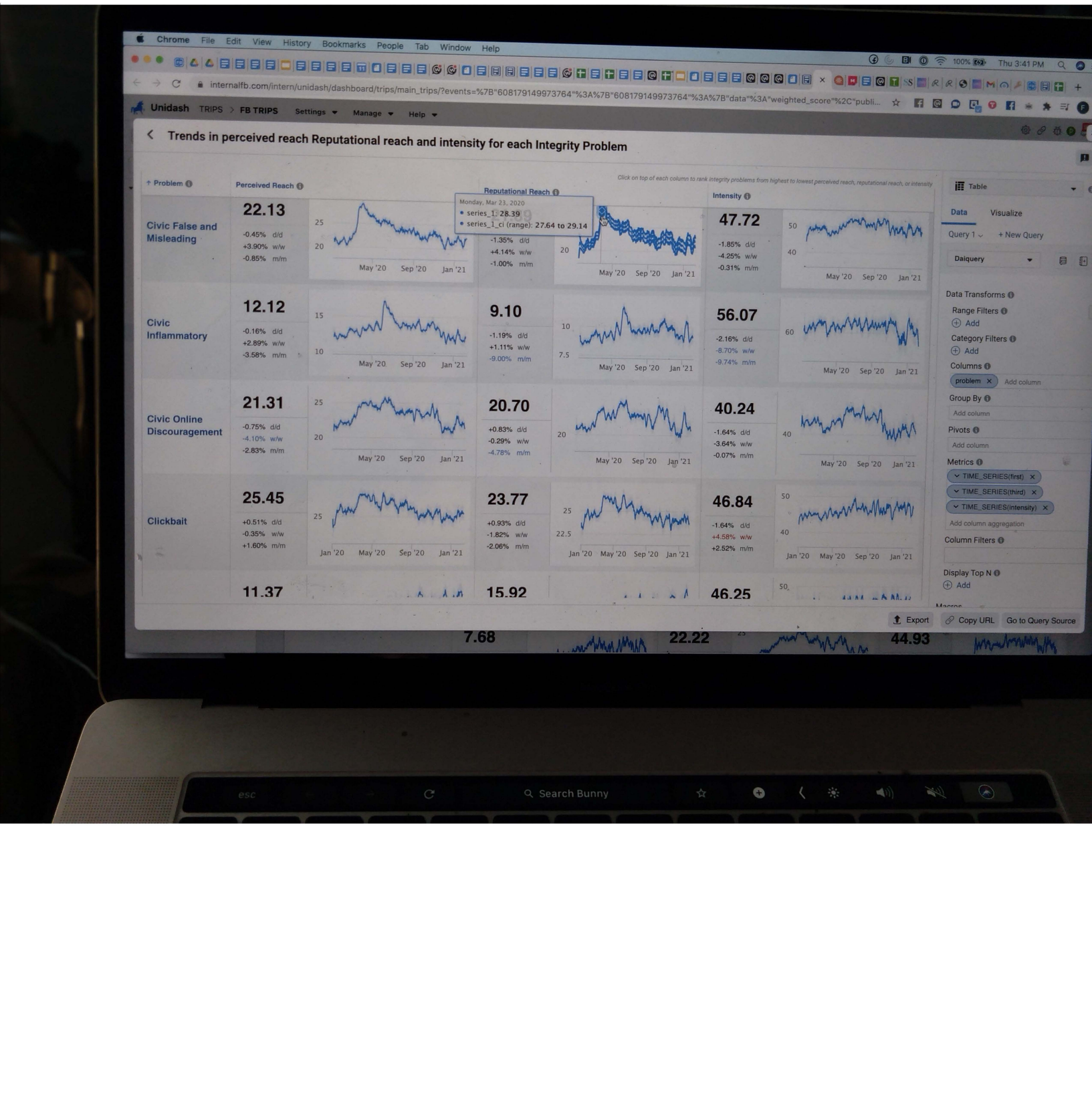

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:39 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
↑ Problem
Perceived Reach
Reputational Reach
Intensity
Civic B & H Witness
21.48
+0.09% d/d
+0.51% w/w
-2.41% m/m
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
55.99
-3.30% d/d
-6.01% w/w
-1.74% m/m
Civic Consumption
38.49
+1.08% d/d
-2.06% w/w
-2.66% m/m
No data
No data
Civic Demobilization
13.31
-1.41% d/d
-3.20% w/w
-4.52% m/m
14.48
+0.56% d/d
-1.36% w/w
-3.79% m/m
46.96
+1.54% d/d
+6.20% w/w
+4.75% m/m
Civic Fakeacct Contact
13.58
0.00% d/d
+7.86% w/w
+1.34% m/m
16.04
+1.65% d/d
+12.25% w/w
+5.60% m/m
49.43
-1.57% d/d
-7.97% w/w
-8.50% m/m
May '20 Sep '20 Jan '21
Nov '20 Dec '20 Jan '21 Feb '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Sou
7.68
22.22
44.93

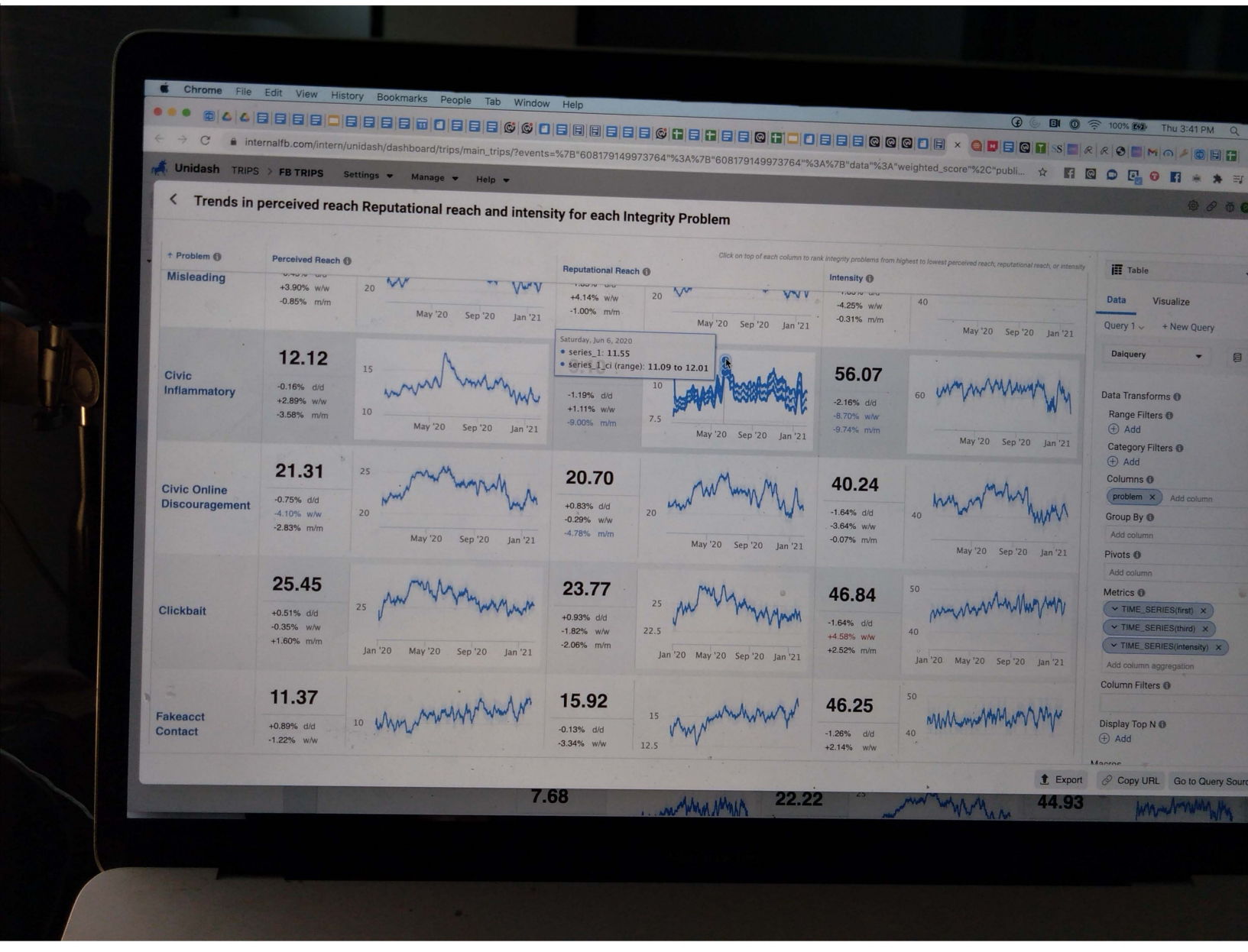

Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Civic B & H Witness
21.48
22.22
55.99
Civic Consumption
38.49
No data
No data
Civic Demobilization
13.31
46.96
Civic Fakeacct Contact
13.58
16.04
49.43
May '20 Sep '20 Jan '21
Nov '20 Dec '20 Jan '21 Feb '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Category Filters
Columns
Group By
Pivots
Metrics
Column Filters
Display Top N
Export
Copy URL
Go to Query Source
7.68
22.22
44.93

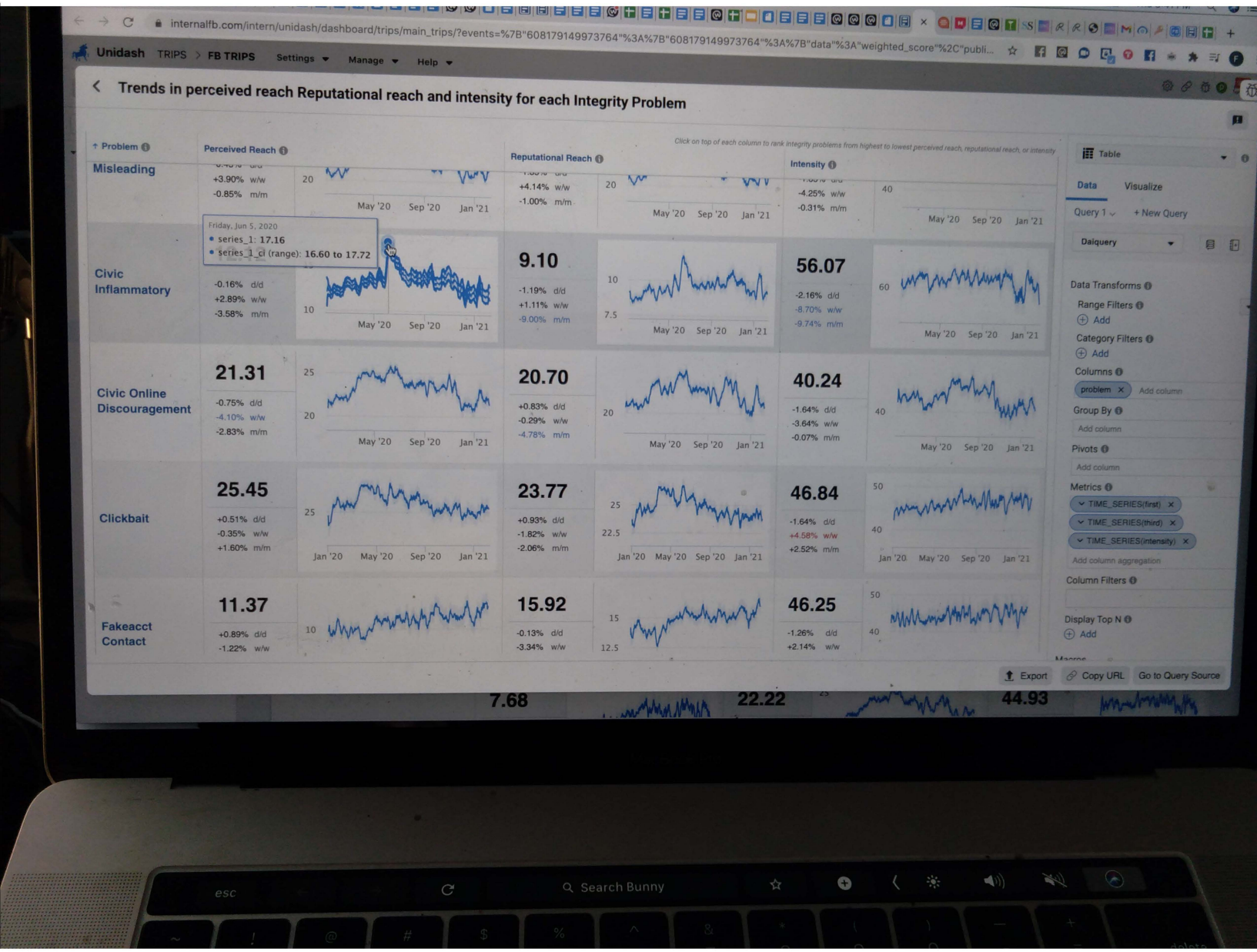

Chrome File Edit View History Bookmarks People Tab Window Help
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Civic B & H Witness
21.48
+0.09% d/d
+0.51% w/w
-2.41% m/m
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
55.99
-3.30% d/d
-6.01% w/w
-1.74% m/m
Civic Consumption
38.49
+1.08% d/d
-2.06% w/w
-2.66% m/m
No data
No data
Civic Demobilization
13.31
-1.41% d/d
-3.20% w/w
-4.52% m/m
14.48
+0.56% d/d
-1.36% w/w
-3.79% m/m
+6.20% w/w
+4.75% m/m
Monday, Nov 9, 2020
series_1: 49.79
series_1_ci (range): 46.79 to 52.79
Civic Fakeacct Contact
13.58
0.00% d/d
+7.86% w/w
+1.34% m/m
16.04
+1.65% d/d
+12.25% w/w
+5.60% m/m
49.43
-1.57% d/d
-7.97% w/w
-8.50% m/m
May '20 Sep '20 Jan '21
Nov '20 Dec '20 Jan '21 Feb '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters Add
Category Filters Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N Add
Export
Copy URL
Go to Query S
7.68
22.22
44.93
esc
Search Bunny

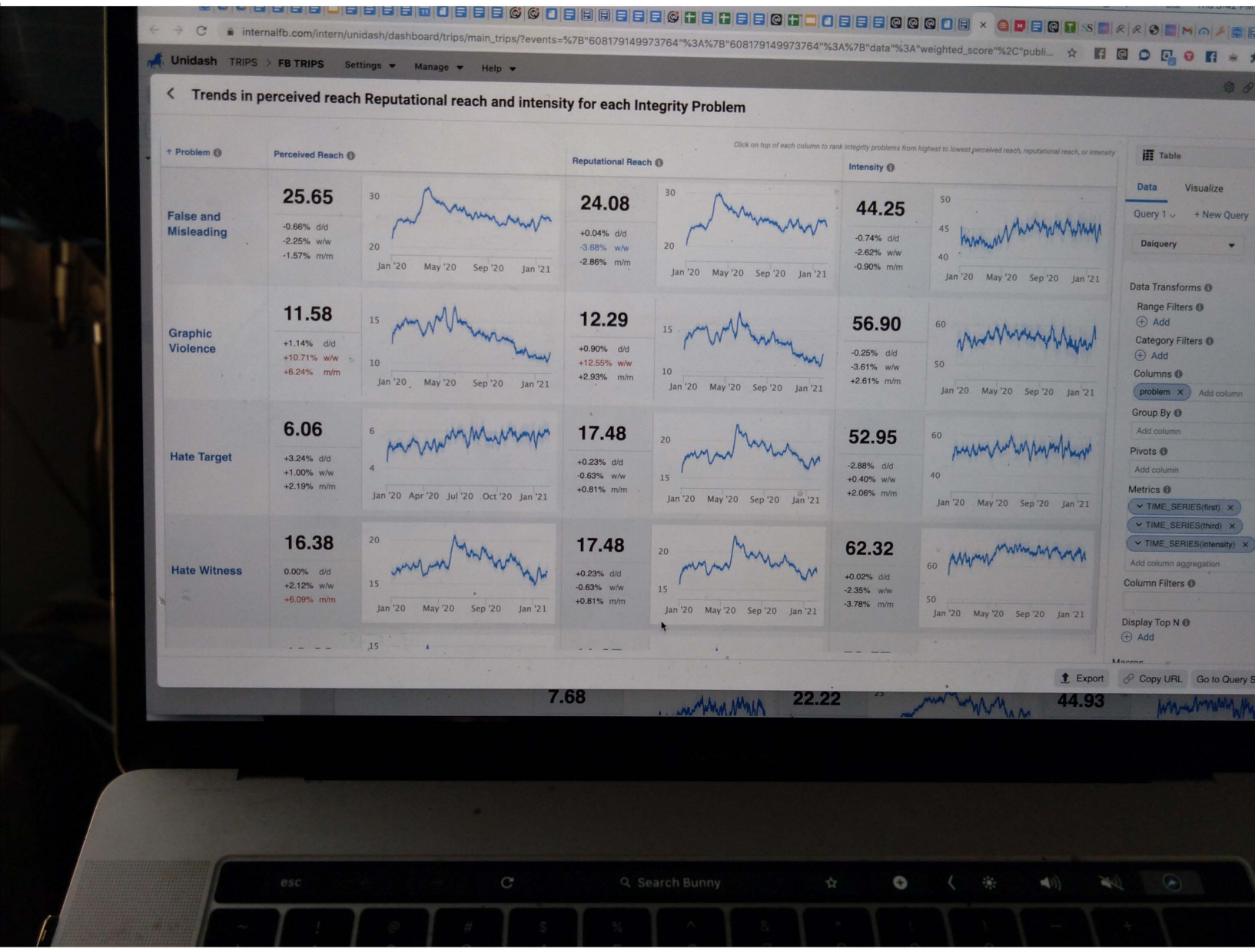

Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Civic B & H Witness
21.48
+0.09% d/d
+0.51% w/w
-2.41% m/m
22.22
-0.27% d/d
-0.76% w/w
-3.18% m/m
55.99
-3.30% d/d
-6.01% w/w
-1.74% m/m
Civic Consumption
38.49
+1.08% d/d
-2.06% w/w
-2.66% m/m
No data
No data
Civic Demobilization
13.31
-1.41% d/d
-3.20% w/w
-4.52% m/m
14.48
+0.56% d/d
-1.36% w/w
-3.79% m/m
46.96
+1.54% d/d
+6.20% w/w
+4.75% m/m
Civic Fakeacct Contact
13.58
0.00% d/d
+7.86% w/w
+1.34% m/m
16.04
Tuesday, Jan 7, 2020
series_1: 18.39
series_1_ci (range): 15.76 to 21.02
49.43
-1.57% d/d
-7.97% w/w
-8.50% m/m
May '20 Sep '20 Jan '21
Nov '20 Dec '20 Jan '21 Feb '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters Add
Category Filters Add
Columns problem Add column
Group By Add column
Pivots Add column
Metrics TIME_SERIES(first) TIME_SERIES(third) TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N Add
Export Copy URL Go to Query Source
7.68
22.22
44.93
esc
Search Bunny

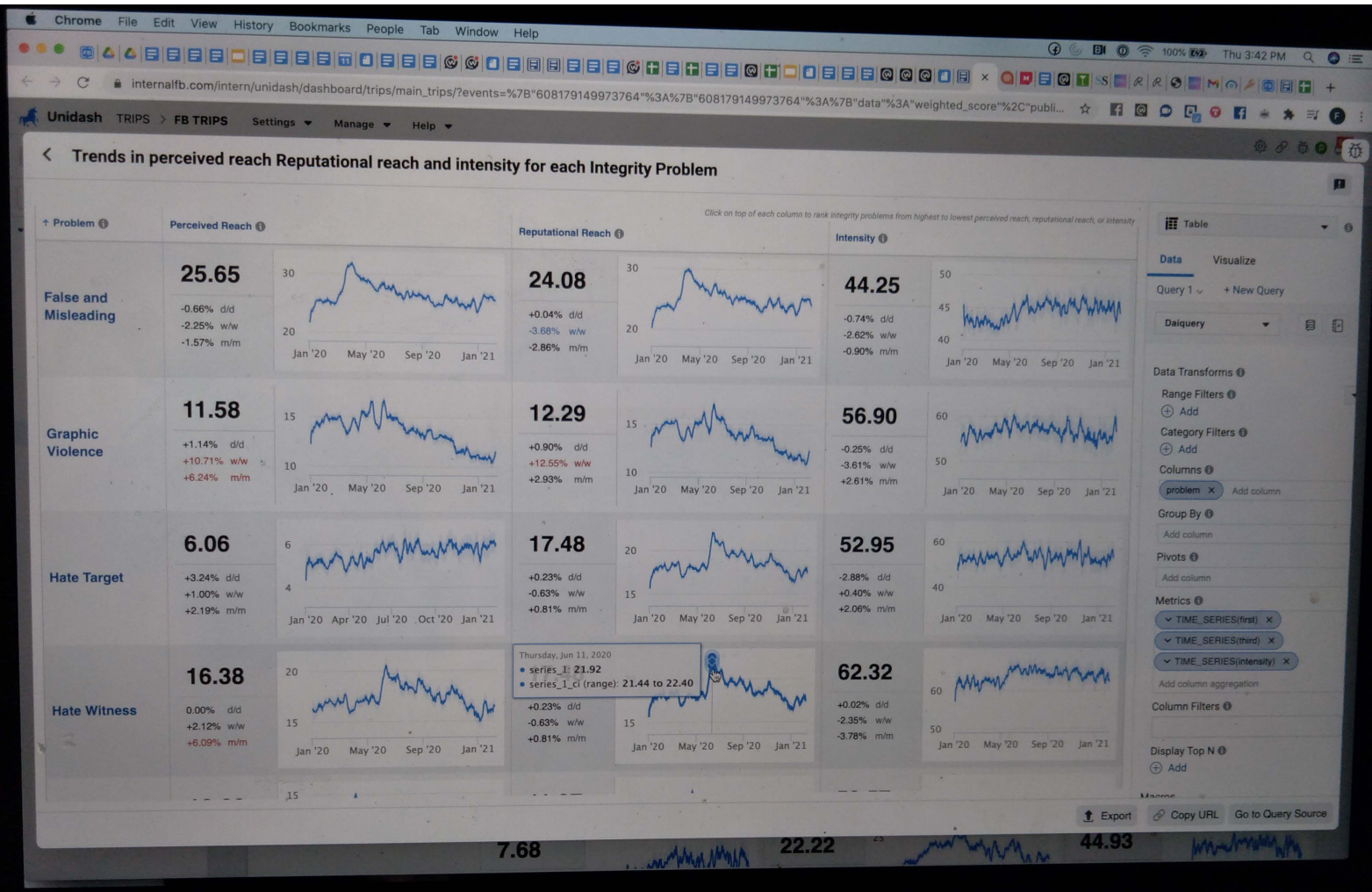

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:40 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Civic False and Misleading
22.13
-0.45% d/d
+3.90% w/w
-0.85% m/m
21.89
-1.35% d/d
+4.14% w/w
-1.00% m/m
47.72
-1.85% d/d
-4.25% w/w
-0.31% m/m
Civic Inflammatory
12.12
-0.16% d/d
+2.89% w/w
-3.58% m/m
9.10
-1.19% d/d
+1.11% w/w
-9.00% m/m
56.07
-2.16% d/d
-8.70% w/w
-9.74% m/m
Civic Online Discouragement
21.31
-0.75% d/d
-4.10% w/w
-2.83% m/m
20.70
+0.83% d/d
-0.29% w/w
-4.78% m/m
40.24
-1.64% d/d
-3.64% w/w
-0.07% m/m
Clickbait
25.45
+0.51% d/d
-0.35% w/w
+1.60% m/m
23.77
+0.93% d/d
-1.82% w/w
-2.06% m/m
46.84
-1.64% d/d
+4.58% w/w
+2.52% m/m
11.37
15.92
46.25
May '20
Sep '20
Jan '21
Jan '20
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query
7.68
22.22
44.93

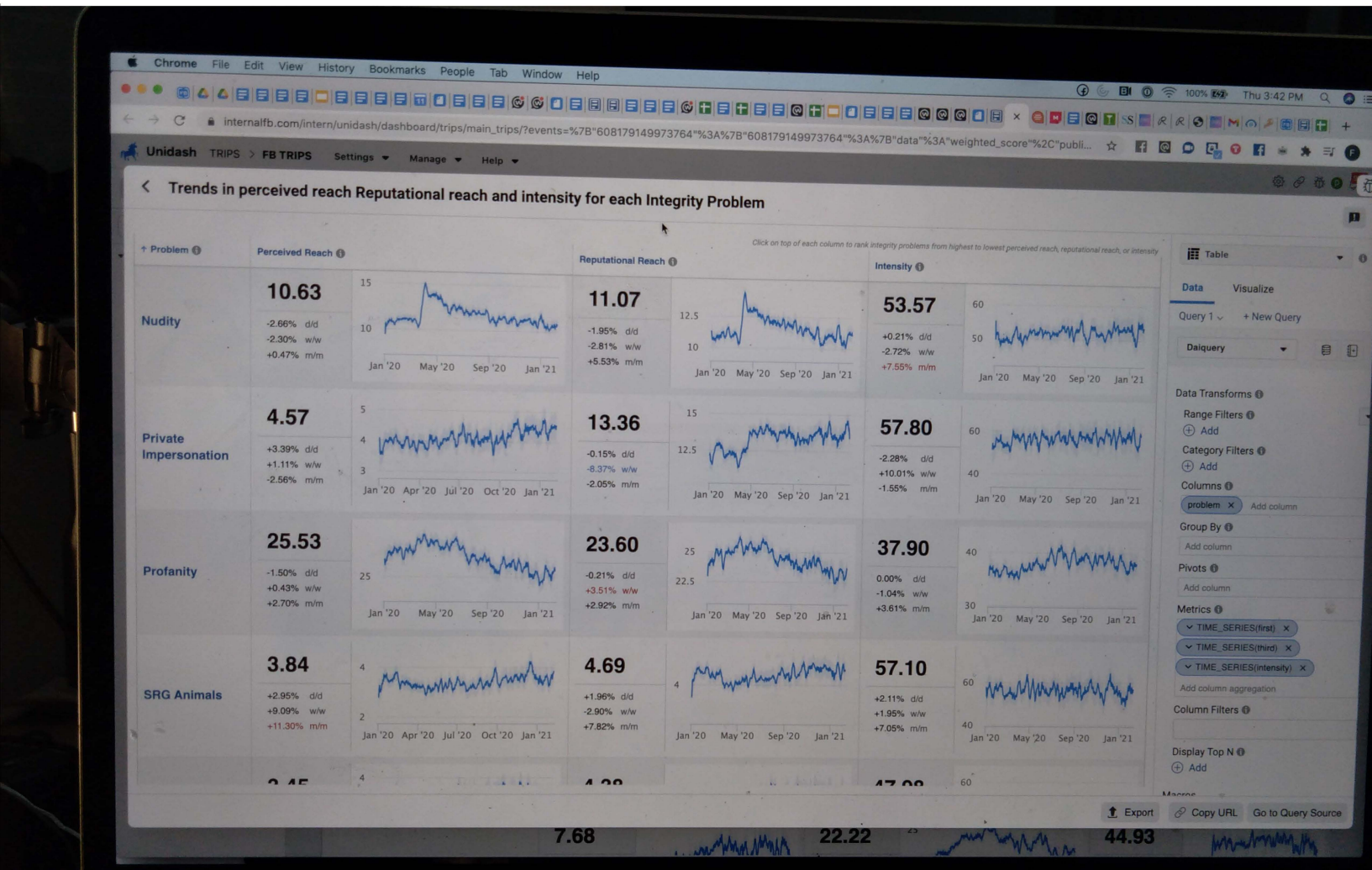

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:41 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Monday, Mar 23, 2020
series_1: 28.39
series_1_ci (range): 27.64 to 29.14
Civic False and Misleading
22.13
-0.45% d/d
+3.90% w/w
-0.85% m/m
-1.35% d/d
+4.14% w/w
-1.00% m/m
47.72
-1.85% d/d
-4.25% w/w
-0.31% m/m
Civic Inflammatory
12.12
-0.16% d/d
+2.89% w/w
-3.58% m/m
9.10
-1.19% d/d
+1.11% w/w
-9.00% m/m
56.07
-2.16% d/d
-8.70% w/w
-9.74% m/m
Civic Online Discouragement
21.31
-0.75% d/d
-4.10% w/w
-2.83% m/m
20.70
+0.83% d/d
-0.29% w/w
-4.78% m/m
40.24
-1.64% d/d
-3.64% w/w
-0.07% m/m
Clickbait
25.45
+0.51% d/d
-0.35% w/w
+1.60% m/m
23.77
+0.93% d/d
-1.82% w/w
-2.06% m/m
46.84
-1.64% d/d
+4.58% w/w
+2.52% m/m
11.37
15.92
46.25
May '20 Sep '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
esc
Search Bunny

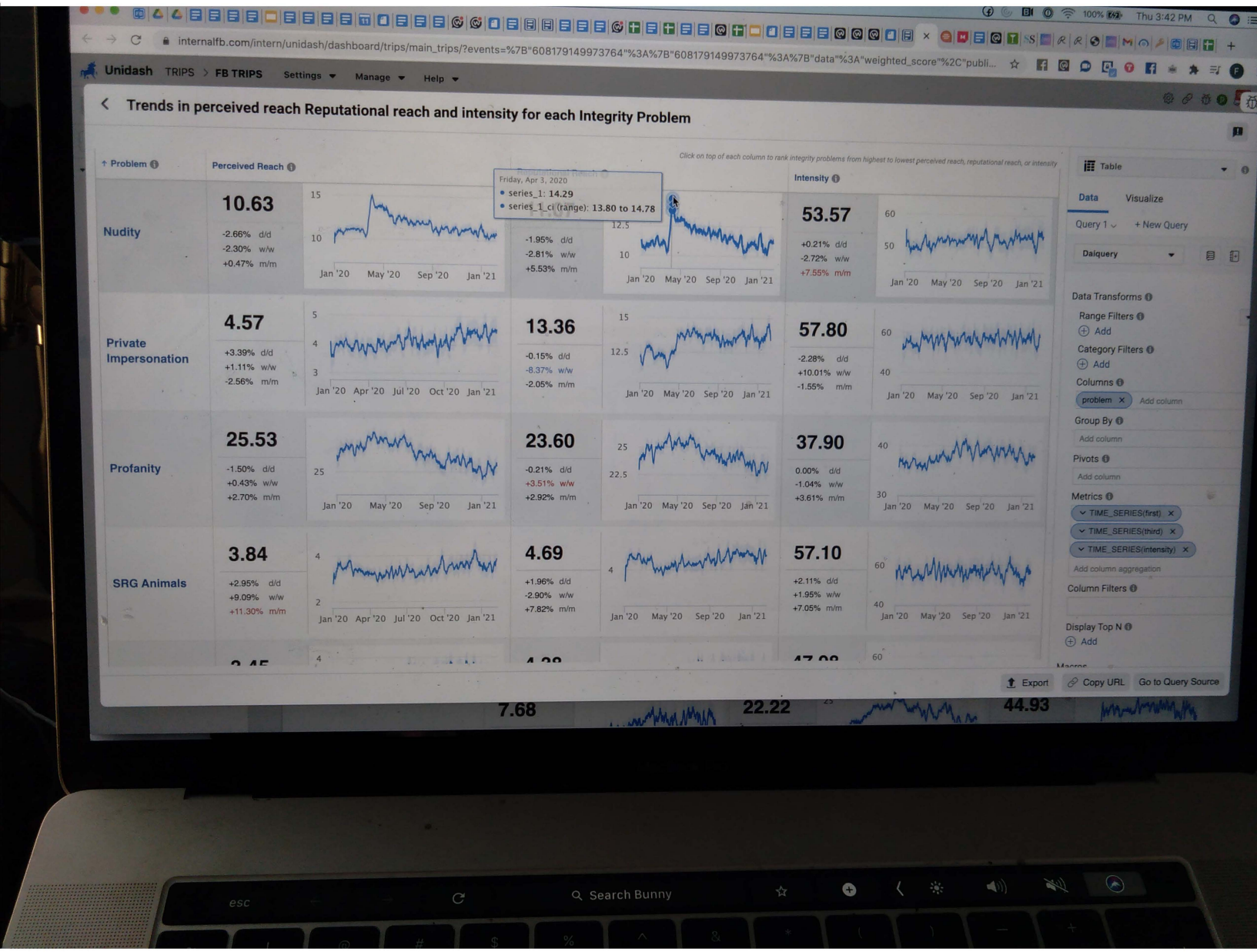

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:41 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Misleading
+3.90% w/w
-0.85% m/m
+4.14% w/w
-1.00% m/m
-4.25% w/w
-0.31% m/m
Civic Inflammatory
12.12
-0.16% d/d
+2.89% w/w
-3.58% m/m
Saturday, Jun 6, 2020
series_1: 11.55
series_1_ci (range): 11.09 to 12.01
-1.19% d/d
+1.11% w/w
-9.00% m/m
56.07
-2.16% d/d
-8.70% w/w
-9.74% m/m
Civic Online Discouragement
21.31
-0.75% d/d
-4.10% w/w
-2.83% m/m
20.70
+0.83% d/d
-0.29% w/w
-4.78% m/m
40.24
-1.64% d/d
-3.64% w/w
-0.07% m/m
Clickbait
25.45
+0.51% d/d
-0.35% w/w
+1.60% m/m
23.77
+0.93% d/d
-1.82% w/w
-2.06% m/m
46.84
-1.64% d/d
+4.58% w/w
+2.52% m/m
Fakeacct Contact
11.37
+0.89% d/d
-1.22% w/w
15.92
-0.13% d/d
-3.34% w/w
46.25
-1.26% d/d
+2.14% w/w
May '20 Sep '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters Add
Category Filters Add
Columns
problem Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N Add
Export
Copy URL
Go to Query Sourc
7.68
22.22
44.93

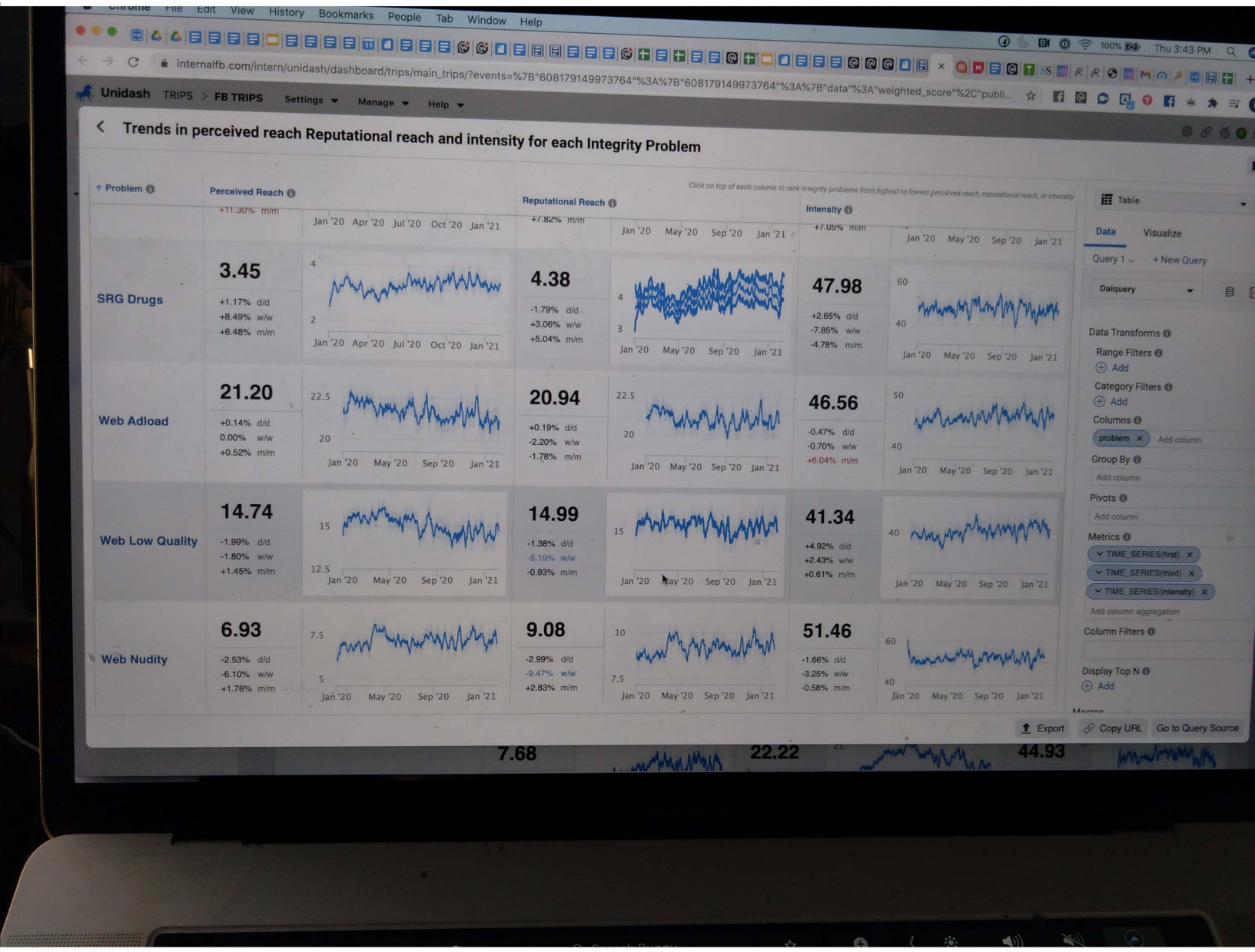

internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Misleading
+3.90% w/w
-0.85% m/m
+4.14% w/w
-1.00% m/m
-4.25% w/w
-0.31% m/m
Friday, Jun 5, 2020
series_1: 17.16
series_1_ci (range): 16.60 to 17.72
Civic Inflammatory
-0.16% d/d
+2.89% w/w
-3.58% m/m
9.10
-1.19% d/d
+1.11% w/w
-9.00% m/m
56.07
-2.16% d/d
-8.70% w/w
-9.74% m/m
Civic Online Discouragement
21.31
-0.75% d/d
-4.10% w/w
-2.83% m/m
20.70
+0.83% d/d
-0.29% w/w
-4.78% m/m
40.24
-1.64% d/d
-3.64% w/w
-0.07% m/m
Clickbait
25.45
+0.51% d/d
-0.35% w/w
+1.60% m/m
23.77
+0.93% d/d
-1.82% w/w
-2.06% m/m
46.84
-1.64% d/d
+4.58% w/w
+2.52% m/m
Fakeacct Contact
11.37
+0.89% d/d
-1.22% w/w
15.92
-0.13% d/d
-3.34% w/w
46.25
-1.26% d/d
+2.14% w/w
May '20 Sep '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters Add
Category Filters Add
Columns problem Add column
Group By Add column
Pivots Add column
Metrics TIME_SERIES(first) TIME_SERIES(third) TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N Add
Export Copy URL Go to Query Source
7.68
22.22
44.93
esc
Search Bunny

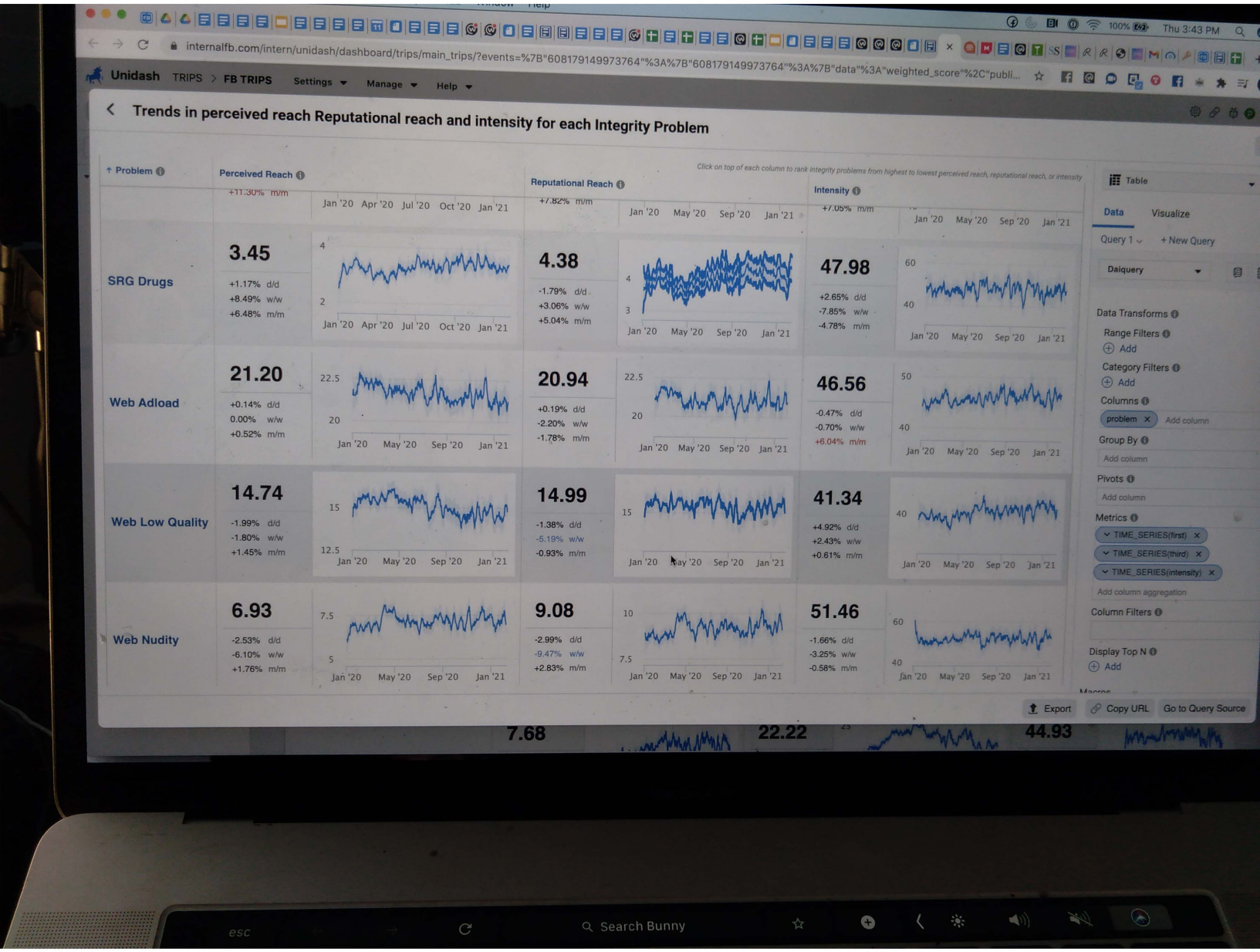

internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash
TRIPS > FB TRIPS
Settings
Manage
Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
False and Misleading
25.65
-0.66% d/d
-2.25% w/w
-1.57% m/m
24.08
+0.04% d/d
-3.68% w/w
-2.86% m/m
44.25
-0.74% d/d
-2.62% w/w
-0.90% m/m
Graphic Violence
11.58
+1.14% d/d
+10.71% w/w
+6.24% m/m
12.29
+0.90% d/d
+12.55% w/w
+2.93% m/m
56.90
-0.25% d/d
-3.61% w/w
+2.61% m/m
Hate Target
6.06
+3.24% d/d
+1.00% w/w
+2.19% m/m
17.48
+0.23% d/d
-0.63% w/w
+0.81% m/m
52.95
-2.88% d/d
+0.40% w/w
+2.06% m/m
Hate Witness
16.38
0.00% d/d
+2.12% w/w
+6.09% m/m
17.48
+0.23% d/d
-0.63% w/w
+0.81% m/m
62.32
+0.02% d/d
-2.35% w/w
-3.78% m/m
Jan '20
May '20
Sep '20
Jan '21
Apr '20
Jul '20
Oct '20
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query S
7.68
22.22
44.93
esc
Search Bunny

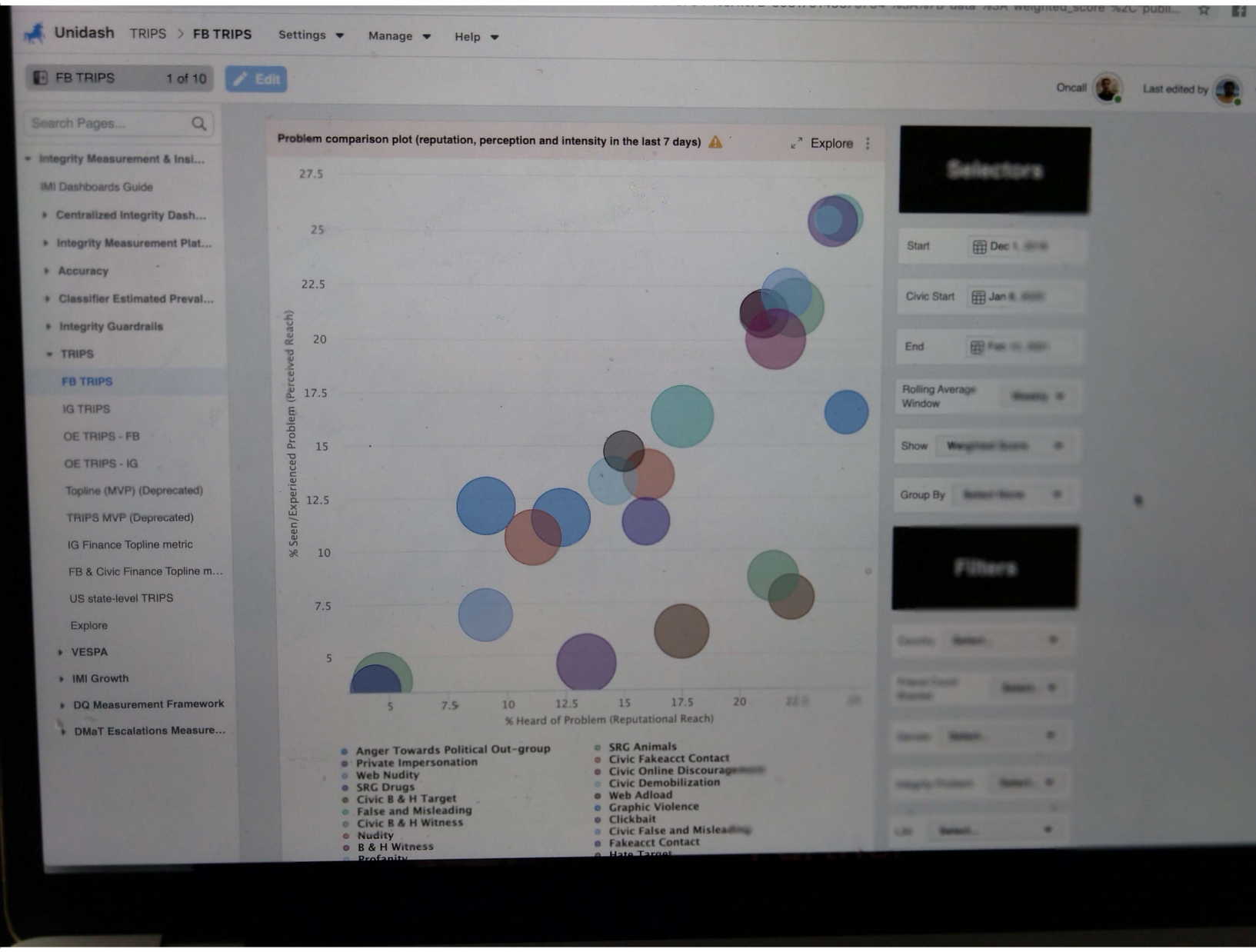

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:42 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
False and Misleading
25.65
-0.66% d/d
-2.25% w/w
-1.57% m/m
24.08
+0.04% d/d
-3.68% w/w
-2.86% m/m
44.25
-0.74% d/d
-2.62% w/w
-0.90% m/m
Graphic Violence
11.58
+1.14% d/d
+10.71% w/w
+6.24% m/m
12.29
+0.90% d/d
+12.55% w/w
+2.93% m/m
56.90
-0.25% d/d
-3.61% w/w
+2.61% m/m
Hate Target
6.06
+3.24% d/d
+1.00% w/w
+2.19% m/m
17.48
+0.23% d/d
-0.63% w/w
+0.81% m/m
52.95
-2.88% d/d
+0.40% w/w
+2.06% m/m
Hate Witness
16.38
0.00% d/d
+2.12% w/w
+6.09% m/m
Thursday, Jun 11, 2020
series_1: 21.92
series_1_ci (range): 21.44 to 22.40
+0.23% d/d
-0.63% w/w
+0.81% m/m
62.32
+0.02% d/d
-2.35% w/w
-3.78% m/m
Jan '20 May '20 Sep '20 Jan '21
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
Table
Data Visualize
Query 1 + New Query
Daiquery
Data Transforms
Range Filters Add
Category Filters Add
Columns problem Add column
Group By Add column
Pivots Add column
Metrics TIME_SERIES(first) TIME_SERIES(third) TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
Search Bunny

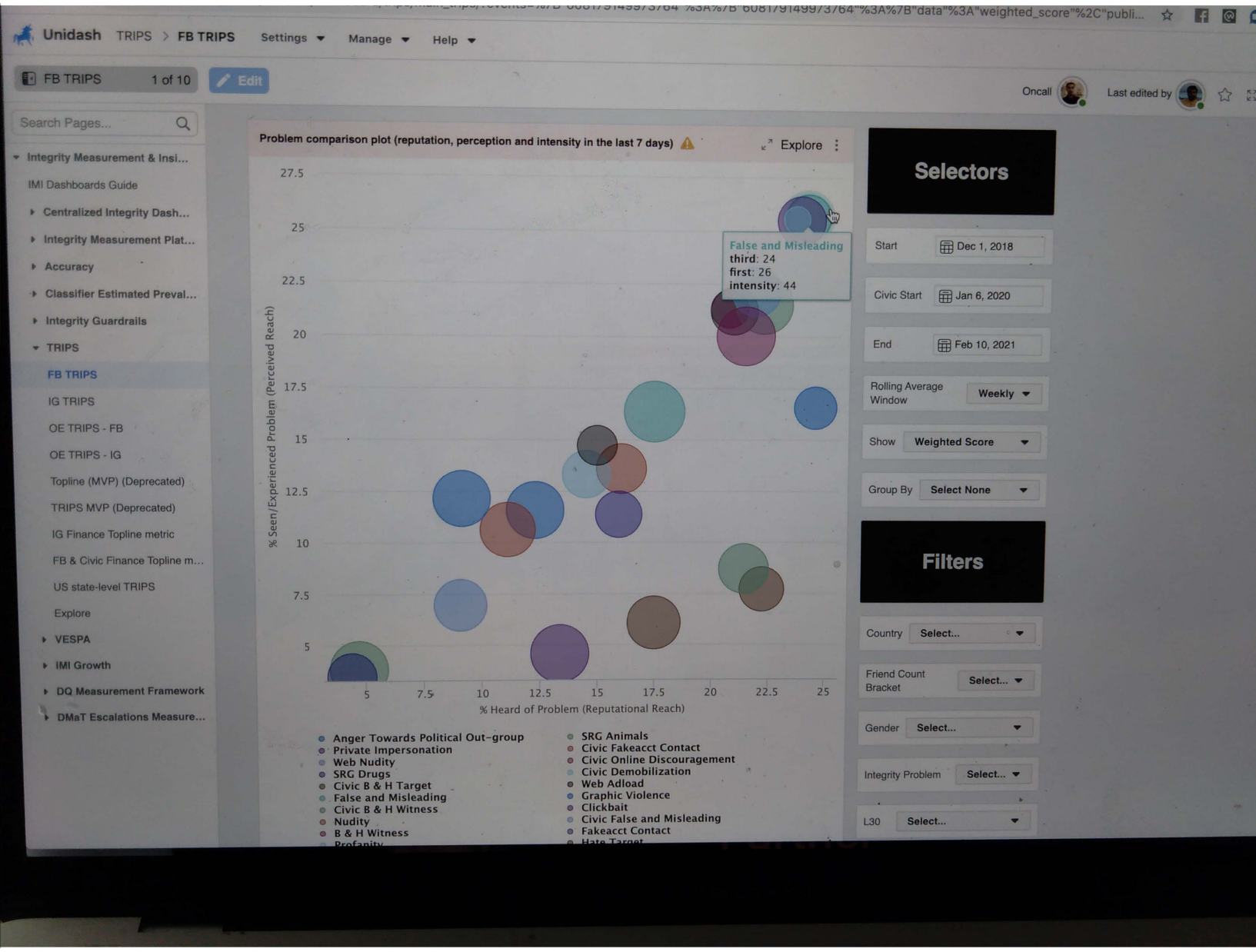

Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:42 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Reputational Reach
Intensity
Nudity
10.63
-2.66% d/d
-2.30% w/w
+0.47% m/m
11.07
-1.95% d/d
-2.81% w/w
+5.53% m/m
53.57
+0.21% d/d
-2.72% w/w
+7.55% m/m
Private Impersonation
4.57
+3.39% d/d
+1.11% w/w
-2.56% m/m
13.36
-0.15% d/d
-8.37% w/w
-2.05% m/m
57.80
-2.28% d/d
+10.01% w/w
-1.55% m/m
Profanity
25.53
-1.50% d/d
+0.43% w/w
+2.70% m/m
23.60
-0.21% d/d
+3.51% w/w
+2.92% m/m
37.90
0.00% d/d
-1.04% w/w
+3.61% m/m
SRG Animals
3.84
+2.95% d/d
+9.09% w/w
+11.30% m/m
4.69
+1.96% d/d
-2.90% w/w
+7.82% m/m
57.10
+2.11% d/d
+1.95% w/w
+7.05% m/m
Jan '20 May '20 Sep '20 Jan '21
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93

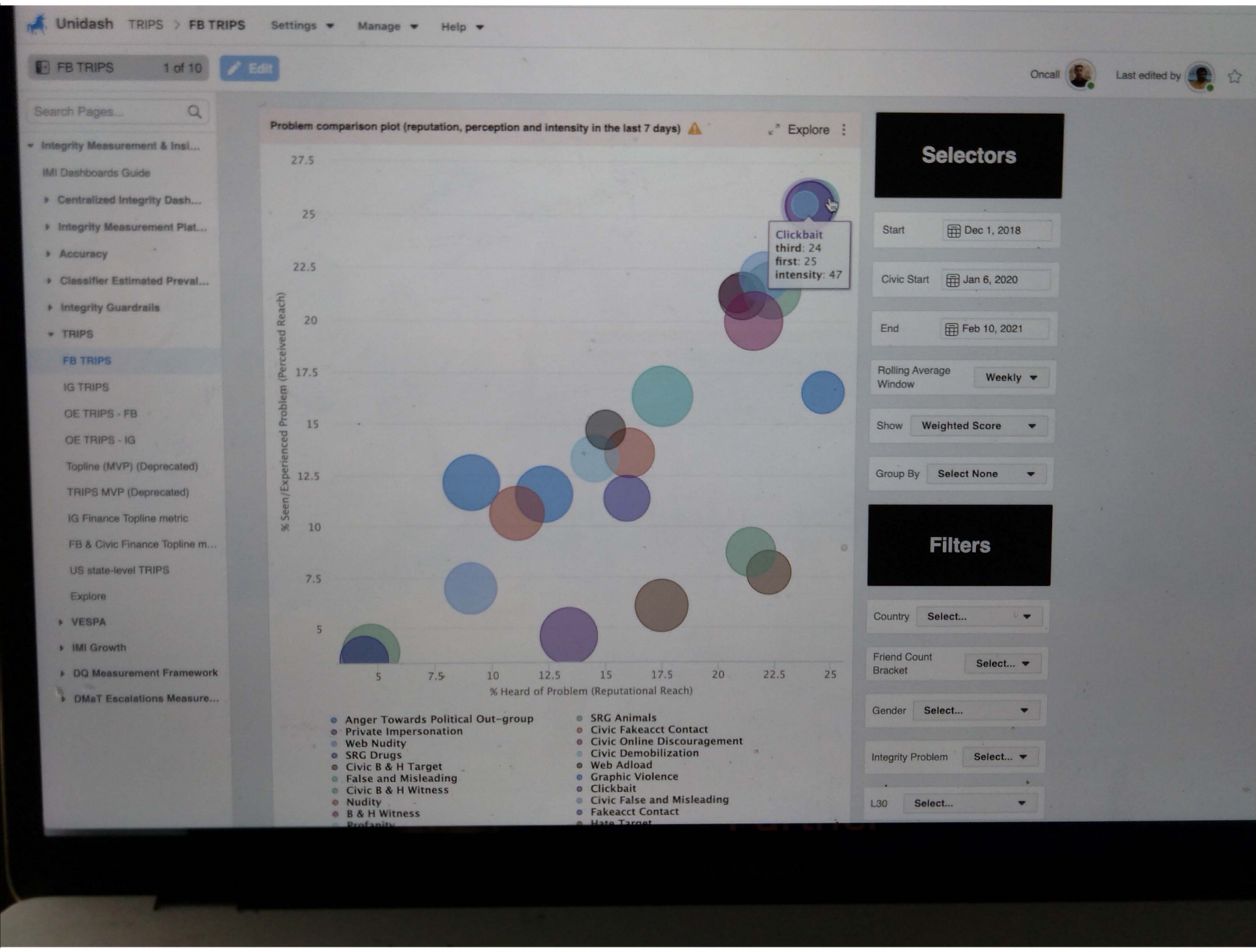

Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
Problem
Perceived Reach
Intensity
Nudity
10.63
-2.66% d/d
-2.30% w/w
+0.47% m/m
-1.95% d/d
-2.81% w/w
+5.53% m/m
53.57
+0.21% d/d
-2.72% w/w
+7.55% m/m
Private Impersonation
4.57
+3.39% d/d
+1.11% w/w
-2.56% m/m
13.36
-0.15% d/d
-8.37% w/w
-2.05% m/m
57.80
-2.28% d/d
+10.01% w/w
-1.55% m/m
Profanity
25.53
-1.50% d/d
+0.43% w/w
+2.70% m/m
23.60
-0.21% d/d
+3.51% w/w
+2.92% m/m
37.90
0.00% d/d
-1.04% w/w
+3.61% m/m
SRG Animals
3.84
+2.95% d/d
+9.09% w/w
+11.30% m/m
4.69
+1.96% d/d
-2.90% w/w
+7.82% m/m
57.10
+2.11% d/d
+1.95% w/w
+7.05% m/m
Friday, Apr 3, 2020
series_1: 14.29
series_1_ci (range): 13.80 to 14.78
Jan '20 May '20 Sep '20 Jan '21
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Dalquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93


Chrome File Edit View History Bookmarks People Tab Window Help
Thu 3:43 PM
internalfb.com/intern/unidash/dashboard/trips/main_trips/?events=%7B"608179149973764"%3A%7B"608179149973764"%3A%7B"data"%3A"weighted_score"%2C"publi...
Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
↑ Problem
Perceived Reach
Reputational Reach
Intensity
SRG Drugs
3.45
+1.17% d/d
+8.49% w/w
+6.48% m/m
4.38
-1.79% d/d
+3.06% w/w
+5.04% m/m
47.98
+2.65% d/d
-7.85% w/w
-4.78% m/m
Web Adload
21.20
+0.14% d/d
0.00% w/w
+0.52% m/m
20.94
+0.19% d/d
-2.20% w/w
-1.78% m/m
46.56
-0.47% d/d
-0.70% w/w
+6.04% m/m
Web Low Quality
14.74
-1.99% d/d
-1.80% w/w
+1.45% m/m
14.99
-1.38% d/d
-5.19% w/w
-0.93% m/m
41.34
+4.92% d/d
+2.43% w/w
+0.61% m/m
Web Nudity
6.93
-2.53% d/d
-6.10% w/w
+1.76% m/m
9.08
-2.99% d/d
-9.47% w/w
+2.83% m/m
51.46
-1.66% d/d
-3.25% w/w
-0.58% m/m
Jan '20 Apr '20 Jul '20 Oct '20 Jan '21
Jan '20 May '20 Sep '20 Jan '21
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93

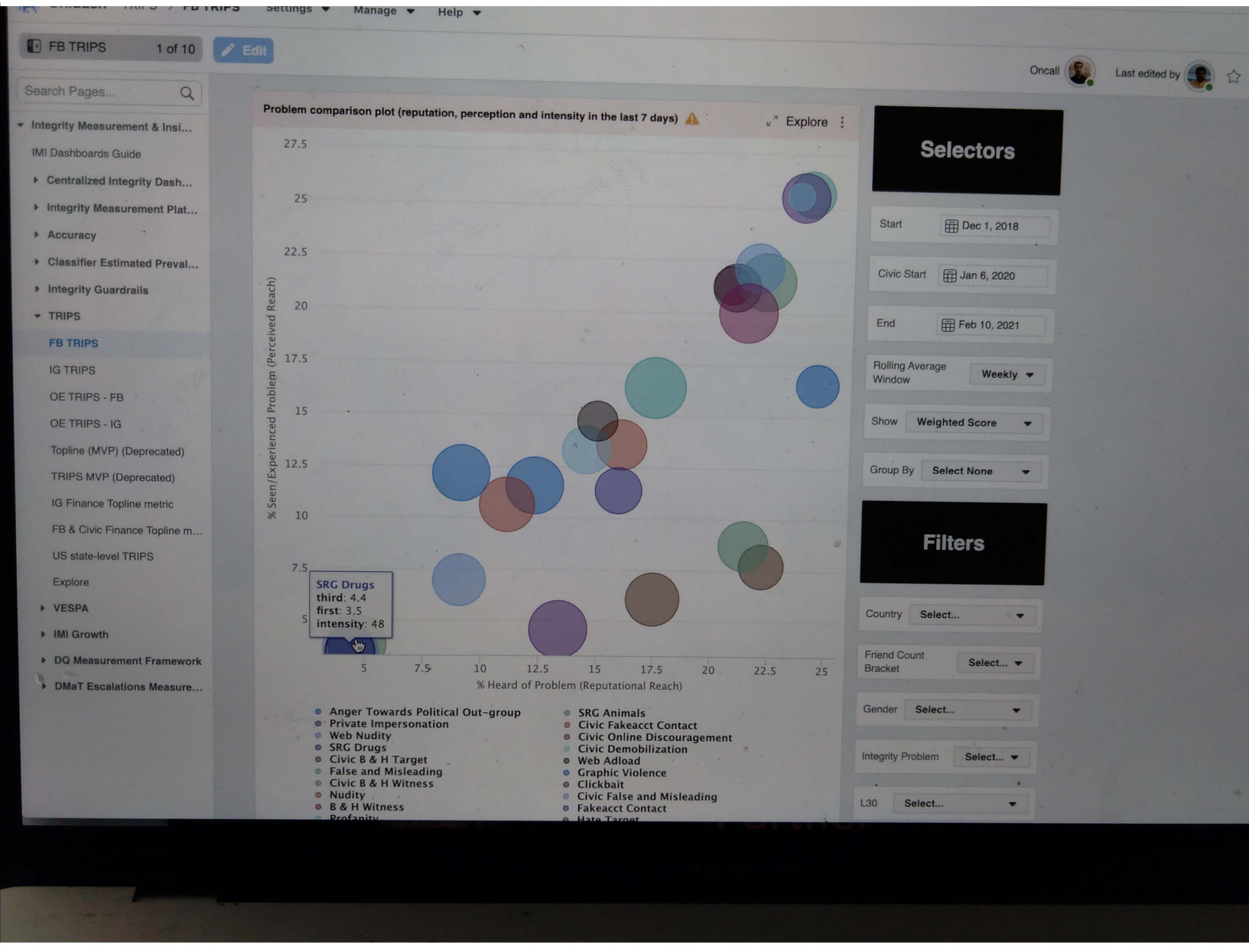

Unidash TRIPS > FB TRIPS Settings Manage Help
Trends in perceived reach Reputational reach and intensity for each Integrity Problem
Problem
Perceived Reach
Reputational Reach
Intensity
Click on top of each column to rank integrity problems from highest to lowest perceived reach, reputational reach, or intensity
SRG Drugs
3.45
+1.17% d/d
+8.49% w/w
+6.48% m/m
4.38
-1.79% d/d
+3.06% w/w
+5.04% m/m
47.98
+2.65% d/d
-7.85% w/w
-4.78% m/m
Web Adload
21.20
+0.14% d/d
0.00% w/w
+0.52% m/m
20.94
+0.19% d/d
-2.20% w/w
-1.78% m/m
46.56
-0.47% d/d
-0.70% w/w
+6.04% m/m
Web Low Quality
14.74
-1.99% d/d
-1.80% w/w
+1.45% m/m
14.99
-1.38% d/d
-5.19% w/w
-0.93% m/m
41.34
+4.92% d/d
+2.43% w/w
+0.61% m/m
Web Nudity
6.93
-2.53% d/d
-6.10% w/w
+1.76% m/m
9.08
-2.99% d/d
-9.47% w/w
+2.83% m/m
51.46
-1.66% d/d
-3.25% w/w
-0.58% m/m
Table
Data
Visualize
Query 1
+ New Query
Daiquery
Data Transforms
Range Filters
Add
Category Filters
Add
Columns
problem
Add column
Group By
Add column
Pivots
Add column
Metrics
TIME_SERIES(first)
TIME_SERIES(third)
TIME_SERIES(intensity)
Add column aggregation
Column Filters
Display Top N
Add
Export
Copy URL
Go to Query Source
7.68
22.22
44.93
esc
Search Bunny

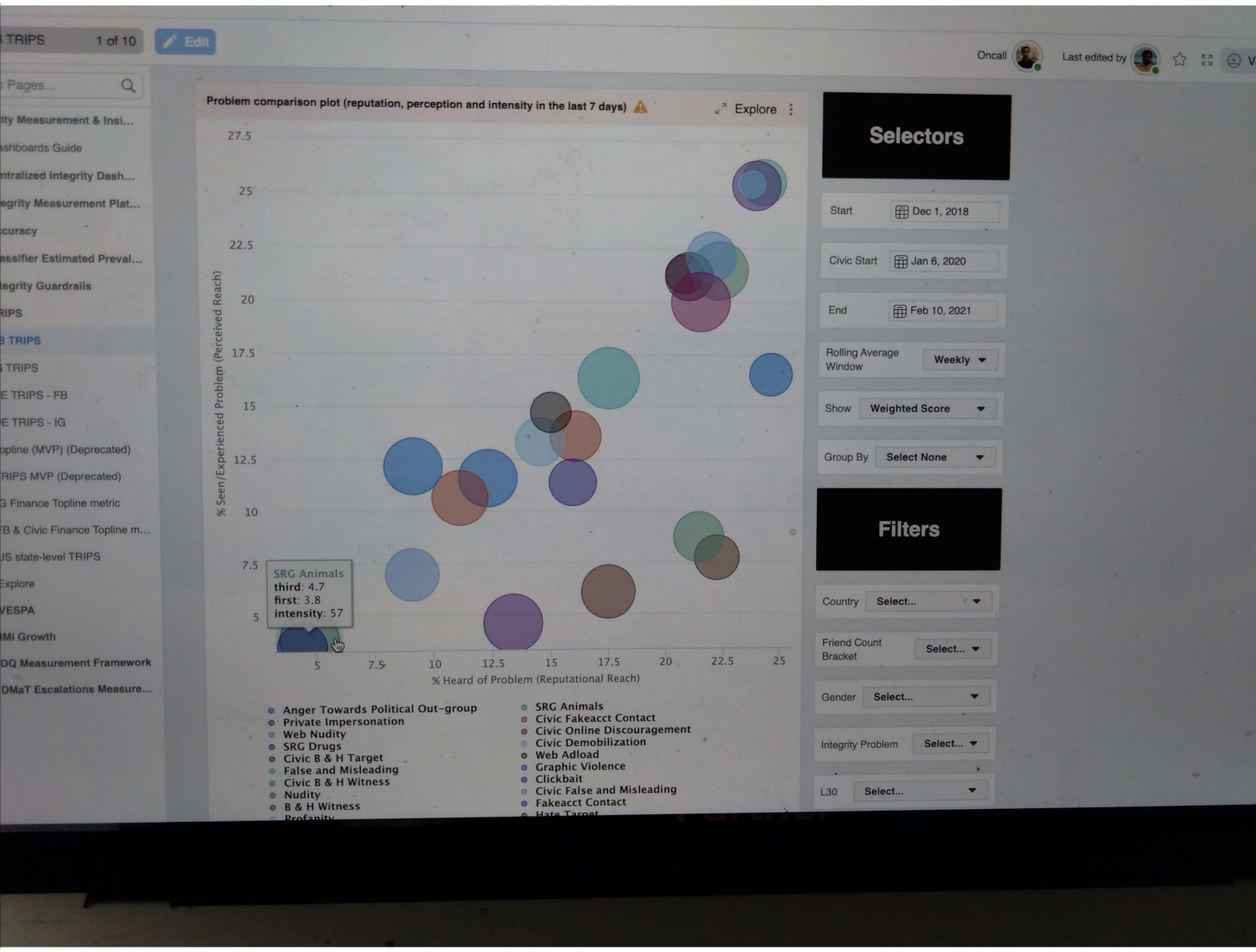

Unidash
TRIPS > FB TRIPS
Settings
Manage
Help
FB TRIPS 1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Selectors
Start
Civic Start
End
Rolling Average Window
Show
Group By
Filters

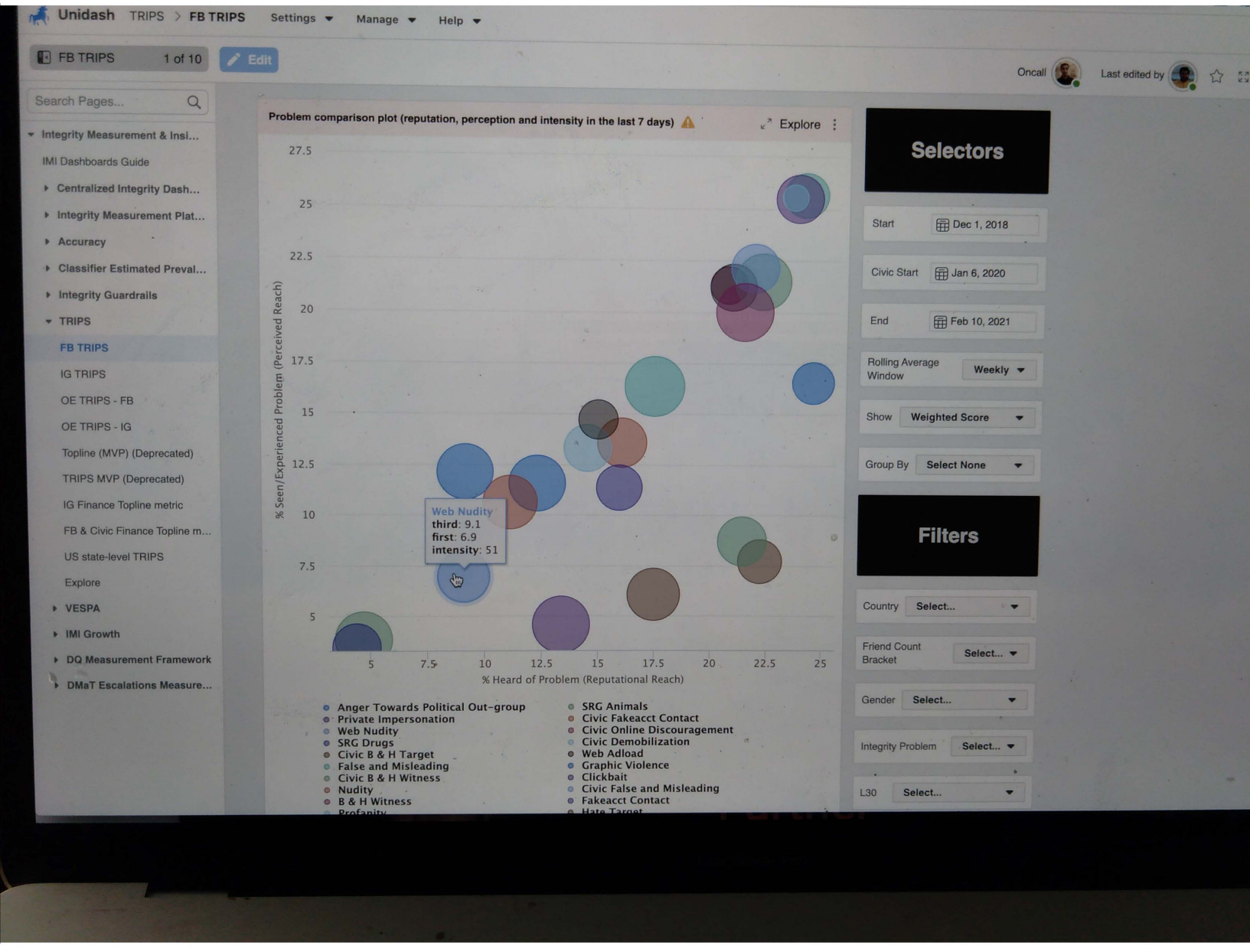

Unidash
TRIPS > FB TRIPS
Settings
Manage
Help
FB TRIPS 1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
False and Misleading
third: 24
first: 26
intensity: 44
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters
Country
Select...
Friend Count Bracket
Select...
Gender
Select...
Integrity Problem
Select...
L30
Select...


Unidash
TRIPS > FB TRIPS
Settings
Manage
Help
FB TRIPS 1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
Clickbait
third: 24
first: 25
intensity: 47
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters
Country
Select...
Friend Count Bracket
Select...
Gender
Select...
Integrity Problem
Select...
L30
Select...


FB TRIPS
1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
Profanity
third: 24
first: 26
intensity: 38
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational
Anger Towards Political Out-group
Private Impersonation
Nudity
B & H Witness
Fakeacct Contact
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters


Settings
Manage
Help
FB TRIPS
1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
5
7.5
10
12.5
15
17.5
20
22.5
25
SRG Drugs
third: 4.4
first: 3.5
intensity: 48
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
Profanity
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Hate Target
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters
Country
Select...
Friend Count Bracket
Select...
Gender
Select...
Integrity Problem
Select...
L30
Select...


TRIPS
1 of 10
Edit
Oncall
Last edited by
Pages...
ity Measurement & Insi...
shboards Guide
ntralized Integrity Dash...
egrity Measurement Plat...
curacy
assifier Estimated Preval...
tegrity Guardrails
RIPS
B TRIPS
TRIPS
E TRIPS - FB
E TRIPS - IG
opline (MVP) (Deprecated)
RIPS MVP (Deprecated)
G Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
MI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
SRG Animals
third: 4.7
first: 3.8
intensity: 57
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters
Country
Select...
Friend Count Bracket
Select...
Gender
Select...
Integrity Problem
Select...
L30
Select...


Unidash
TRIPS > FB TRIPS
Settings
Manage
Help
FB TRIPS 1 of 10
Edit
Oncall
Last edited by
Search Pages...
Integrity Measurement & Insi...
IMI Dashboards Guide
Centralized Integrity Dash...
Integrity Measurement Plat...
Accuracy
Classifier Estimated Preval...
Integrity Guardrails
TRIPS
FB TRIPS
IG TRIPS
OE TRIPS - FB
OE TRIPS - IG
Topline (MVP) (Deprecated)
TRIPS MVP (Deprecated)
IG Finance Topline metric
FB & Civic Finance Topline m...
US state-level TRIPS
Explore
VESPA
IMI Growth
DQ Measurement Framework
DMaT Escalations Measure...
Problem comparison plot (reputation, perception and intensity in the last 7 days)
Explore
% Seen/Experienced Problem (Perceived Reach)
% Heard of Problem (Reputational Reach)
Web Nudity
third: 9.1
first: 6.9
intensity: 51
Anger Towards Political Out-group
Private Impersonation
Web Nudity
SRG Drugs
Civic B & H Target
False and Misleading
Civic B & H Witness
Nudity
B & H Witness
SRG Animals
Civic Fakeacct Contact
Civic Online Discouragement
Civic Demobilization
Web Adload
Graphic Violence
Clickbait
Civic False and Misleading
Fakeacct Contact
Selectors
Start
Dec 1, 2018
Civic Start
Jan 6, 2020
End
Feb 10, 2021
Rolling Average Window
Weekly
Show
Weighted Score
Group By
Select None
Filters
Country
Select...
Friend Count Bracket
Select...
Gender
Select...
Integrity Problem
Select...
L30
Select...

This post is a problem

Arturo, Jake, Pete, Emma, Andy, Diane,
Josh, Charles, Travis, Tijana

Marc A. Brackett, Robin S. Stern, & Andrés Richner
Health, Emotion, & Behavior Laboratory
Yale University

Relationships Matter!

Why are we here?

Bullying is a real problem!

- *Definition:* An intentional act of aggression, based on an imbalance of power, that is meant to harm a victim either psychologically or physically. Bullying usually occurs repeatedly and over time, but sometimes can be identified in a single event.
- Over 50% of kids say someone has said mean or hurtful things to them online.
- Over 50% of kids admit saying something mean to another person online.

We have a responsibility to provide students with tools so they can be both psychologically and physically safe online and in everyday life.

Overview

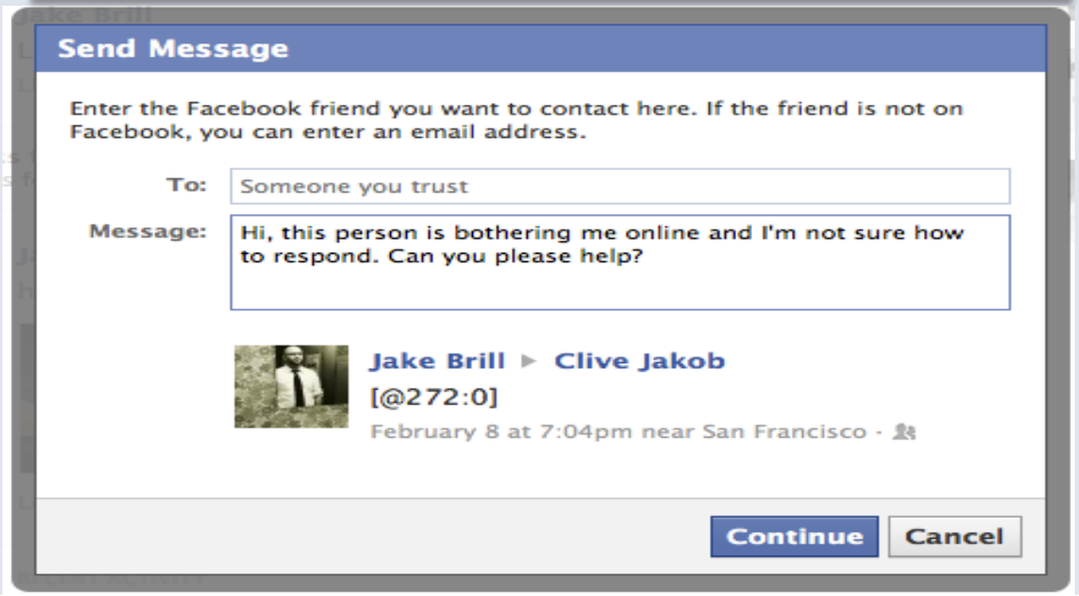
- Forming the Facebook/Yale relationship
- The original report flow
- Ideas for improving the report flow
- Methods for developing new flow
- The new report flow (v1.1)
- What do the data say?
- Next steps

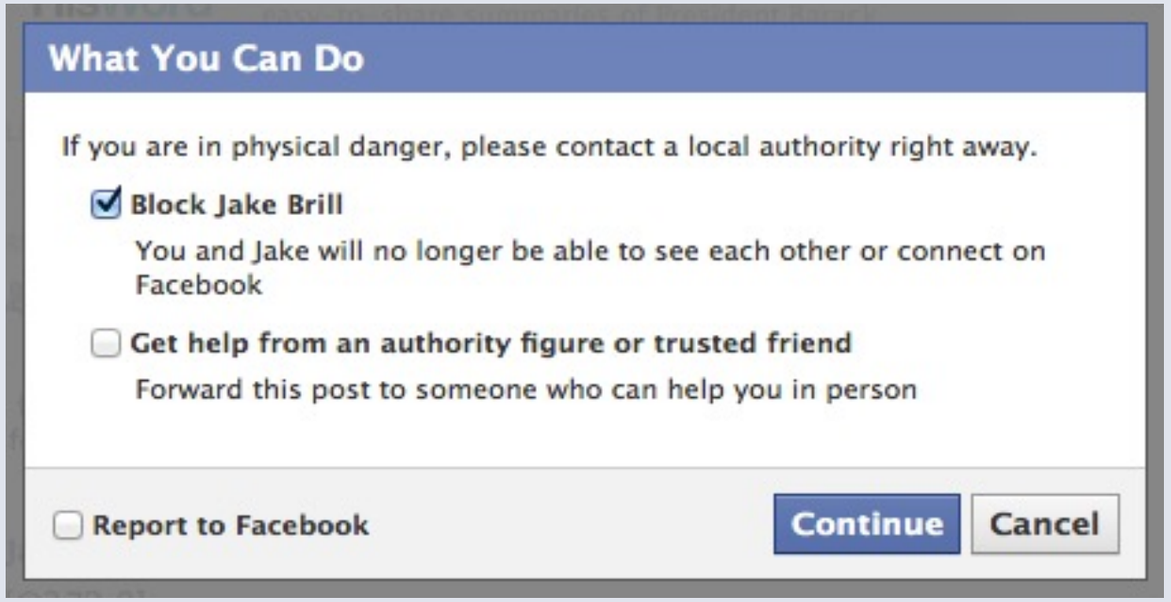
Facebook / Yale

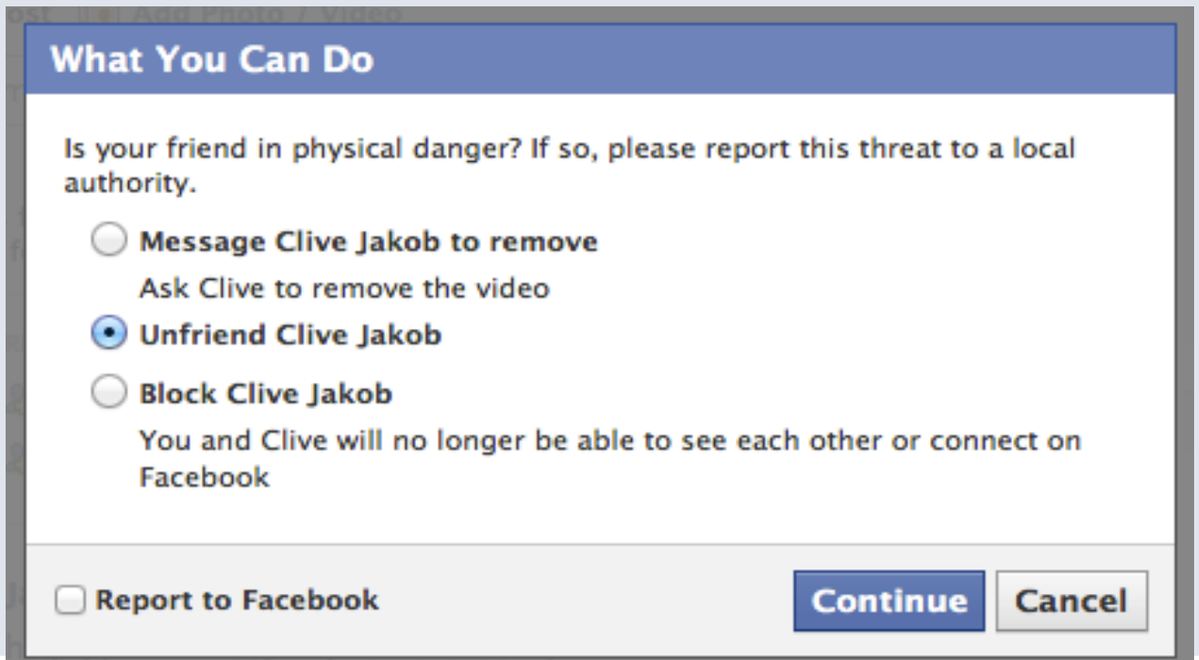
- Compassion Day 1

- Initial conversations about Facebook’s needs and what the Yale team could provide

Original Report Flow







Ideas for improving report flows

- **Infuse developmental science**
  - 13/14 year olds are different from high school and college students
- **Use more kid-friendly language**
  - “Report” vs. “This post is a problem”
- **Enhance logic of the flow**
  - ‘What happened?’ to ‘how are you feeling?’ to ‘what can you do?’
- **Differentiate the experience so we could tailor support**
  - Move from just “harassing me” to real experiences of this age group
- **Empower youth to take a positive and safe action**
  - Provide simple, effective guidance (e.g., “don’t be alone with this person”)
- **Help youth to get more help from their community**
  - Encourage kids to reach out to a trusted adult

Methods for developing new flows

- **Iterative process between Yale Team and Facebook Team:**
  - Review of existing research
  - Focus groups with diverse students in public and private schools
  - Interviews with children who experienced cyberbullying
  - Interviews with parents, school principals, teachers, and counselors
  - Integration of best clinical practices
  - Taking Facebook design into consideration (e.g., writing, editing, and making sure we got everything we possibly could into the limited space).

Focus Groups and Interviews

- **Participants**
  - Public and private school students (N = 50; 13 to 15 year olds; 8 groups total) from diverse backgrounds (east and west coast)

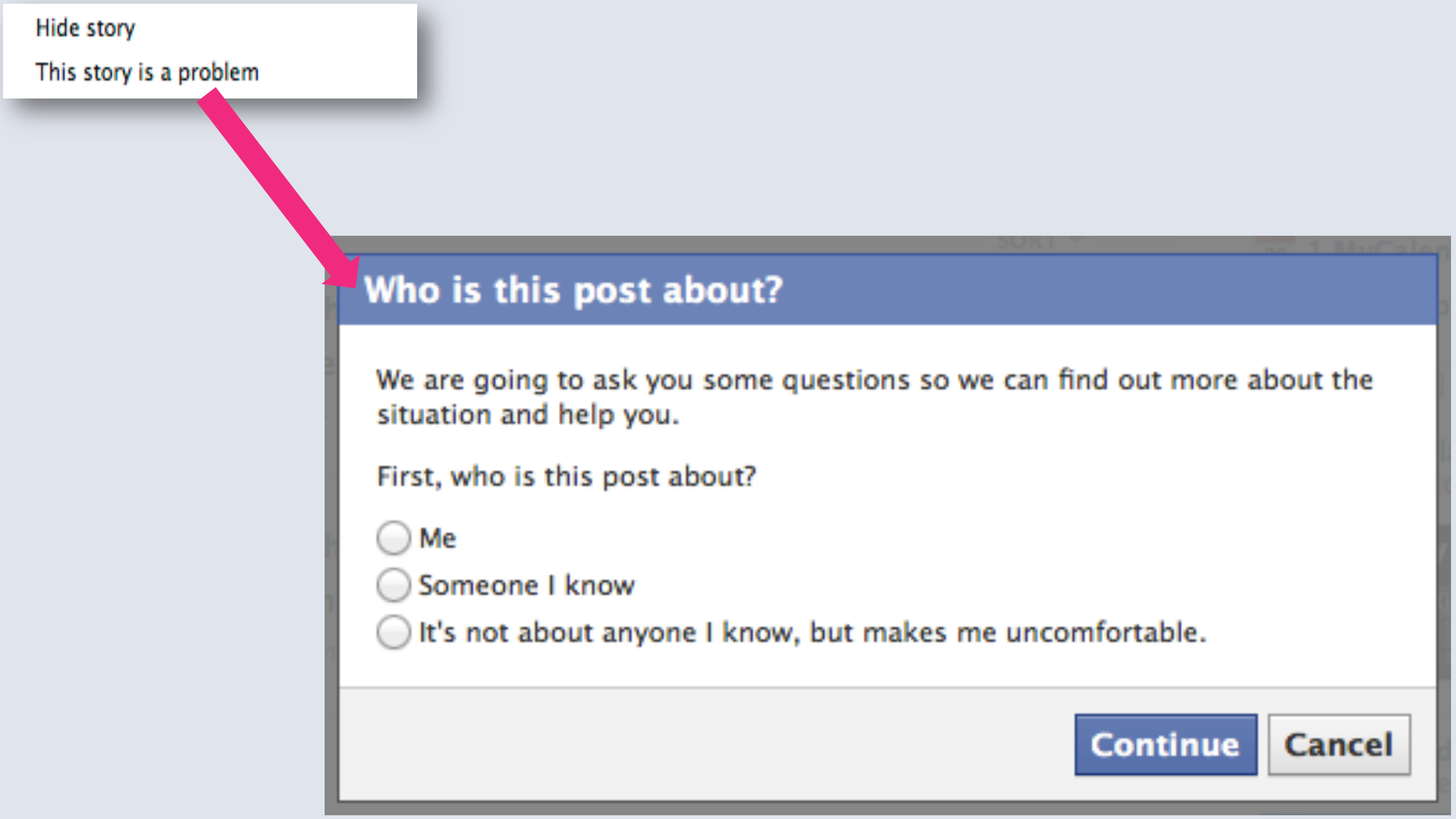
- **Takeaways from focus groups and interviews**
  - Kids were particular about the language we used
    - E.g., *report* – meant ‘authority’ or ‘trouble’ or ‘evaluated,’ ‘get help’ suggested ‘technical problem’
  - Kids helped us to differentiate the bullying experiences
  - Kids wanted Facebook to do something about it, but were not sure what
  - If questions were meaningful and kids believed they would be helpful, they would be more motivated to complete the flow
  - Kids said they wanted help crafting messages
  - Kids said not everything needed to be reported b/c they would just tell their friend…

Focus Groups and Interviews

- **Takeaways from interviews with parents**
  - Parents were mixed on whether they should be the trusted adults
  - (Some) parents enabled kids to fake their age
  - Parents wanted more resources for their kids to get help

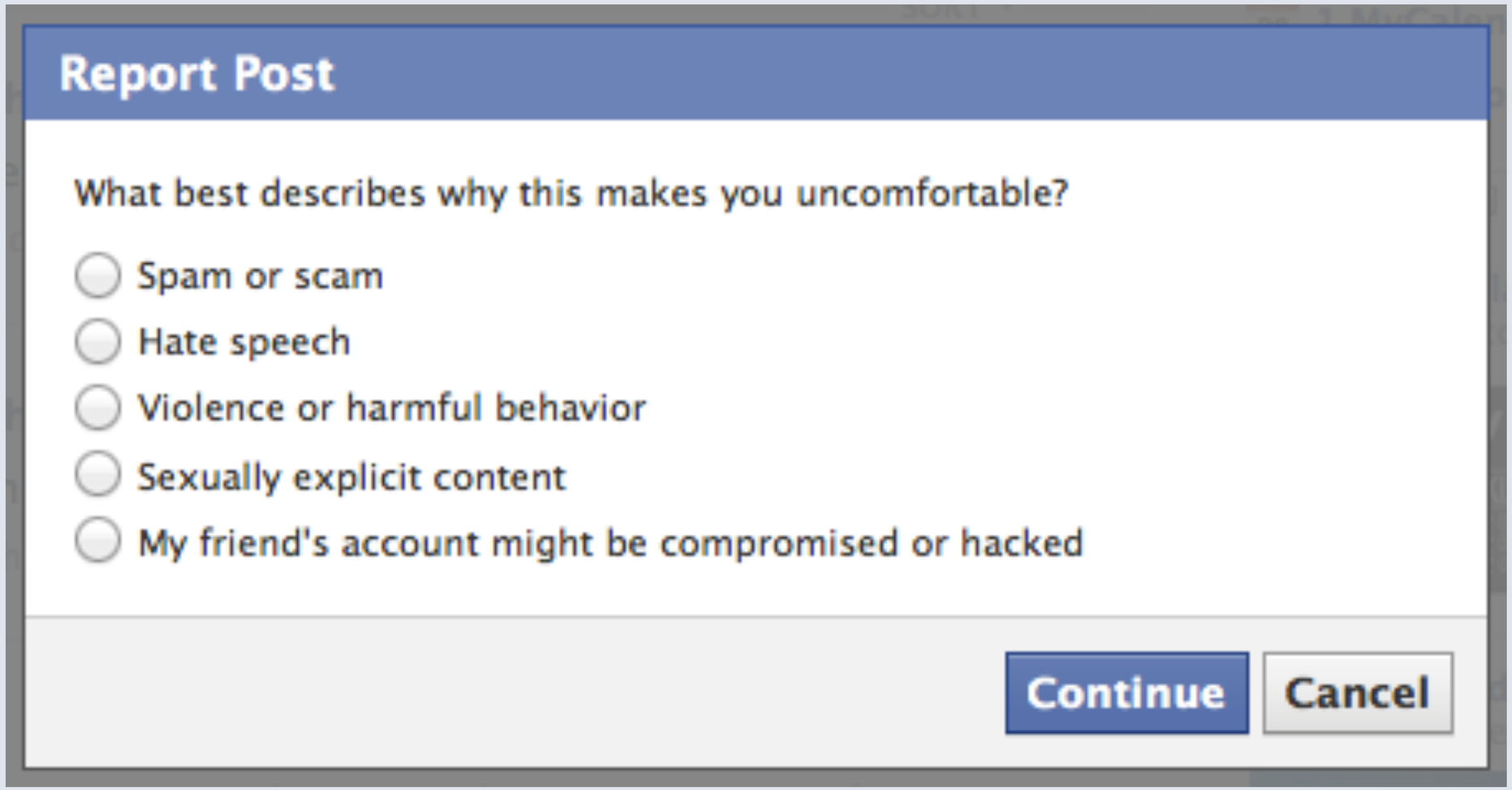
- **Our own takeaway**
  - Had to be a balance between what kids wanted and what we believed they need
    - Threatened – may not want to tell trusted adult, but they need help
  - We needed to provide children with more direct help

The new Report Flow (v 1.1)



The new Report Flow (v 1.1)

If “makes me uncomfortable,” go into Community Standards flow

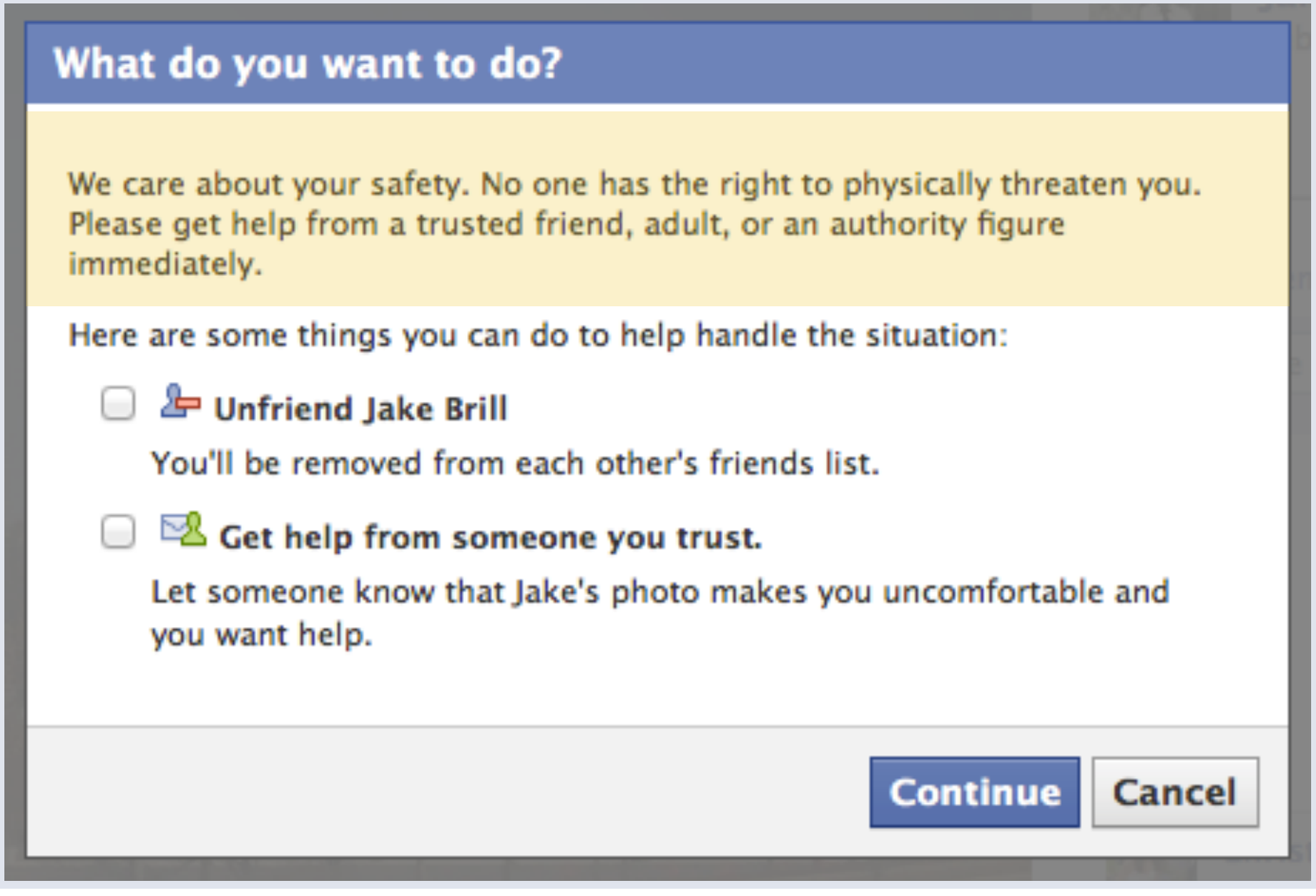


The new Report Flow (v 1.1)

However, if "about me" or "someone I know," go into…

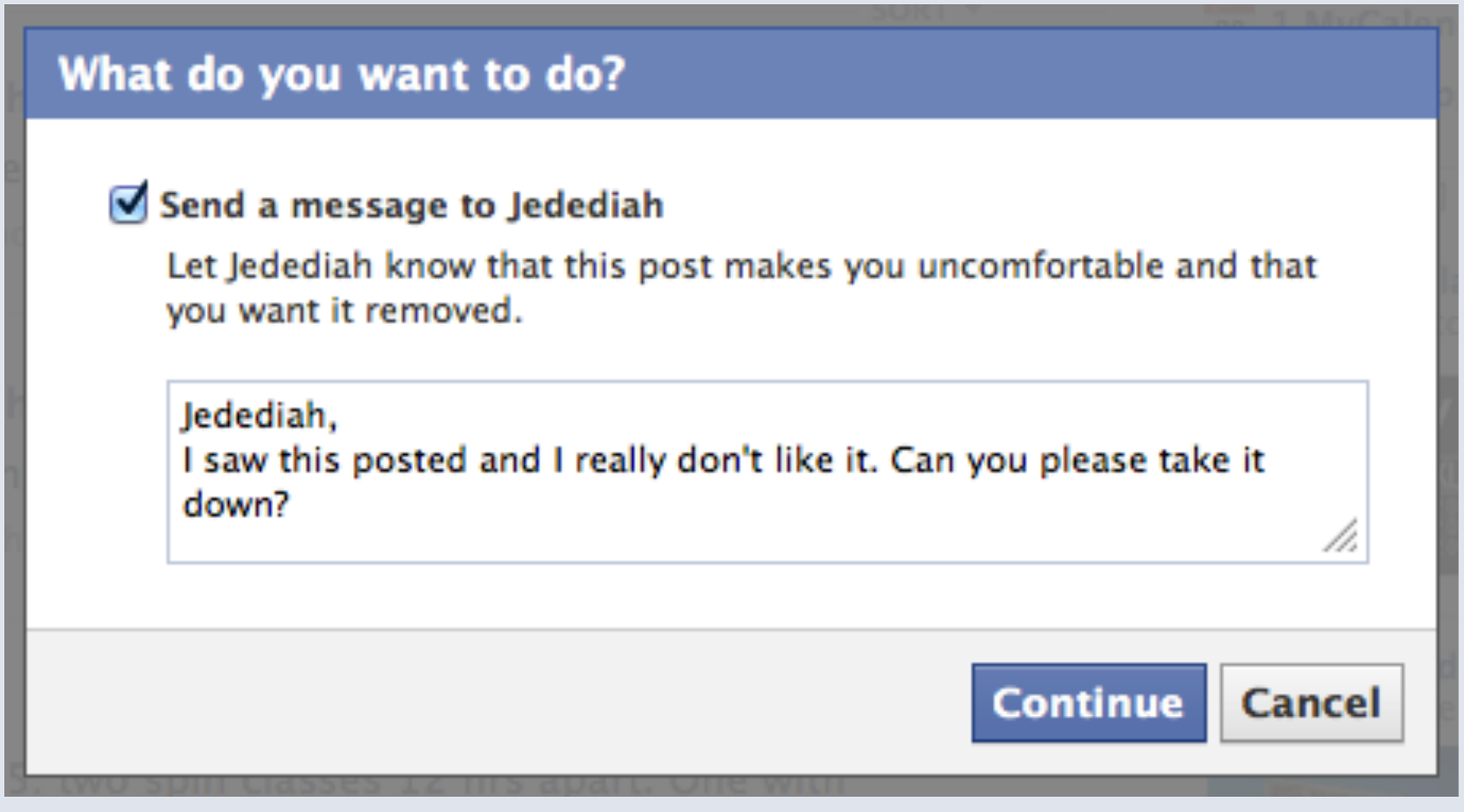
The new Report Flow (v 1.1)

If “threatening,” we lead to social resolution with extra messaging about safety:



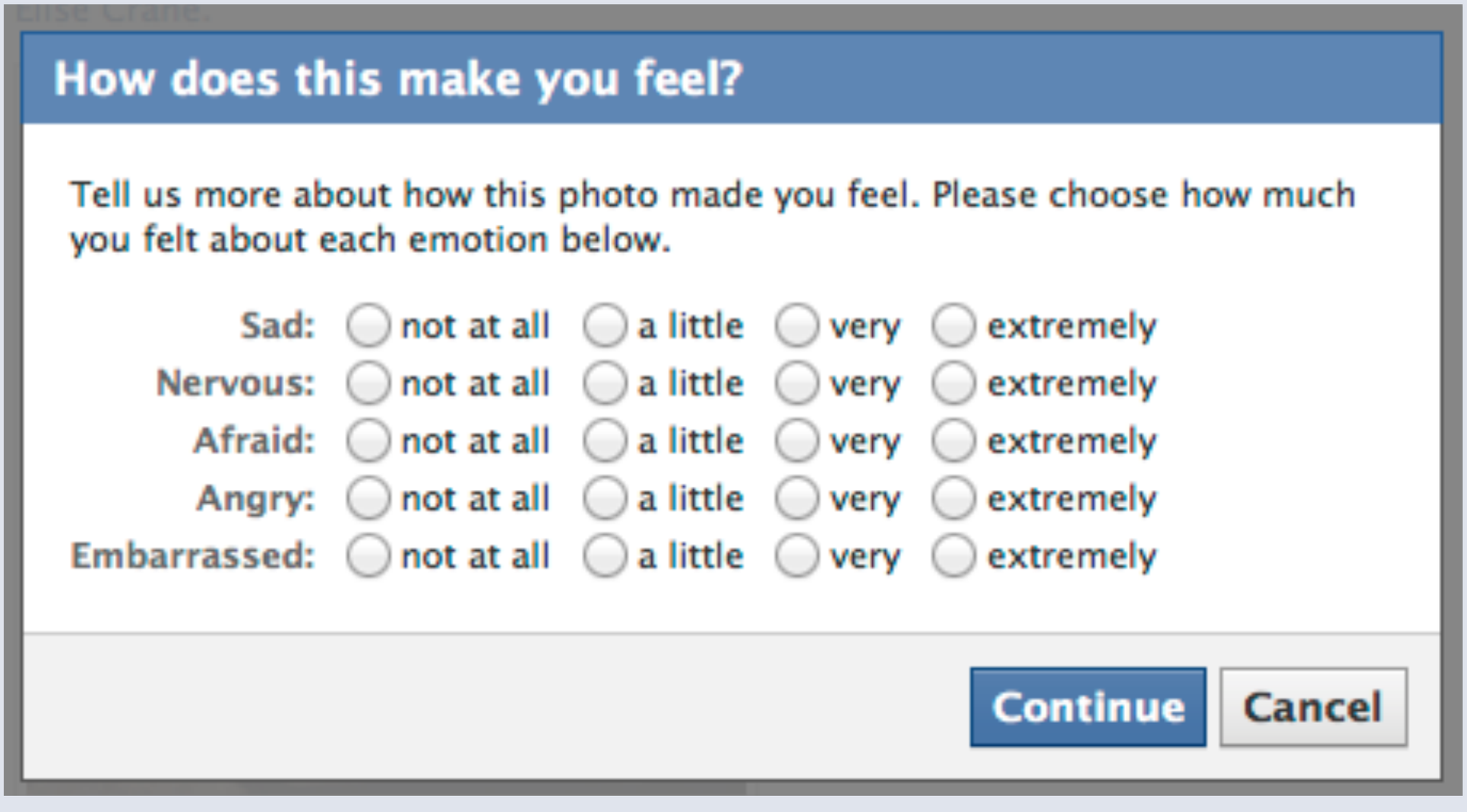
The new Report Flow (v 1.1)

If “posted something I don’t like,” send message with pre-populated text:



The new Report Flow (v 1.1)

Other options (e.g.,“uncomfortable” / “said mean things” / “won’t leave me alone”) lead to emotions slide:

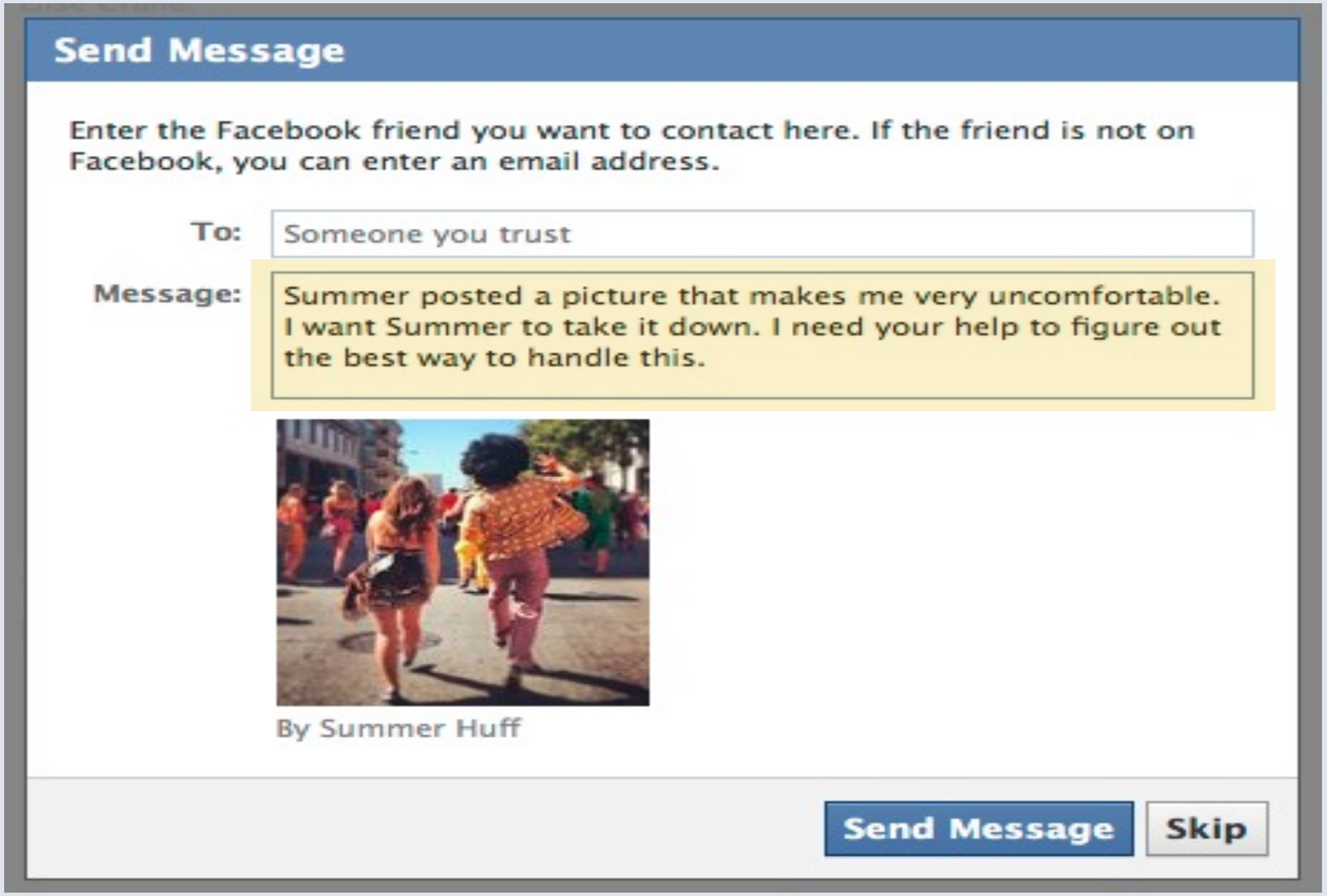


The new Report Flow (v 1.1)

After identifying emotion, lead to social resolution with text/options that vary as a function of the situation and intensity of emotion:

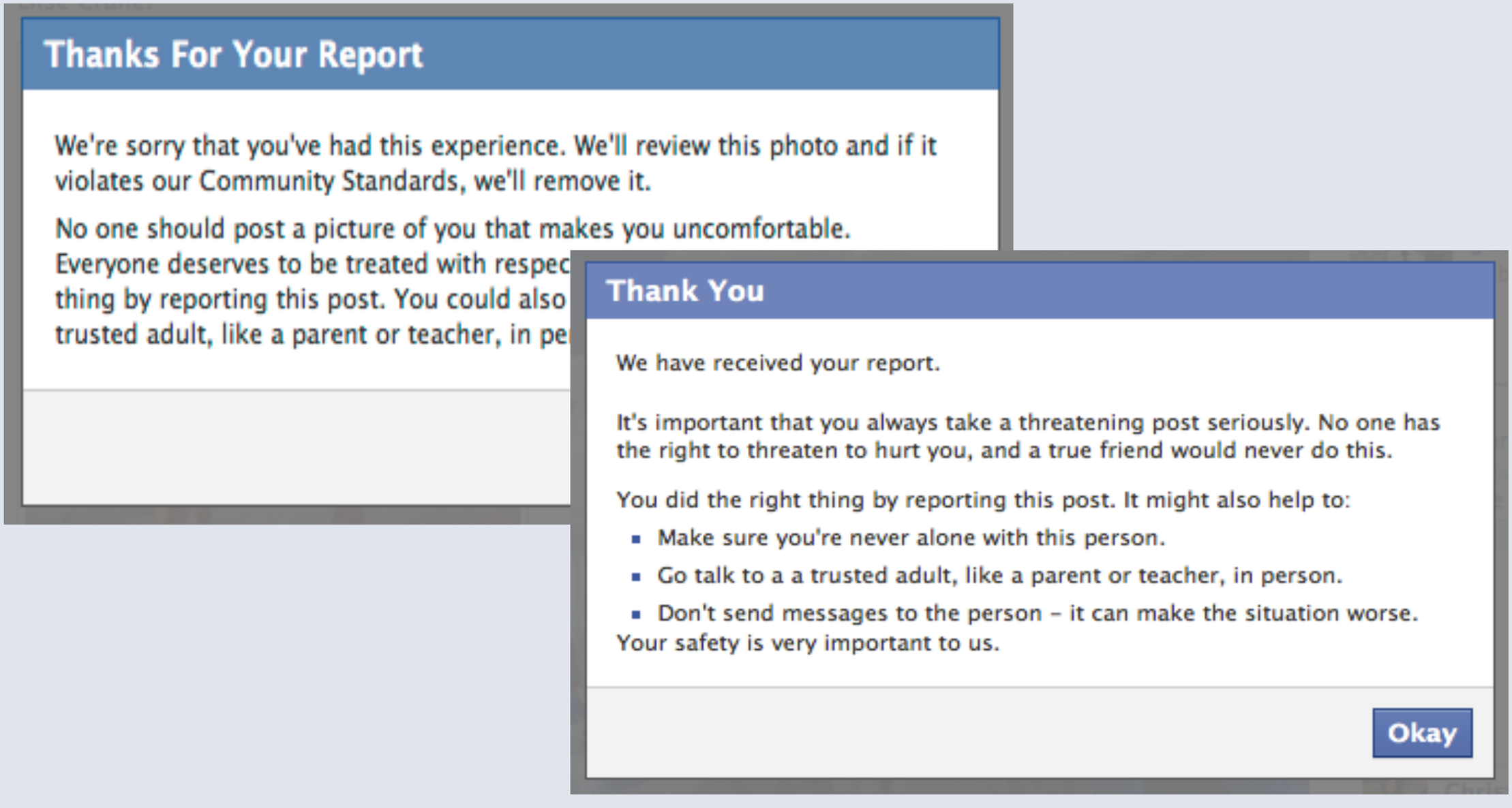
The new Report Flow (v 1.1)

Similarly, the option to message someone you trust is pre-populated with text that also varies as a function of the situation and intensity of emotion:



The new Report Flow (v 1.1)

Thank you slides are differentiated by experience



What do the data say?

(Female/male)

"The post is a problem" 5/25-6/03
23600 (unique users)
66%/34%

'Me'
62%
70%/30%

'Uncomfortable'
17%
62%/38%

'Someone I know'
6%
62%/38%

Don't like
53%
72%/28%

Photo
13%
70%/30%

Said mean
4%
60%/40%

Won't leave alone
7%
60%/40%

Threat
3%
55%/45%

Msg
59%

Unfriend
2%

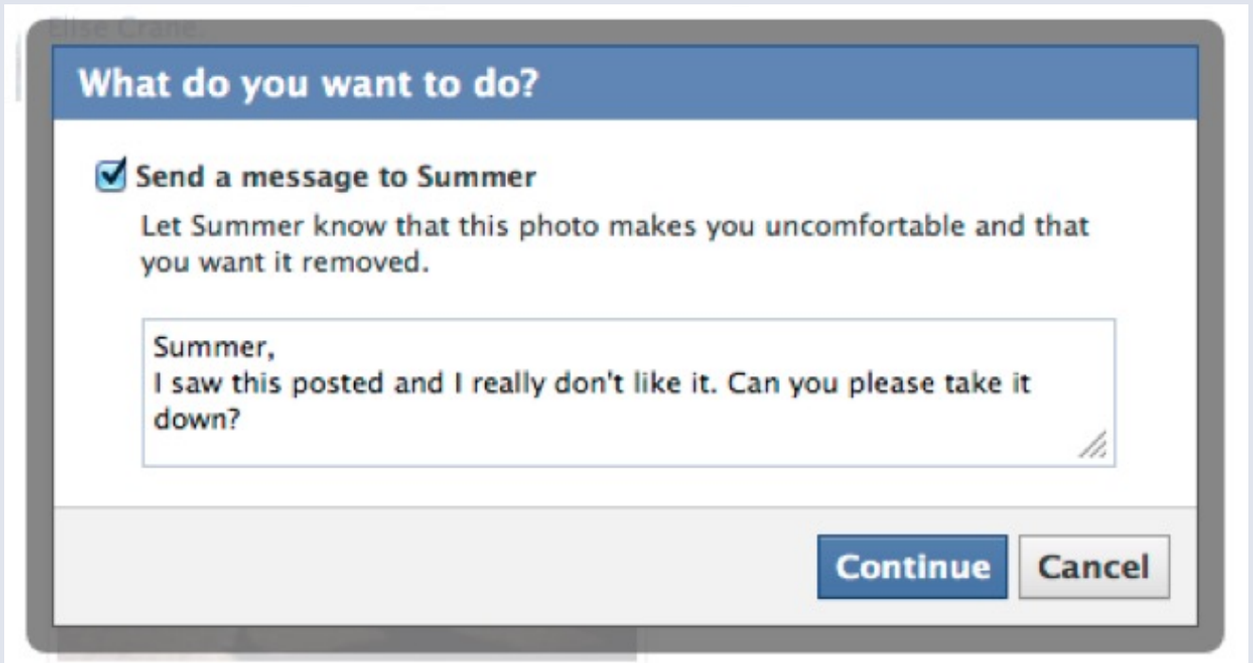
Block
6%

Tr Msg
10%

Category breakdown

Area	Sub-categories
Just don't like (77%)	• Photos: Awk pics, screenshots, vs photos, tag besties, spam • Text: call out person, relationship post, tag besties
Posted a photo that makes me uncomfortable (17%)	• Mostly bad (candids, funny face) • Screenshots, porn, relationship, making fun
Said mean things (5%)	• Photos: Tag besties, political, vs photos, screenshots, joking/mean comments • Text: family conflict, fights, passive aggressive posts
Won't leave me alone (9%)	• Screenshots, vs photos, spam • 'Pestering' as opposed to 'stalking'
Threatening (4%)	• Photos: Bad photos, vs photos, screenshots, spam • Text: rating girls (top ten), harassment

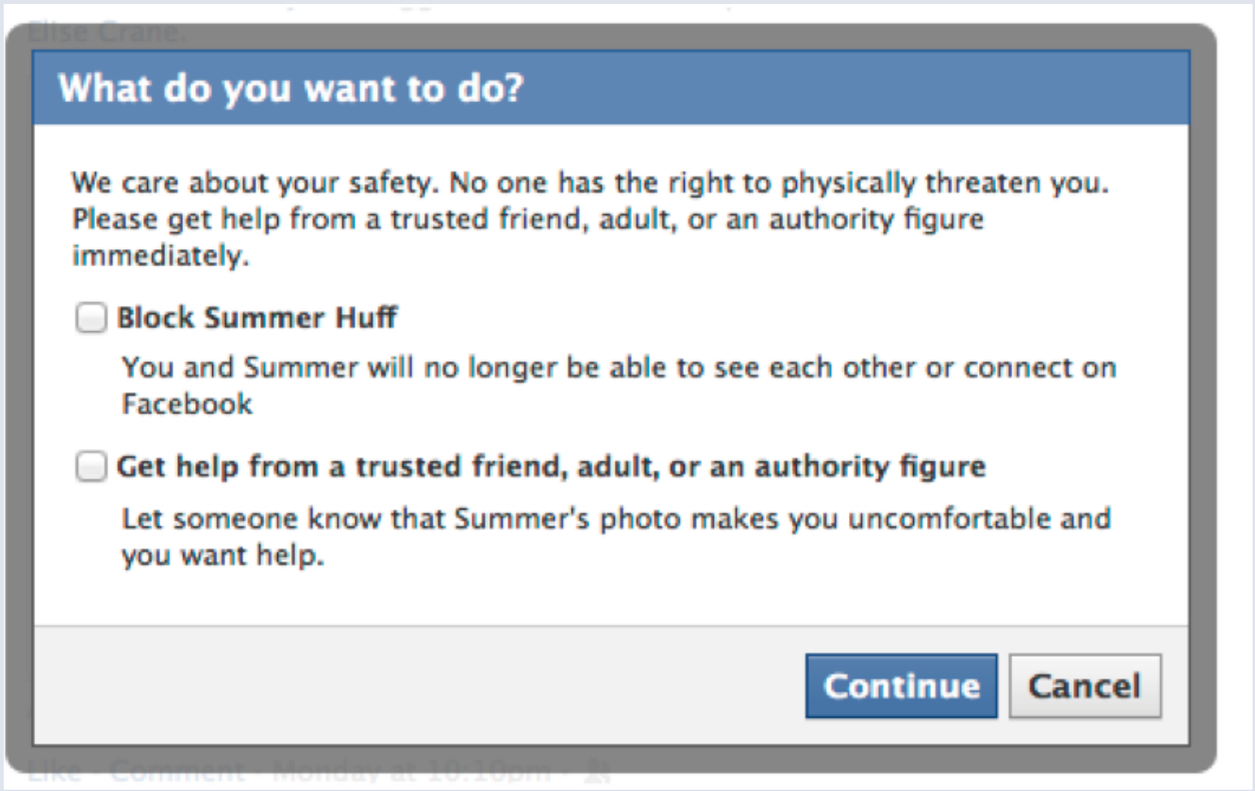
The new Report Flow (v 1.1)



For those who picked 'posted something that I just don't like'

- **60% send msg**

The new Report Flow (v 1.1)



For those who pick 'threatened to hurt me':

- **What % block: 6%**
- **What % unfriend: 3%**
- **What % choose trusted msg: 11%**
  - **What % end up sending msg: 14% (2% overall)**
- **What % cancel: 22%**
- **What % choose no option: 27%**
- **What % navigate away: 31%**

The new Report Flow (v 1.1)

- **For users who picked ‘said mean things to me’ / ‘won’t leave me alone’ / ‘posted a photo that makes me uncomfortable’**
  - **What % completed overall: 85%**
  - **What % completed that were forced: 96%**
  - **What % completed that were unforced: 73%**

Distribution of emotions

	Sad	Nervous	Afraid	Angry	Embarrassed
No answer	27%	30%	30%	24%	25%
Not at all	42%	42%	46%	35%	30%
A little	11%	10%	9%	12%	16%
Very	5%	5%	4%	10%	8%
Extremely	13%	12%	10%	19%	21%

‘said mean things’

	Sad	Nervous	Afraid	Angry	Embarrassed
No answer	28%	30%	30%	24%	26%
Not at all	26%	35%	36%	19%	24%
A little	14%	12%	11%	8%	11%
Very	9%	7%	6%	12%	10%
Extremely	24%	17%	16%	36%	30%

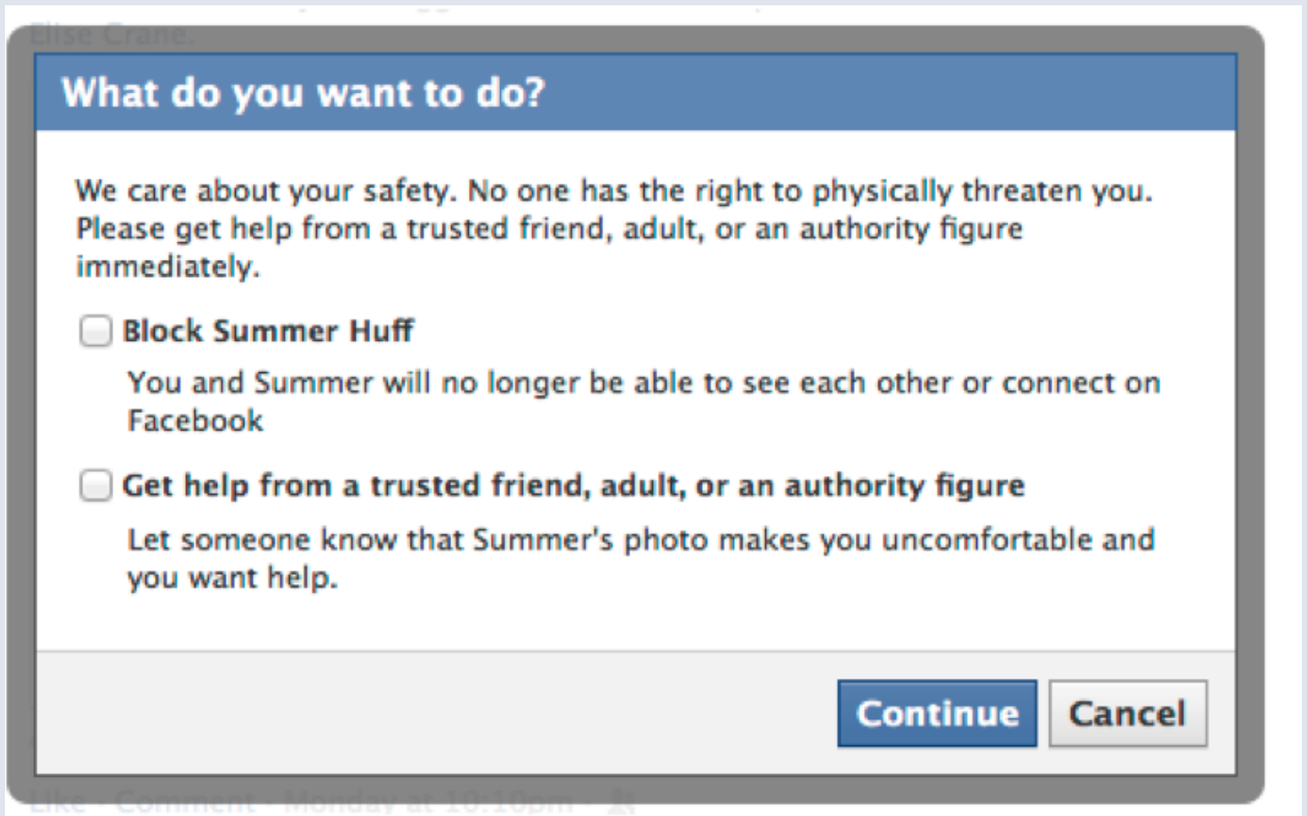
‘won’t leave me alone’

	Sad	Nervous	Afraid	Angry	Embarrassed
No answer	34%	40%	37%	32%	37%
Not at all	41%	40%	41%	33%	37%
A little	9%	7%	8%	11%	9%
Very	5%	3%	4%	8%	4%
Extremely	10%	9%	9%	16%	13%

‘posted a photo that makes me uncomfortable’

	Sad	Nervous	Afraid	Angry	Embarrassed
No answer	24%	26%	26%	22%	19%
Not at all	46%	44%	50%	40%	27%
A little	12%	12%	10%	14%	20%
Very	5%	5%	4%	7%	10%
Extremely	12%	12%	10%	17%	24%

The new Report Flow (v 1.1)



For those who complete emotion slide:

- **What % block: 7%**
- **What % unfriend: 6%**
- **What % choose trusted msg: 14%**
  - **What % end up sending msg: 24% (3% overall)**
- **What % cancel: 9%**
- **What % choose no option: 11%**
- **What % navigate away: 53%**

Comparing Old and New Flows

- **Were users more or less satisfied with the new report flow?**
  - **One concern was that kids would be less satisfied with the new flow compared to the old flow because the new flow was longer**
  - **There were no significant differences**

	New Flow	Old Flow
How easy?	1.89	1.92
How helpful?	2.23	2.18
How comfortable?	2.23	2.17
How satisfied?	2.19	2.22

Comparing Old and New Flows

- **Did we change actual behavior? YES!**
  - **Of those who completed the report (for more extreme instances), a greater number of users in the new flow reached out to a trusted adult**

	New Flow	Old Flow
Reaching out to trusted adult	**43%**	**19%**
Blocking	**28%**	**44%**

Comparing Old and New Flows

- **Two-days later: Was the trusted adult helpful?**

- **Trusted adults were perceived as being more helpful in the new versus old flows (Ns are small; more data necessary)**

	New Flow	Old Flow
Was the trusted adult helpful?	**2.08**	**2.77**

(lower number means more helpful 1-4 scale)

Next steps

- Tweak v1.1 and release v2.
  - Have fewer cancelations
  - Get more children to reach out to trusted adult (or friend)
  - Provide even more specialized help to youth who are in danger
- Analyze data more carefully; publish findings
  - Categorical analysis: Do reported posts map onto categories
  - Dive deeper into each category. Some numbers are alarming (e.g., physical threats). More categories likely are necessary
  - Learn more from kids about what they need to navigate their lives online
- Start helping older age groups
- Build a comprehensive help center for teens and parents
- Prevention is the key!

Let’s Imagine…

Thank you!

Today we are faced with the preeminent fact that, if civilization is to survive, we must cultivate the science of human relationships... the ability of all peoples, of all kinds, to live together, in the same world, at peace.

Franklin D. Roosevelt
1945

Relationships Matter!

Emotionally Intelligent Bullying Prevention

The 3rd Compassion Research Day

Arturo, Jake, Pete, Charles,
Emma, Josh, Diane, Dan, Andy,
Tijana, & Aileen

Yale Center for Emotional Intelligence
Marc Brackett, Robin Stern,
Zorana Ivcevic-Pringle, Andrés Richner,
& Diana Divecha

Our Team

Cyberbullying

- **Cyberbullying and "traditional" bullying are similar in many ways:**
  - An intentional act of aggression, based on an imbalance of power, that is meant to harm the victim (physically or psychologically).
  - Tends to occur repeatedly and over time, but sometimes can be identified in a single event.

- **But Cyberbullying also has unique characteristics:**
  - It's more easily replicated
  - It has limitless scalability
  - It's permanent

Prevalence of Cyberbullying

- **50% of middle and high school students say they have been cyberbullied and 33% report bullying someone online (Mishna et al. 2010).**

- **Adolescents report that cyberbullying spills into 'real life'**
  - 25% had experiences on SNS that lead to a face-to-face argument,
  - 22% had an experience that ended a friendship,
  - 13% got in trouble with parents, and
  - 6% got in trouble at school (Lenhart, 2012).

- **Why study cyberbullying?**
  - 20% of teens think that people their age are mostly unkind on SNS (Lenhart, 2012)
  - Adolescents say cyberbullying is more serious than face-to-face bullying (Mishna et al. 2009)
  - Cyberbullying is related to higher anxiety and depression, lower grades (Tokunaga, 2010) and higher rates of suicidal thinking and suicide attempts in adolescents (Hinduja, & Patchin, 2010)
  - 80% of US teens use social networking sites; 93% of them have Facebook accounts (Rideout, Foehr, & Roberts, 2010).

The big question:
Can social media can
incorporate design
that integrates
emotional intelligence and developmental science to
promote pro-social behavior
– both on- and off-line?

Two (seemingly) disparate fields

Emotional Intelligence

- EI introduced to psychology in 1990; reaches public in 1995
  - EI is the *ability* to reason with and about emotions to enhance decision making and promote both personal growth and pro-social behavior.
- Hundreds of studies demonstrating that EI is associated with positive outcomes for young adolescents
- Our EI program, RULER, has demonstrated positive results in shifting school climate and children's prosocial behavior

Technology/Social media

- Internet reaches the public in 1994
- Social media evolves out of the chat room and into popular networks
- Internet keeps getting blamed for social and psychological problems that are not new
- Facebook recognizes the potential power of integrating emotional intelligence principles into reporting systems

The life of a 13-14 year old

Young Adolescent Development

Biological Changes

- Onset of puberty leads to hormonal instability
- Executive network that allows self-regulation, planning, and overall monitoring, are “under development”
- Social excitement literally overwhelms the ability to control behavior.

Cognitive Changes

- Improvements in thought complexity makes kids more vulnerable to what others think. “Imaginary audience” (thinking that everyone sees them) makes them especially self-conscious and vulnerable to embarrassment.

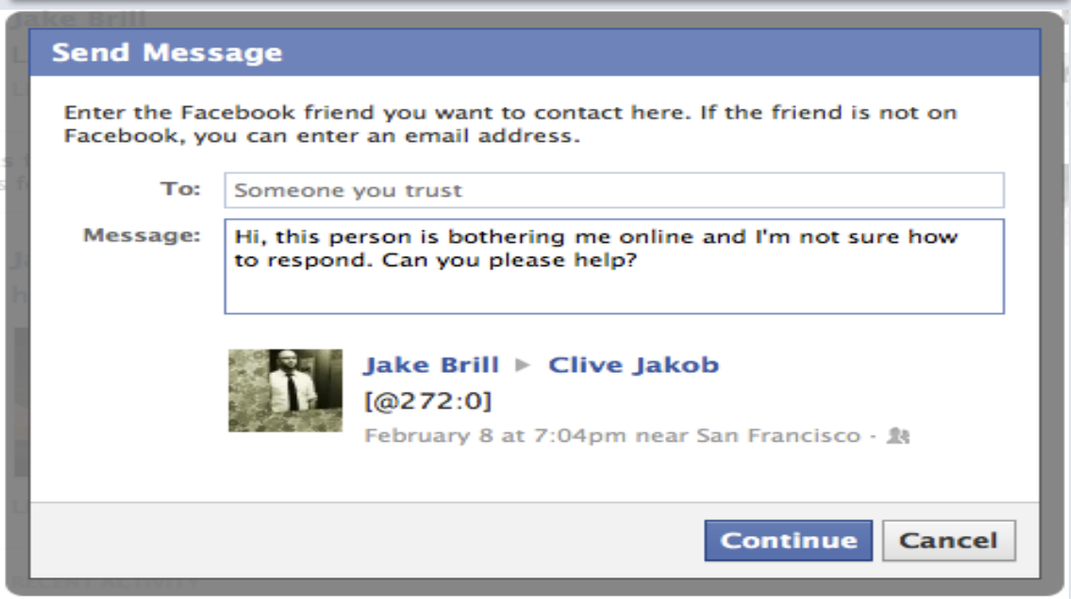
Self and Identity

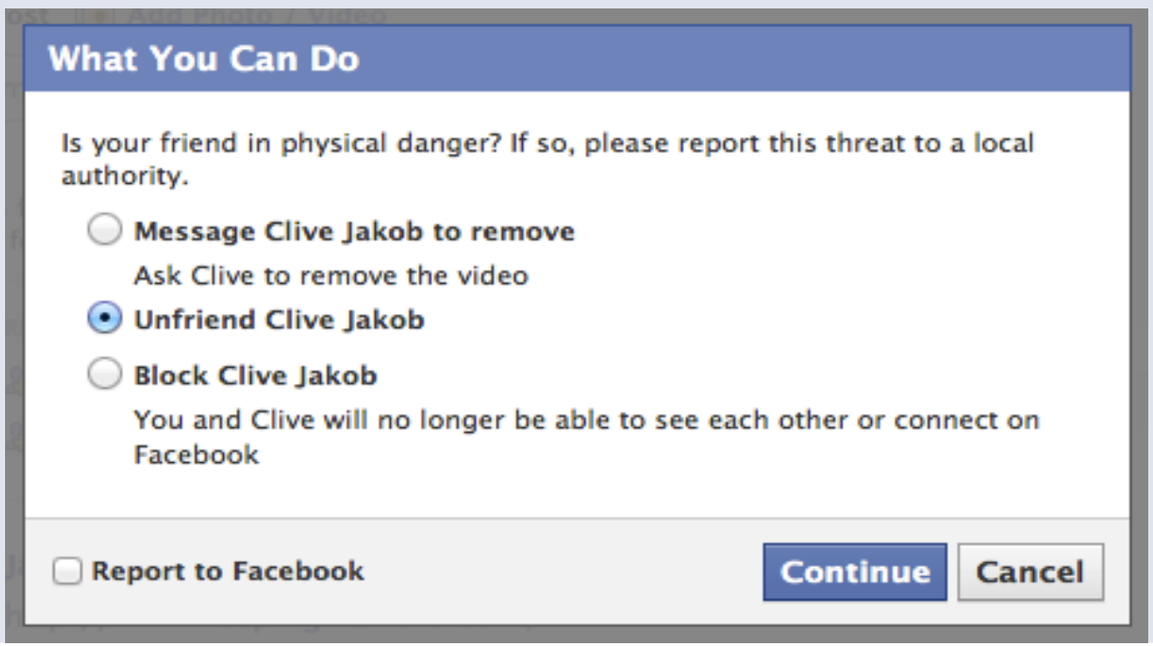
- Separation/individuation from parents; peer group offers temporary identity so they can become “autonomous”
- Young adolescents are especially sensitive to peer relationships – power dynamics and increased risk-taking especially in presence of peers.

Overview

- The original report flows (13-14 year olds)
- Infusing emotional intelligence
- What we learned from v1.1
- Version v2.0
- What the data reveal
- What’s next?

The original report flows





Infusing emotional intelligence

- **Takeaways from initial focus groups and interviews**
  - Kids were particular about the language we used
    - E.g., *report* – meant ‘authority’ or ‘trouble’ or ‘evaluated,’ whereas ‘get help’ suggested ‘technical problem’
  - Kids helped us to differentiate bullying and non-bullying experiences
  - Kids wanted Facebook to do something about it, but were not sure what that was; wanted a ‘conversation’
  - If questions were meaningful, specific, and helpful, they would be more motivated to complete the flow
  - Kids said they wanted help crafting messages
  - Kids did not believe everything needs to be reported b/c they would just tell (call, text) someone they trusted

Infusing emotional intelligence

- **Takeaways from interviews with parents**
  - Parents were mixed on whether they should be the ‘trusted’ adults
  - Some parents enabled kids to fake their age
  - If their child was threatened, they wanted to know
  - Parents wanted more resources for their kids
- **Our own takeaways**
  - Had to be a balance between what kids wanted and what we believed they need
    - E.g., Threatened – may not want to tell trusted adult, but they need help
  - A conversational approach was ideal
  - We needed to provide children, parents, and educators with more direct help

Infusing emotional intelligence

- **Infuse developmental emotion science – more adolescent-friendly language, enhanced logic, more relevant)**
  - 13/14 year olds prefer “This post is a problem” to “Report”
  - ‘What happened?’ to ‘how are you feeling?’ to ‘what can you do?
  - Move from just “harassing me” to “saying mean things to me”
- **Integrate emotional intelligence**
  - How did the post/photo make you feel? (both emotion and intensity)
- **Empower youth to take a positive, safe action both on- and off-line**
  - Provide simple, effective guidance for less versus more threatening posts
  - Develop positive pre-populated messages to content creator/trusted adults or friends

The Present Study

Version 2.0

DEMOGRAPHICS

What we learned from v1.1

- **Most reports were about ‘self’ as opposed to others**
- **Most kids just want to be ‘untagged’ from posts/photos**
- **Photo and post report systems needed to be separated**
- **We wanted to increase messaging to content creator and trusted friends/adults and decrease blocking/unfriending**
- **We needed to improve pre-populated messages to help teens communicate with content creators and trusted friends and adults,**
- **We also wanted to help trusted friends and adults communicate with the reporter**
- **We wanted to increase completion rates**
- **Gender was a variable that needed to be explored**

Descriptive Statistics

- **Reporter Information**
  - *N* = 402,269 13-14 year olds (distinct users)
    - Girls = 68%: Boys = 32%
    - All reports are between 9/1/12 to 12/31/12
  - Median # friends = 295; Girls = 332; Boys = 229
  - 1.5 reports, on average
  - Reporters were assigned randomly to old versus new flows
  - Based on approximately 4,000 follow up surveys:
    - Reports completed mostly by kids (85%), although some were completed by kids with their parents and parents alone (15%).

Descriptive Statistics

- **Content Creator Information**
  - Girls = 70%; Boys = 30%
  - Median # friends = 405; Girls = 438; Boys = 351

- **Reporter/Content Creator Mix**
  - Boy Reporters – Content Creators are: 55% (girl), 45% (boy)
  - Girl Reporters – Content Creators are 75% (girl), 25% (boy)

Version 2.0

PHOTOS

Photo Report Flow 2.0

Photo Report Flow 2.0

“I just don’t like it”

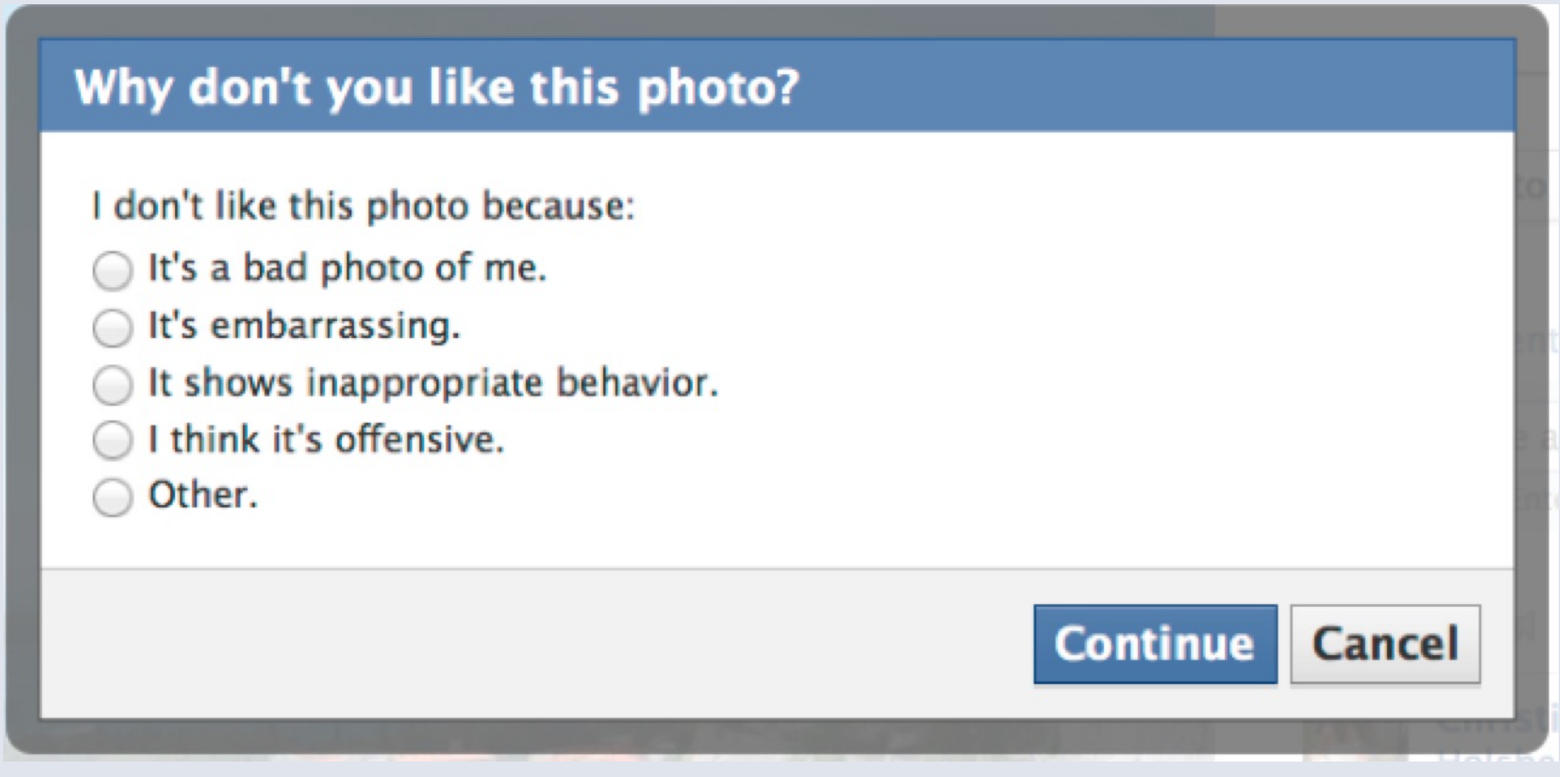


Photo Report Flow 2.0

“I just don’t like it”

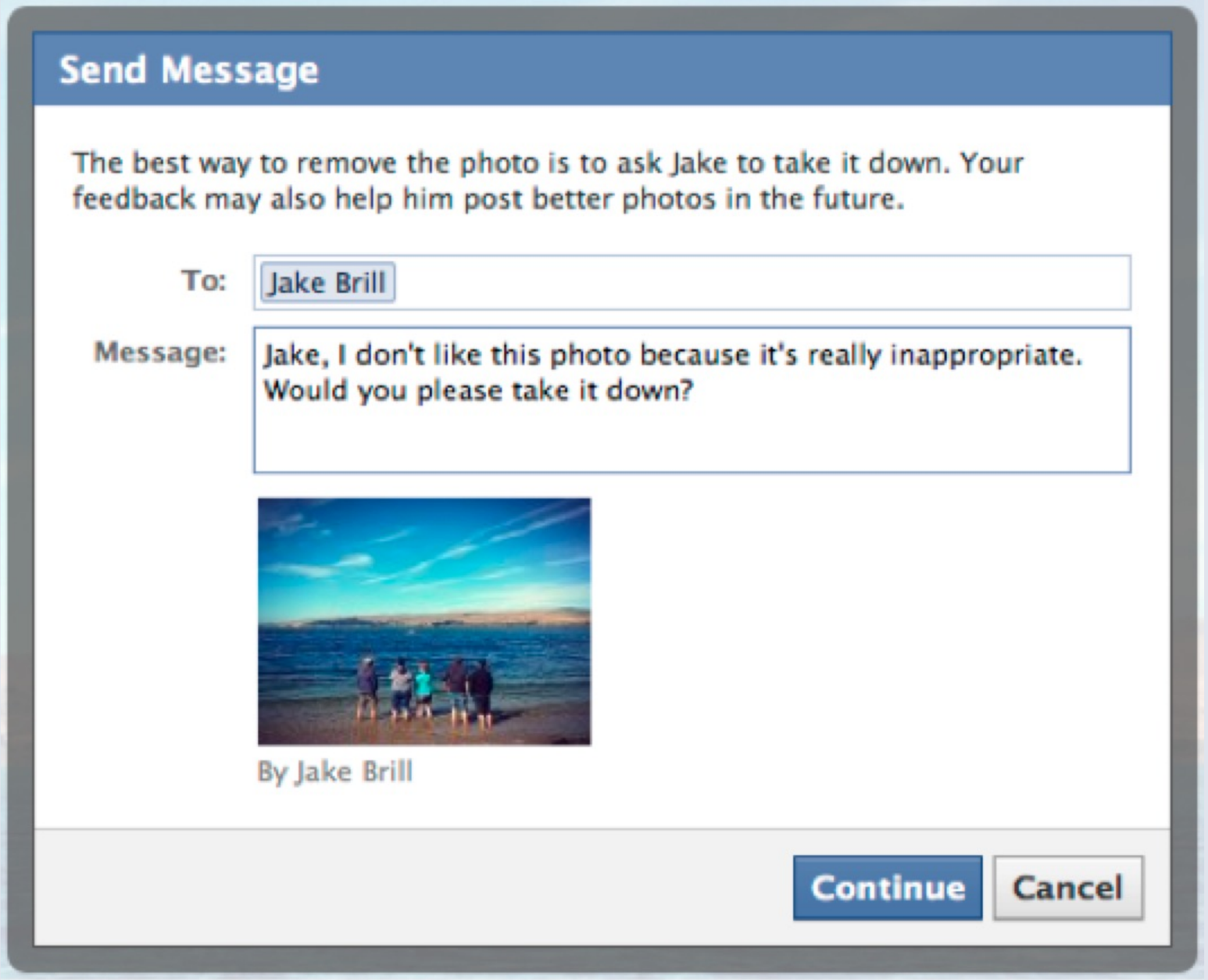


Photo Report Flow 2.0

“It’s harmful and might affect my reputation”

Photo Report Flow 2.0

“It’s harmful and might affect my reputation”

- *Messages are tailored to emotion intensity*
- *Can also send message via email*

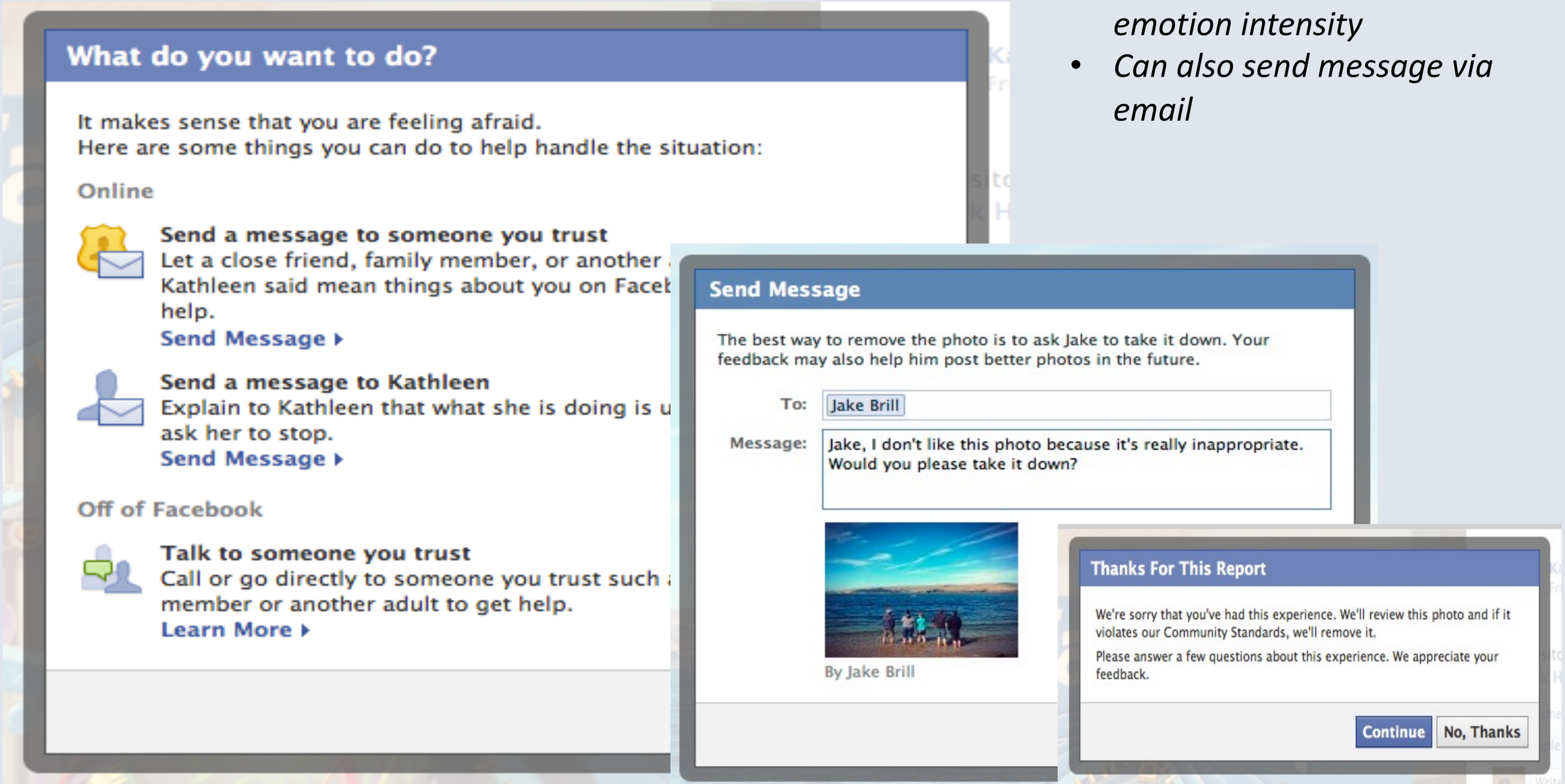


Photo Report Flow 2.0

“It’s harmful and might affect my reputation”
- spreading rumors -

<u>Low Intensity</u>

Jake, I don’t appreciate the rumors being spread about me. They make me uncomfortable. Please stop and take this post down.

<u>High Intensity</u>

Jake, I really don’t appreciate the rumors being spread about me. They make me very uncomfortable. Please stop and take this post down.

Photo Report Flow 2.0

"This PHOTO is a problem"
721,670
(Girl = 73%/Boy = 27%)

I just want to untag myself/spam
71%
77%/23%

I would like this PHOTO removed from Facebook
15.5%
62%/38%

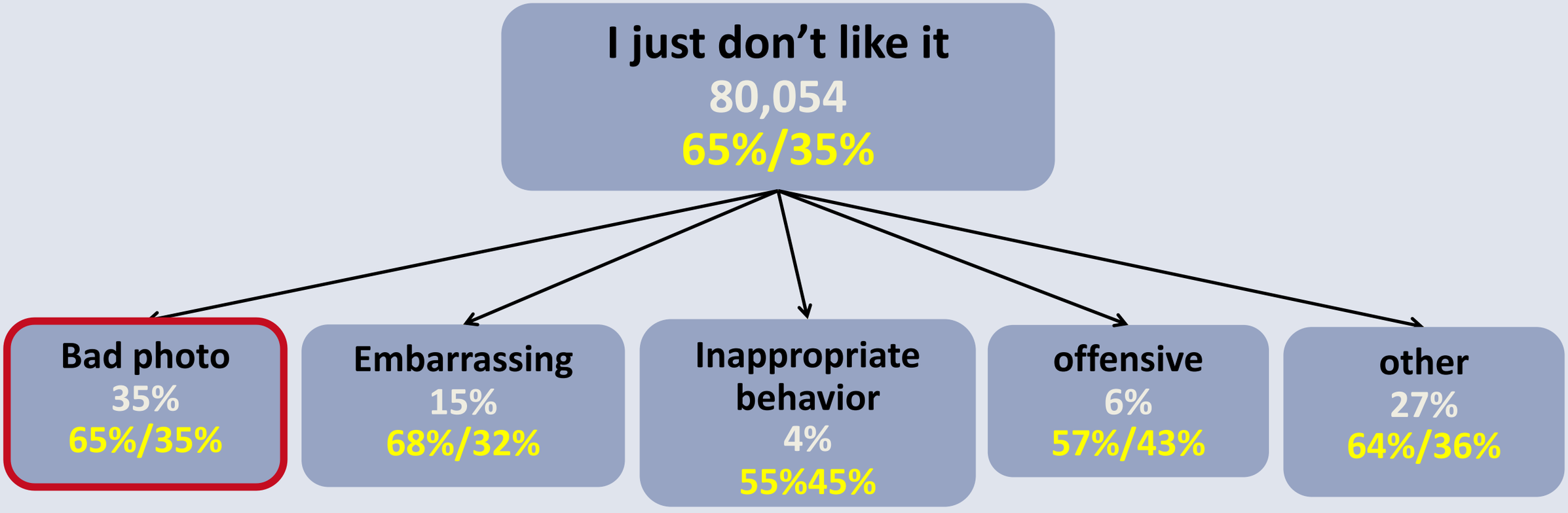
I want to help someone else
0.4%
65%/35%

I just don't like it
71%
65%/35%

It's harmful and might affect my reputation
16%
56%/44%

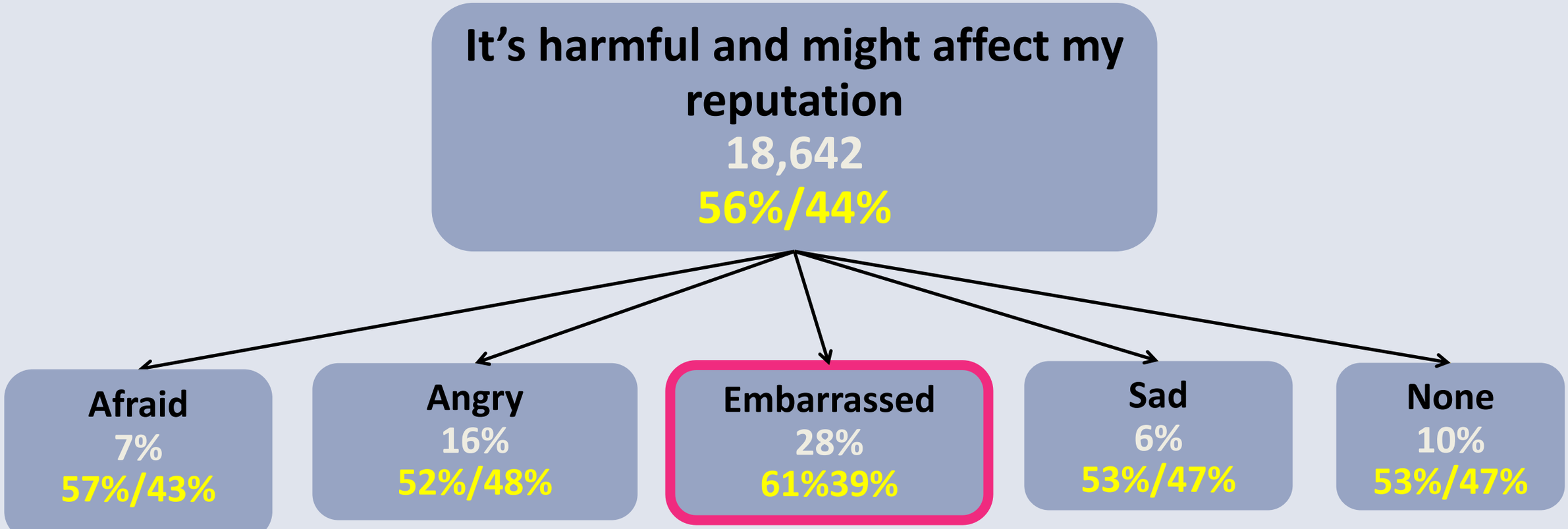
It shouldn't be on Facebook (TOS)
13%
60%/40%

Photo Report Flow 2.0



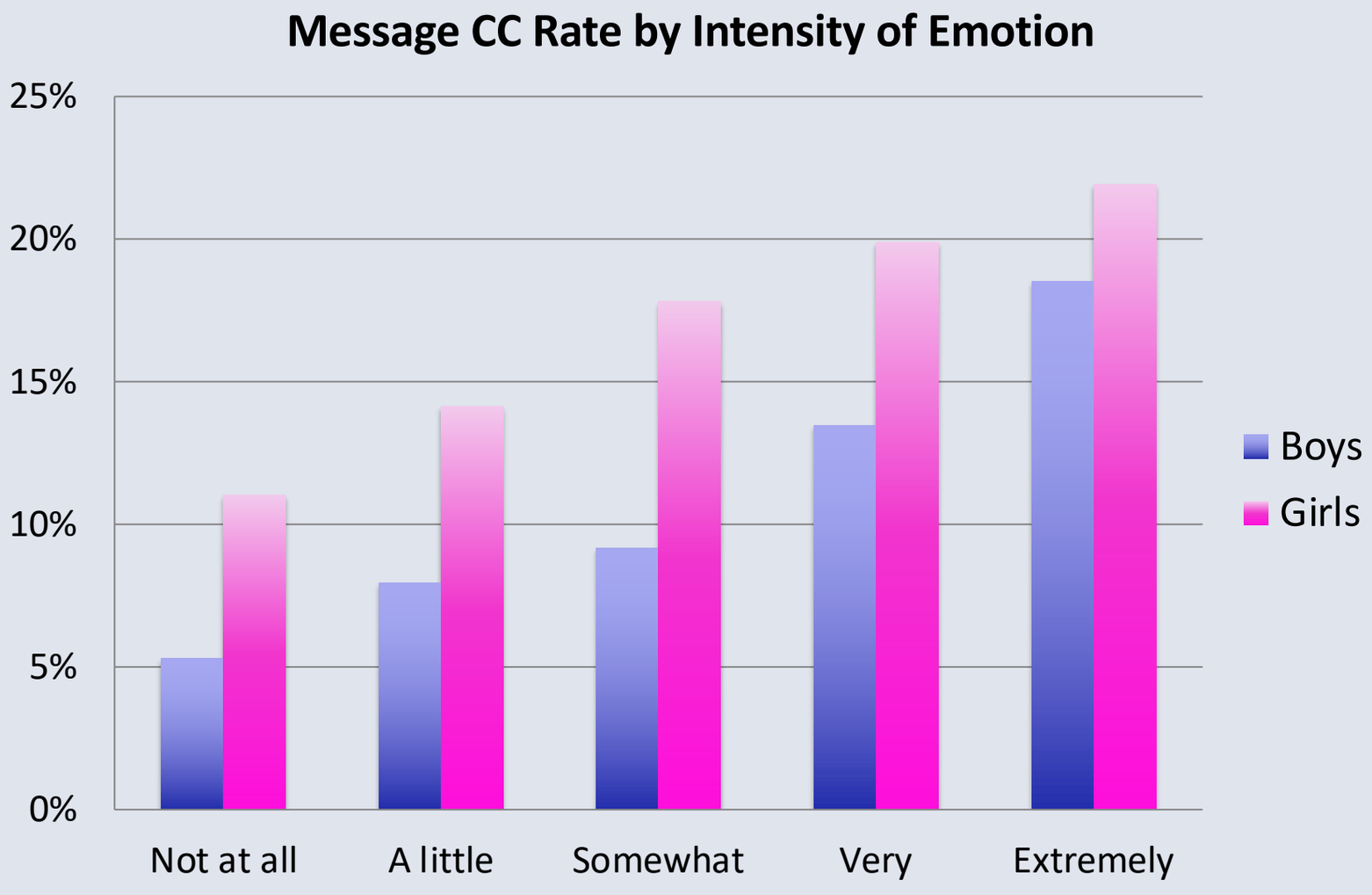
- On average, 58% of kids send messages to content creator
- Girls are more likely than boys to send messages for bad or embarrassing photos

Photo Report Flow 2.0



- **Embarrassment is the most frequently experienced emotion**
- **Embarrassment results in more messaging (18%) compared to all other emotions**
- **Girls are more likely than boys to send messages when embarrassed or afraid (7:3)**

Photo Report Flow 2.0



- **Emotion intensity is correlated with messaging, especially for girls**
  - ***Importantly, 84% of kids use our prepopulated (positive) messages***

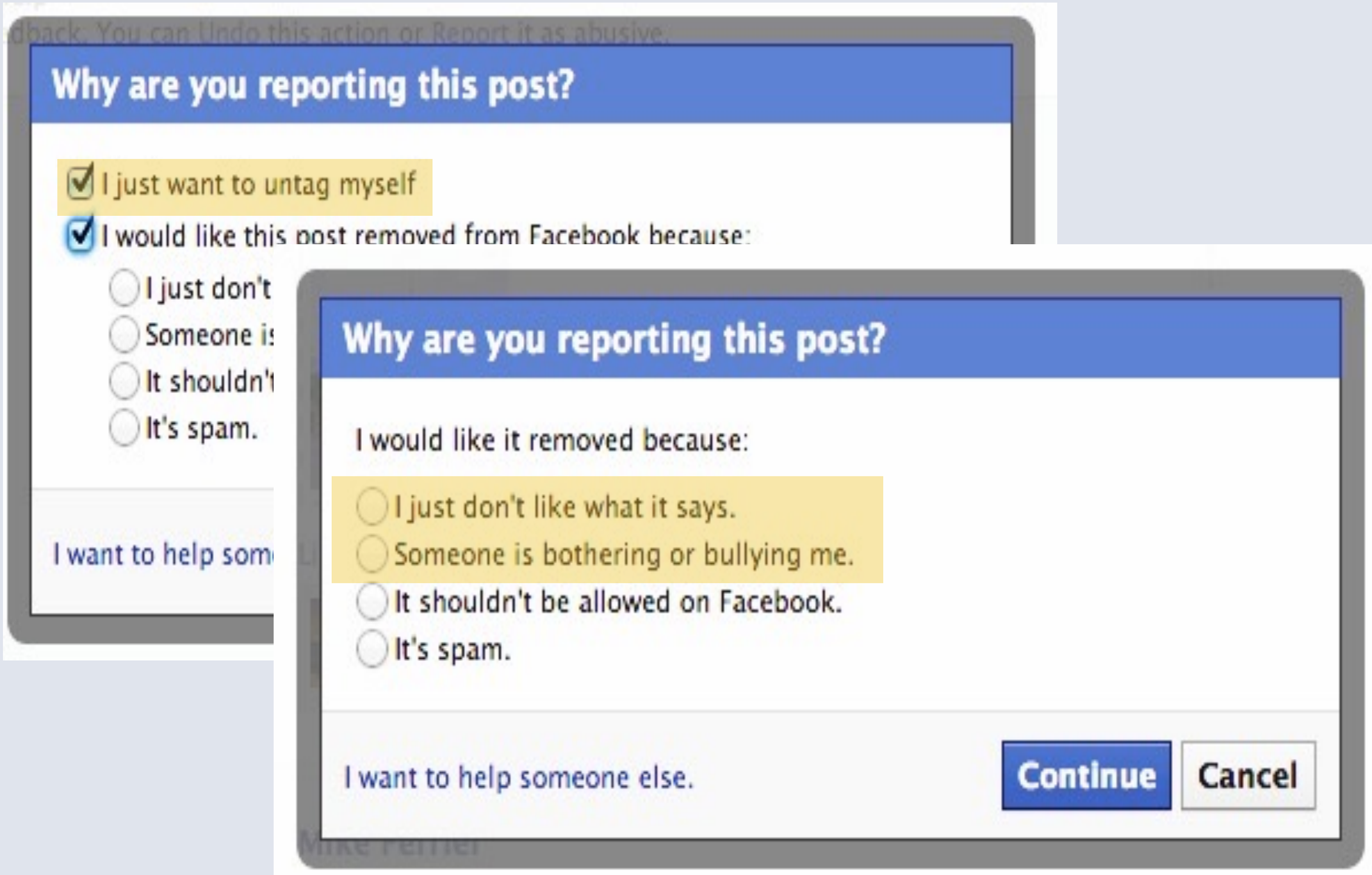
Reporting photos: Summary

- **Most adolescents simply wish to ‘untag’ themselves**
- **When wanting to remove photo, most adolescents ‘just don’t like it’ because ‘it’s a bad photo’**
- **Harmful photos are largely associated with embarrassment**
- **Stronger emotions result in greater likelihood of sending messages**
- **Gender differences are noteworthy**

Version 2.0

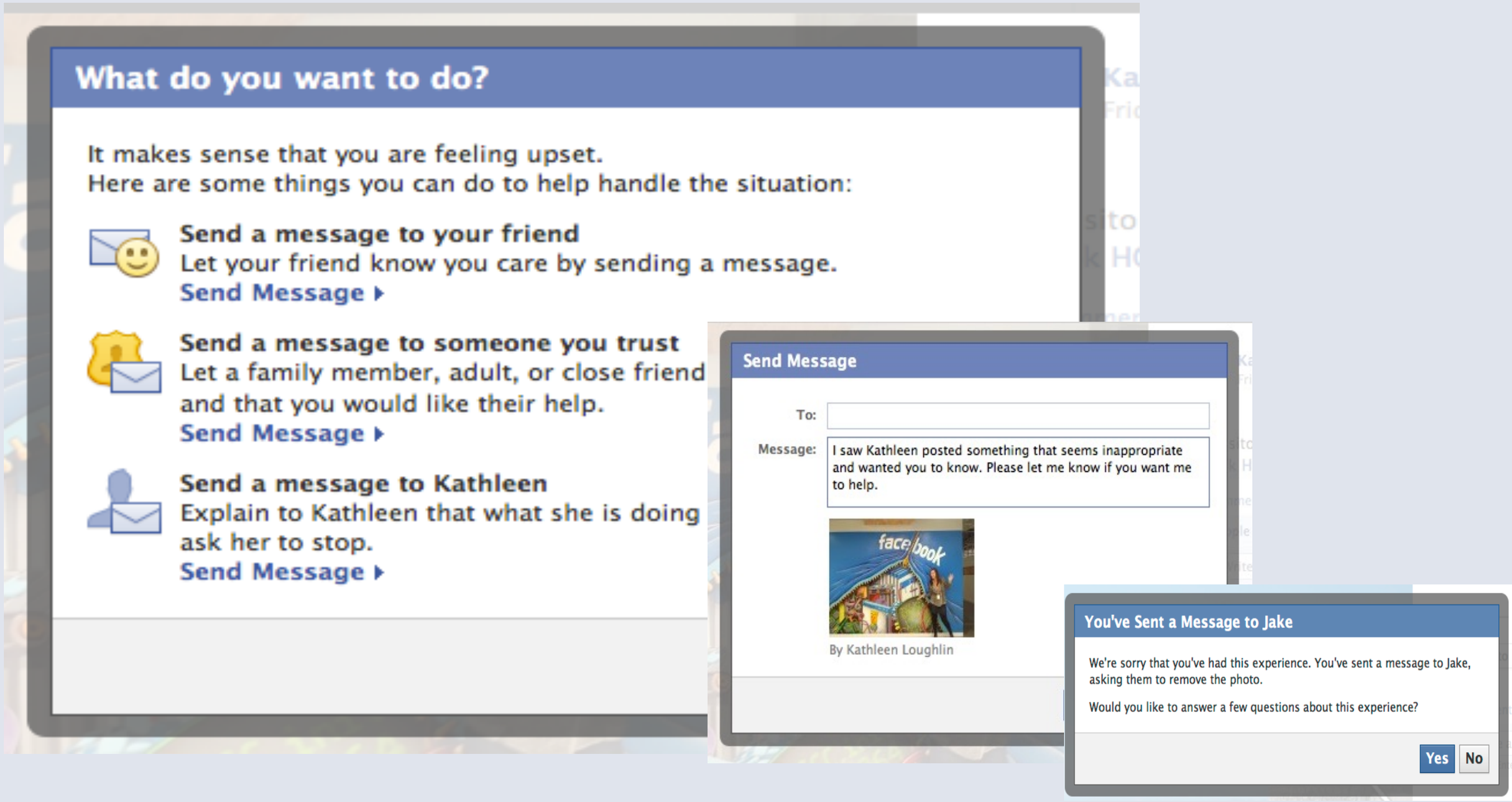
(POSTS)

Post Report Flow 2.0



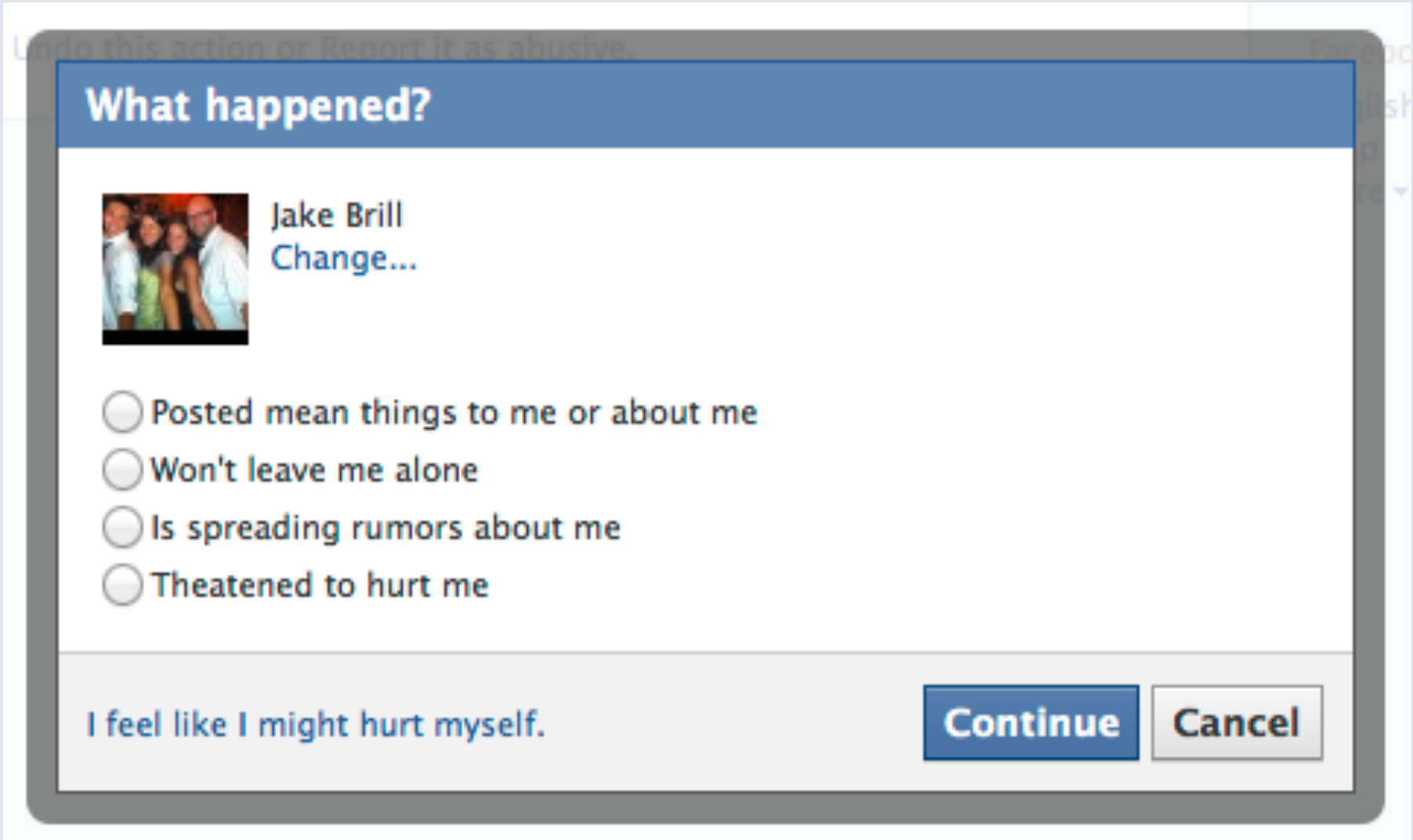
Post Report Flow 2.0

“I just don’t like what it says”



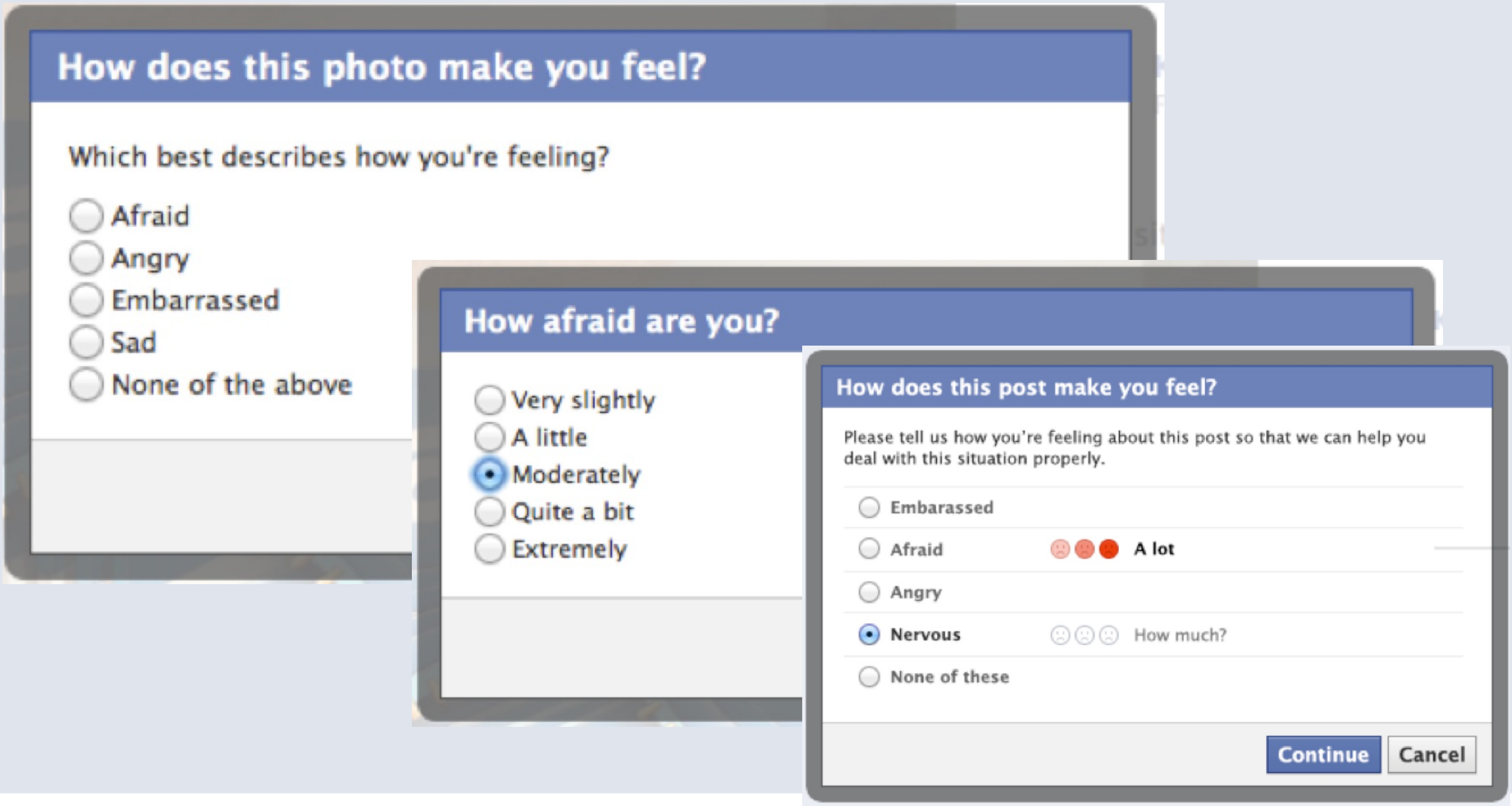
Post Report Flow 2.0

“Someone is bothering or bullying me”



Post Report Flow 2.0

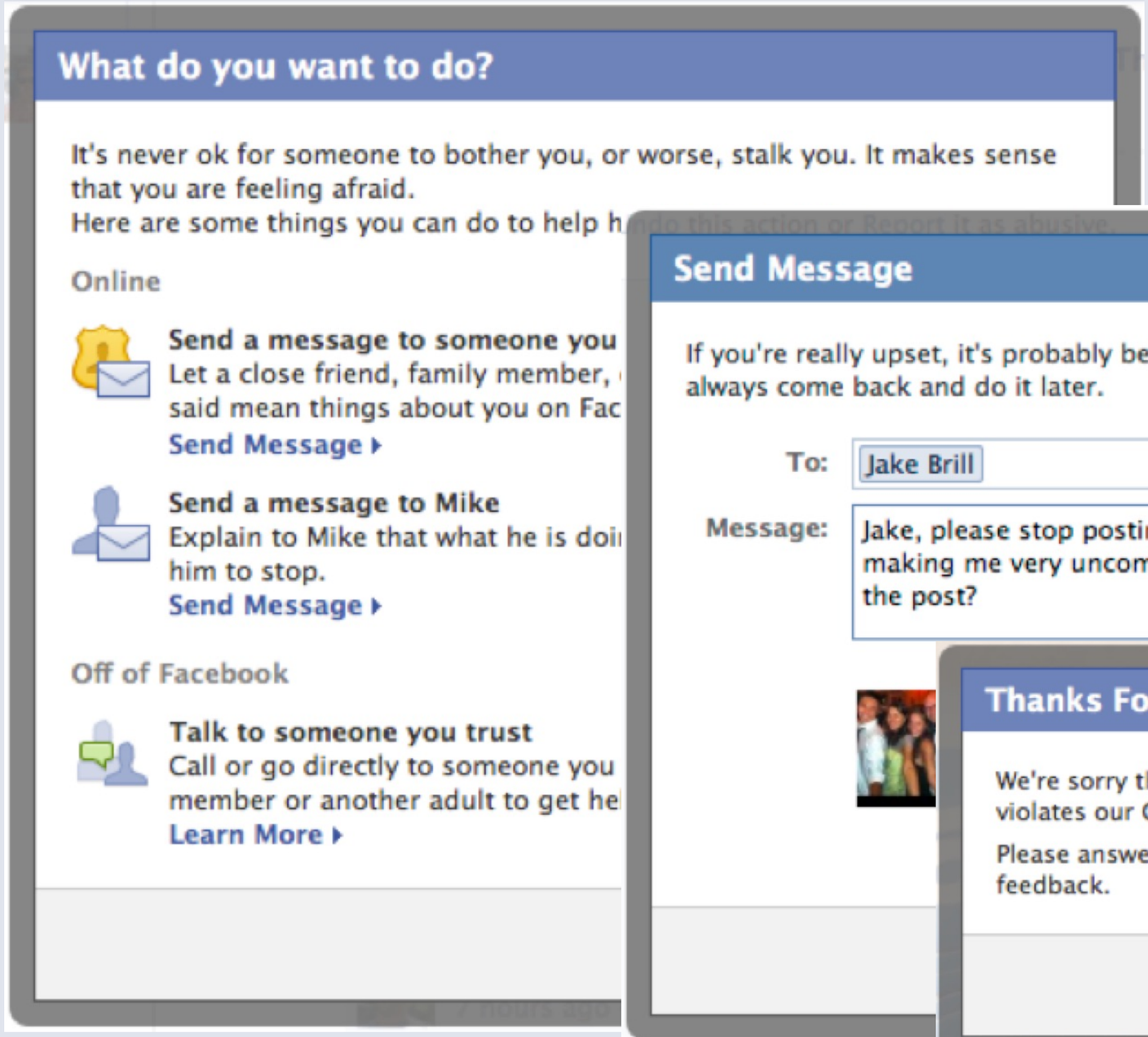
“Someone is bothering or bullying me”



Post Report Flow 2.0

“Someone is bothering or bullying me”

- *Messages are tailored to emotion intensity*
- *Can also send message via email*



Post Report Flow 2.0

"The POST is a problem"
61,305
(Girl = 61%/Boy = 39%)

I just want to untag myself/spam
39%
66%/34%

I would like this POST removed from Facebook
27%
61%/39%

I want to help someone else
1.4%
60%/30%

I just don't like what it says
59%
60%/40%

Message CC = 63%
57%/43%

Someone is bothering or bullying me
22%
60%/40%

It shouldn't be on Facebook (TOS)
19%
65%/35%

Post Report Flow 2.0

Someone is bothering or bullying me
6,499
60%/40%

- **Posted mean things** — 20% — 64%/36%
- **Won't leave me alone** — 10% — 60%/40%
- **Is spreading rumors** — 12% — 55%45%
- **Threatened to hurt me** — 6% — 60%/40%
  - Msg Other = 8%
  - Block = 4.4%
  - Unfriend = 12.5%

- **Anger is the most experienced emotion across all categories**
- **Non-significant gender difference on emotion "pick"**
- **Girls report having more intense emotions than boys**
- **On average, 10% of kids send messages to content creator and 3% to trusted friends or adults**

Post Report Flow 2.0

Area	Low intensity	High intensity
Said mean things	• Mocking reporter for over engagement with FB • Accusing reporter of being fake	• Negative post about unnamed individual • Targeted insults (e.g. fat, gay, slut)
Won't leave me alone	• Mocking reporter for over engagement with FB • Jokes about appearance	• Re-sharing reporter's content • Top 10 lists
Spreading rumors	• Negative post about unnamed individual • Relationship gossip	• Slurs • Top 10 lists • Sexually derogatory comments
Threatening *(emotion not asked)*	• Aggressive • Name calling • References to offline activity and situations	

Post Report Flow 2.0

- **Said mean things**
  - “[He] is gay as hell !!!! Dont be his friend !!!!”
- **Won’t leave me alone**
  - “Get some proactive , and a better attitude , THEN we'll talk . (;”
- **Spreading rumors**
  - “[She] is such a whore.. she's told me she slept with 5 different guys and she's willing to do more. What a whore.”
- **Threatening**
  - “Watch your back you little bitch (; your going to wish you never fucked with me.”

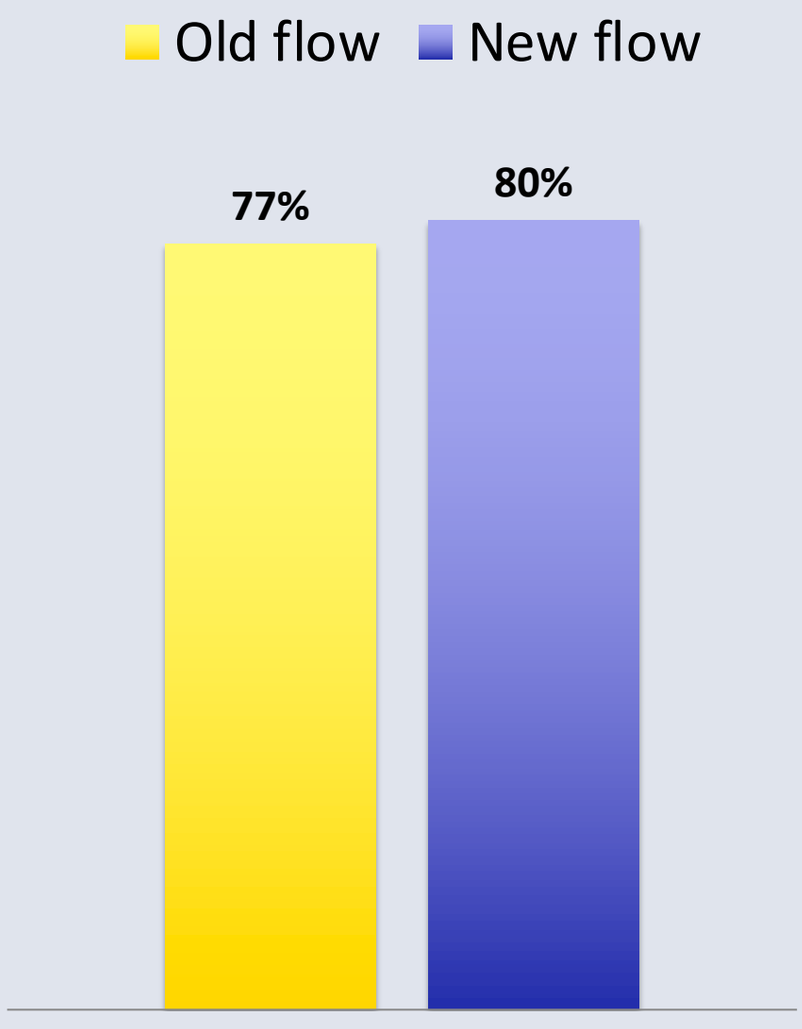
Reporting Posts: Summary

- **Similar to photo reports, most young adolescents simply wish to ‘untag’ themselves from posts**
- **When wanting to remove the post, most young adolescents ‘just don’t like what it says’**
- **When being bothered or bullied, most report ‘mean things’ being posted, resulting in anger**
- **We need to unpack more what’s happening for kids who report that someone is threatening to hurt them**
- **Again, there are noteworthy gender differences**

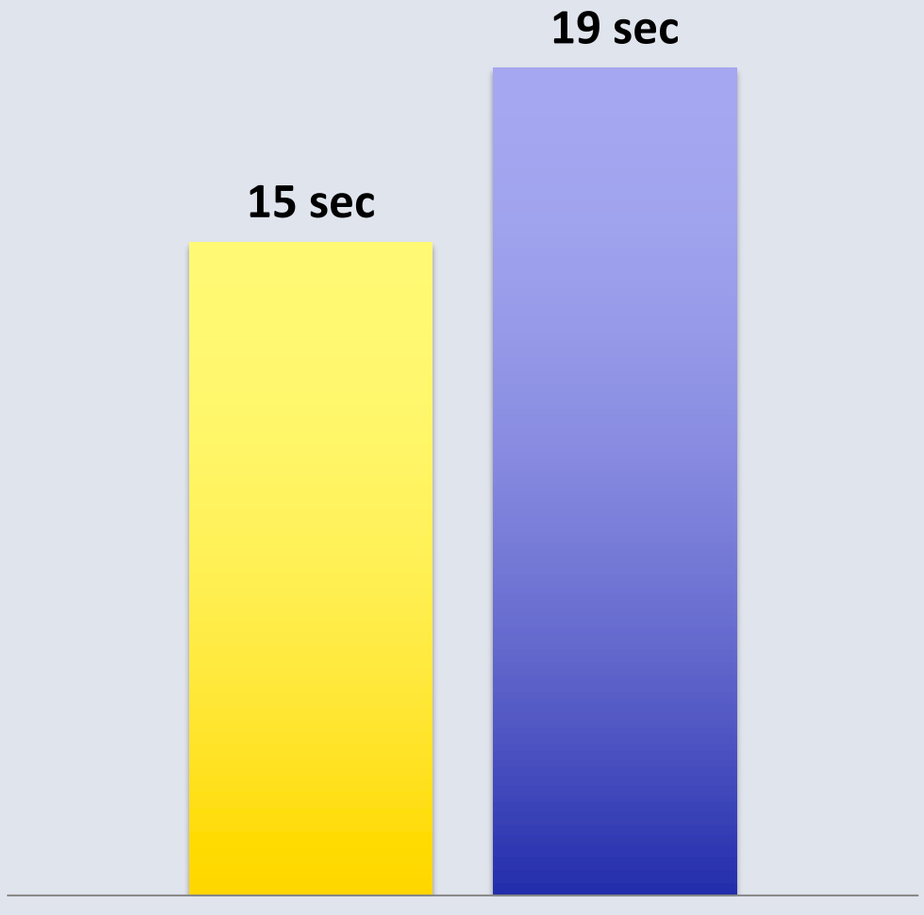
Experimental Findings:

Original vs. v2.0

Old Flow vs. 2.0 Flows

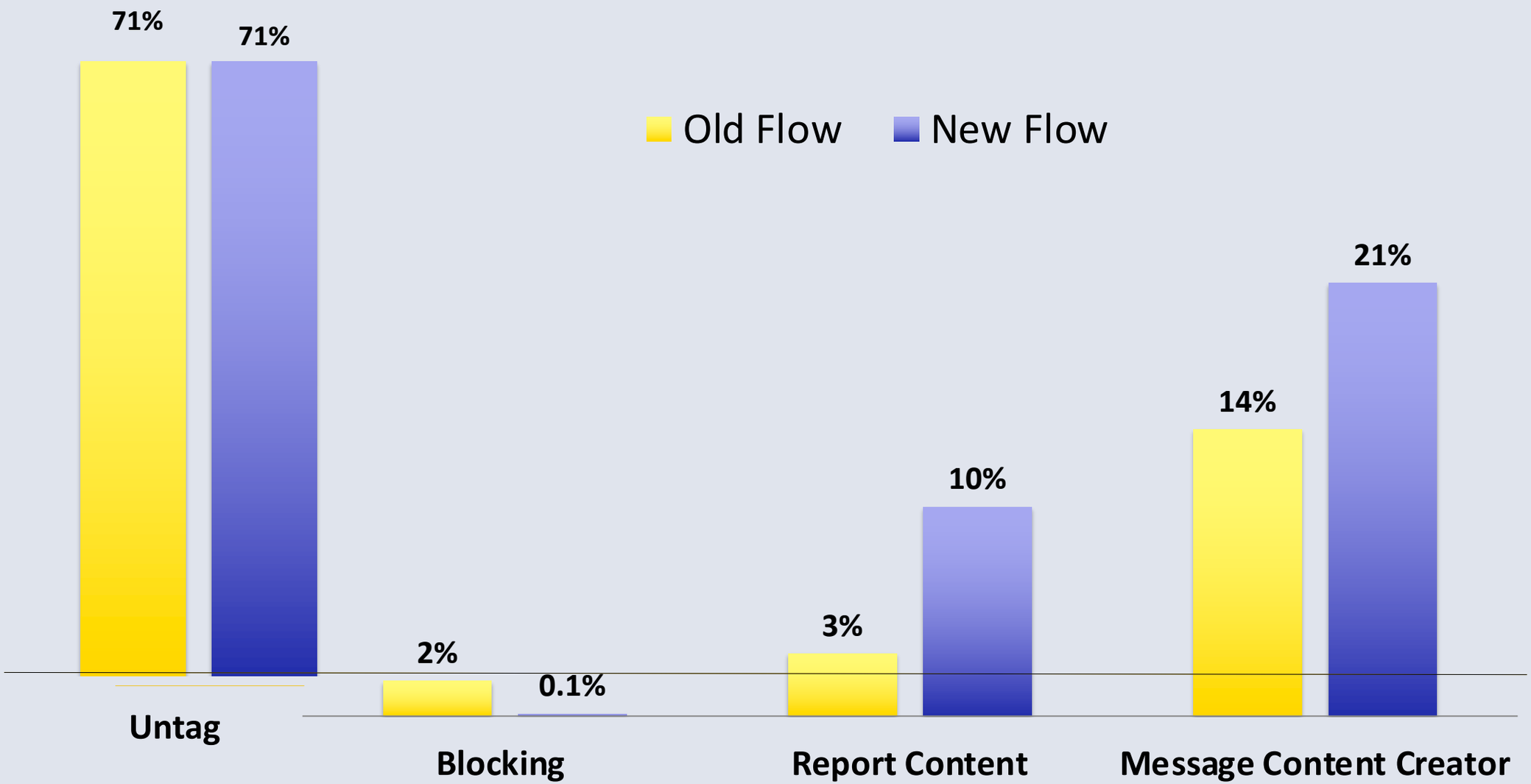


Completion rate

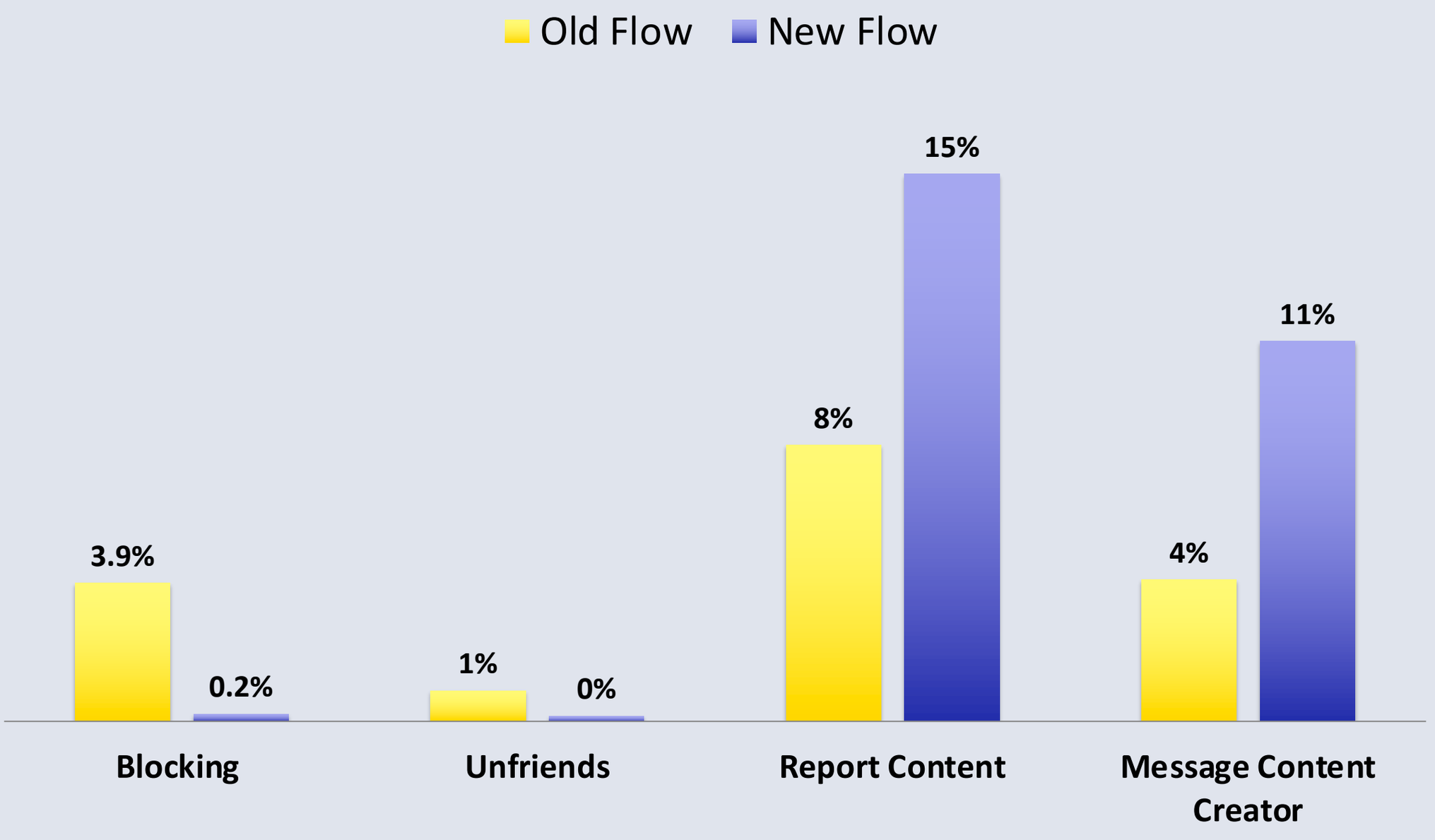


Time spent in flow (not just untag)

Old Flow vs. Photo 2.0 Flow



Old Flow vs. Post 2.0 Flow



Discussion

- **Gender matters**
  - Reporting behavior – girls report more than boys
  - Bullying behavior – girls are more likely than boys to be the ‘content creators’
- **Embarrassment is most frequent emotion associated with photos**
  - Kids are self-conscious about the way they look
- **Anger is most frequent emotion associated with posts**
  - Kids “say mean things” which is perceived of as an injustice
- **Emotion intensity is associated with behavior (messaging)**
  - Emotions drive decision making and action
- **Providing kids with a more emotionally intelligent report flow helps to have more positive interactions**
  - Kids are more likely to “stay in the relationship” and make constructive decisions like sending positive messages as opposed to blocking
  - In essence, we have eliminated ‘blocking’ – likely an ineffective coping strategy

Limitations and Next steps

- **These data only represent kids who reported; many kids do not know about the reporting system so there is a need to get the word out.**
- **We need to try new methods for follow-up survey data – satisfaction, resolution? Follow up on content creator, trusted friends/adults**
- **Confirm findings in a fresh sample with some tweaking to the flow**
- **Qualitative Analyses**
  - Gender differences
  - Mapping posts onto categories
  - Examine posts preceding and following report
- **Help Center for kids, parents, and educators**
- **15-16 Year old flows coming soon**

Thank you!

Emotionally Intelligent Bullying Prevention

The 4th Compassion Research Day
December 5, 2013

facebook

Yale *Center for Emotional Intelligence*

Yale Team

Marc Brackett
Mrinalini Rao
Charlie Sherman
Robin Stern
Diana Divecha
Zorana Ivcevic
Cynthia Dickason-Scott

Facebook Team

Bhal Agashe
Rob Boyle
Charles Gorintin
Cheryl Lowry
Diane Murphy
Dave Steer
Arturo Bejar
Tessa Cafiero
Samantha Gruskin
Mojtaba Mehrara
Mamal Poladia
Siqui Yan
Emiliana Simon-Thomas
Pete Fleming
Jennifer Guagagno
Dan Muriello
Nikki Staubli

Creating Evidence-Based Tools for Teens

- Part I – Social resolution flows for teens (ages 13-16)
  - Provide kids with tools to help them manage unpleasant experiences

- Part II – Bullying Prevention Hub
  - Provide kids, parents, and educators with high-quality resources to manage and prevent bullying

Adolescence and Social Media

"If you didn't have Facebook when you were a kid, how did you know who your friends were?"

Adolescence and Peer Relations

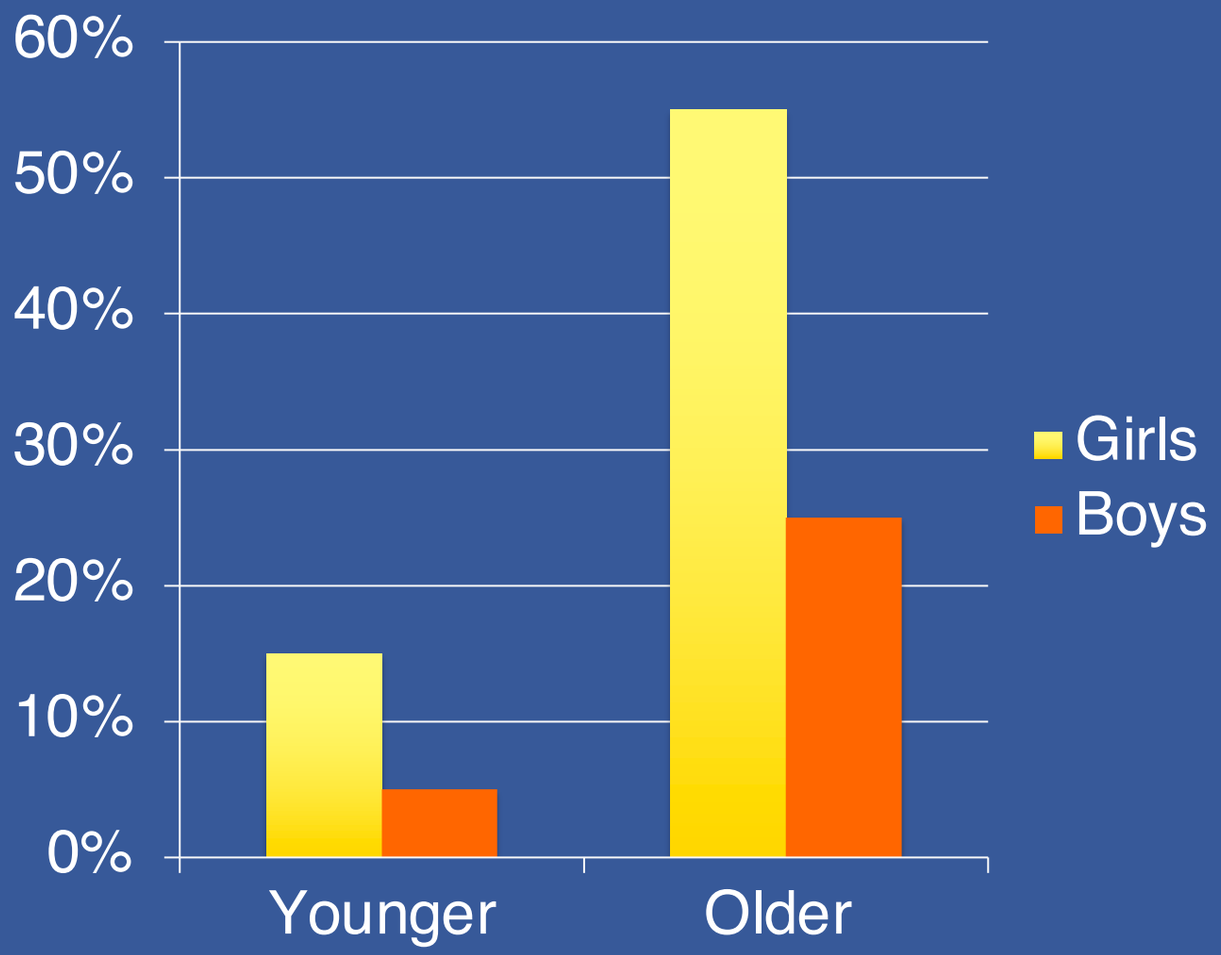
- Peer relationships are a central focus for teens
- Creating and maintaining positive relationships doesn't happen automatically
- The adolescent brain is different
- Emotion skills matter

Emotionally Intelligent Bullying Prevention

- Infused a developmental framework
- Incorporated age appropriate/conversational language
- Integrated emotional intelligence skills
- Empowered youth to take positive action

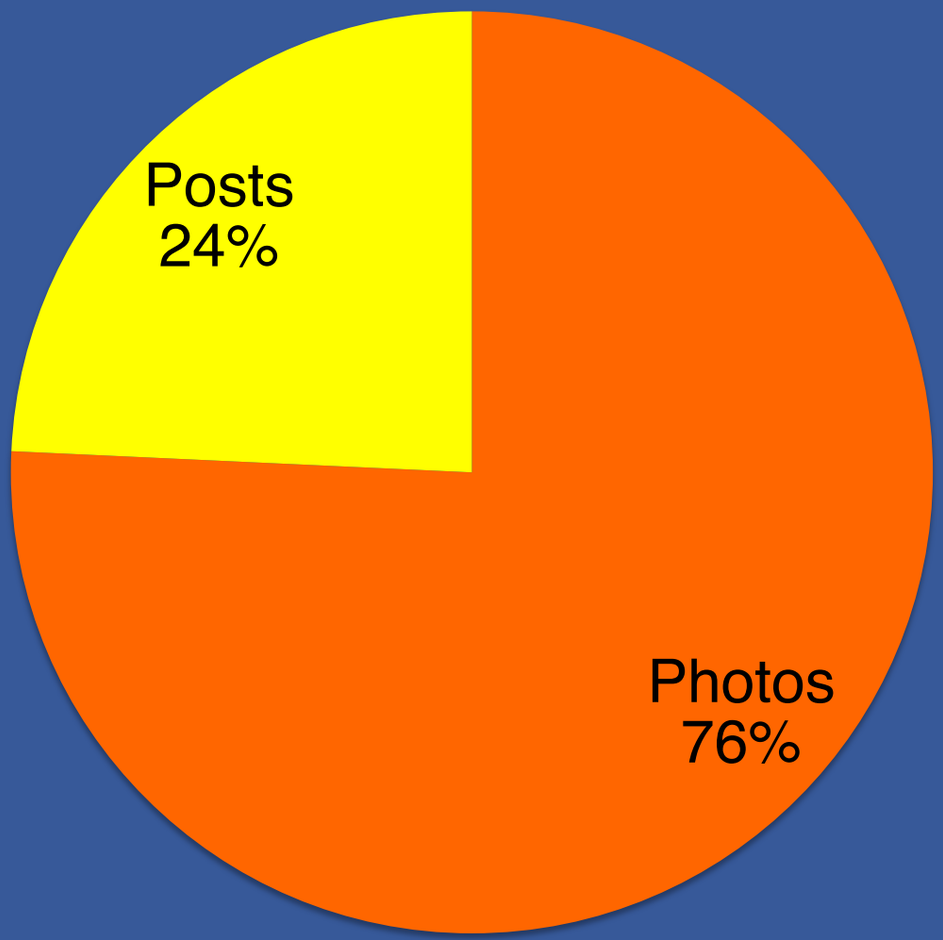
The Resolution Tools…

The Present Sample



- *N* = all 13 -16 year olds in U.S. who entered resolution tool within a 30-day period
- Older girls use the tool the most and also are reported more

What are the resolution tools being used for?



Posts:

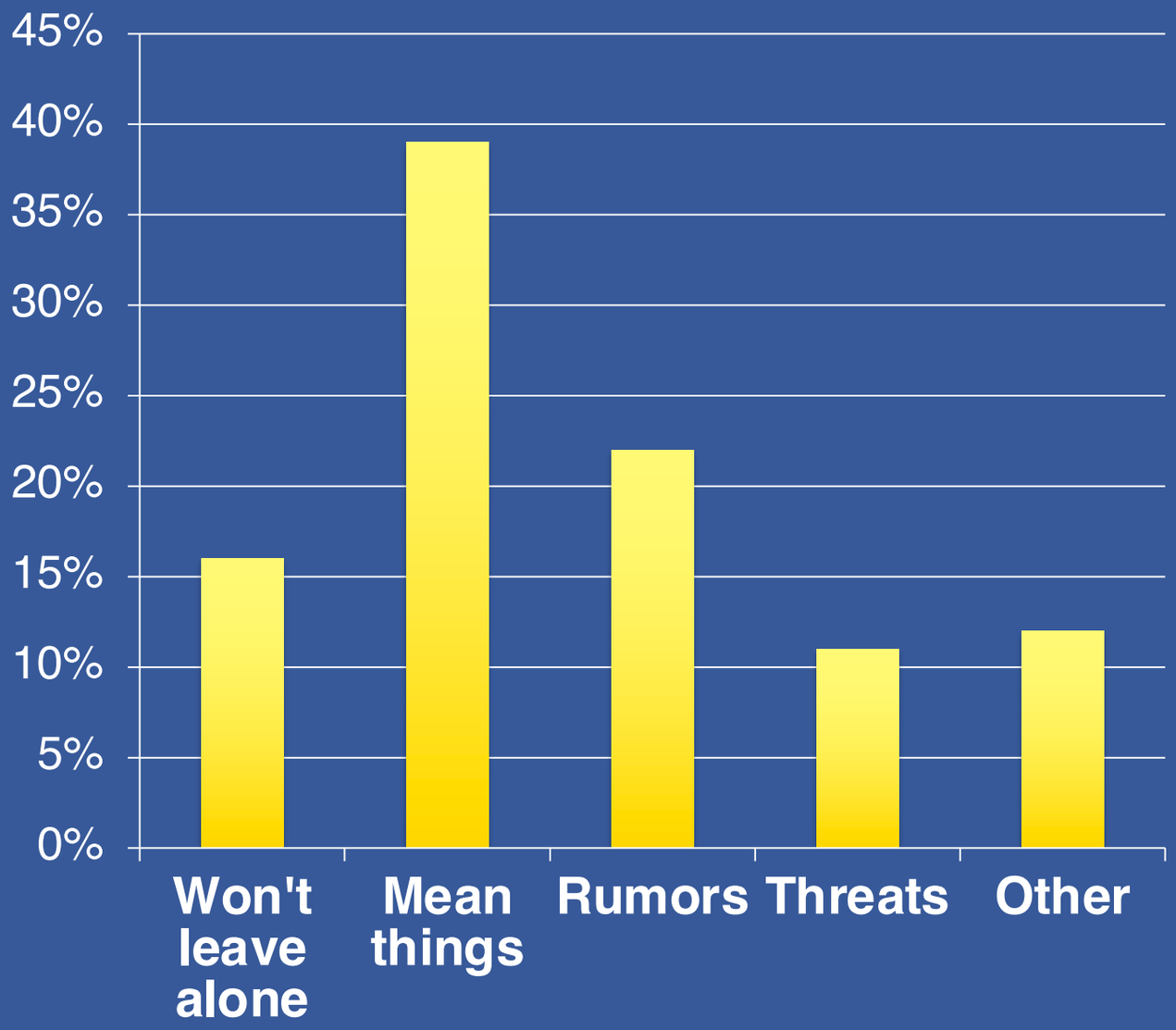
"Someone is bothering or bullying me"

Photos:

"It's harmful or might hurt my reputation"

*Of all teens entering the flows, 15% select 'bullying.' Most (66%) select 'annoying.'

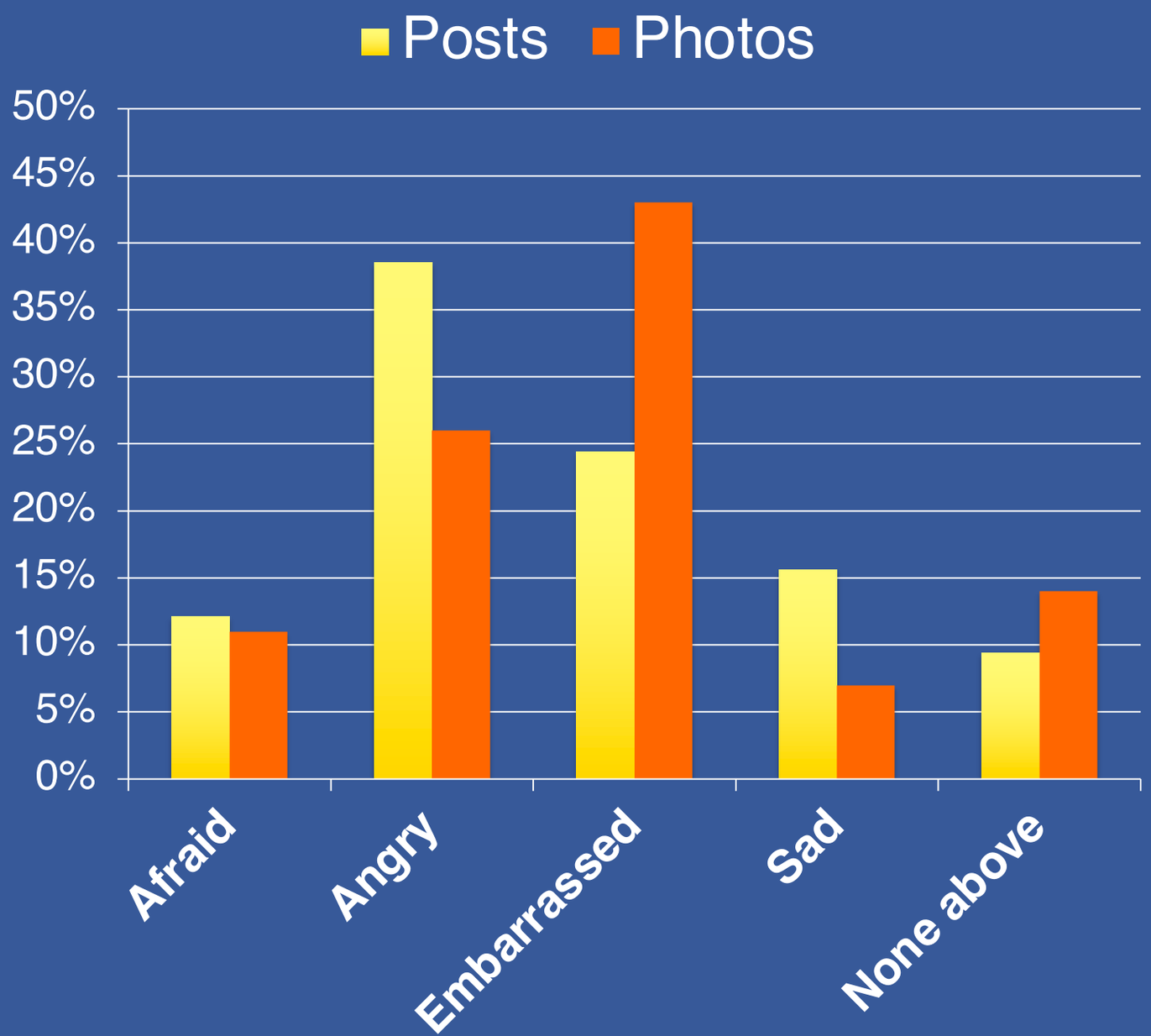
“What happened?”



- No gender differences
- No age differences
- Younger boys report more threats. Older boys report more mean things.

*Breakdown of 15% who select ‘bullying.’

“How does this post/photo make you feel?”



- Girls report more sadness and embarrassment than boys
- Boys use “none” more than girls
- No age differences
- Younger boys report being more afraid and more threats than older boys

Sometimes it’s clear why teens label posts as bullying

He was crying today lol

- Reported by 14 year old girl

You better watch yo back I'm going to knock you out tomorrow.

- Reported by 14 year old boy

I feel like he just used me! But I also thought he loved me. I should have known better. Maybe one day we'll get back together!

- Reported by 17 year old boy

And sometimes it’s not

So ready to go home and go to bed!

- Reported by 13 year old boy

I got contacts. No more glasses.

- Reported by 16 year old boy

What actions do teens take?

- 25% of teens message - person who posted the content (90%) or a trusted adult/friend (10%)
- 75% of teens use the pre-populated (positive) messages
- Younger teens message more. However, younger boys who report ‘afraid’ send more reports to Facebook.
- “Won’t leave me alone” ➔ use pre-populated messages; “rumors” ➔ tailor messages

What happens next?

- Content creator behavior:
  - 75% reply to the message
  - 37% delete content

- Parental involvement:
  - 38% of younger teens vs. 23% of older teens' have parent involvement

Summary of findings

- Like face-to-face bullying, online bullying results in a range of emotional experiences for both boys and girls (embarrassment/anger are dominant)

- Teens' online lives look similar to their offline lives:
  - More girls than boys report being sad and embarrassed
  - More boys report 'threats' than girls
  - Boys are less willing to disclose feelings

- Age differences in the 'content' of online bullying are consistent with face-to-face bullying (e.g., homophobic bullying)

- When given effective "tools" teens appear to send messages – and when they learn have done something 'wrong,' they tend to respond

- These results helped to inform us about other tools kids and adults needed

Part 2: The Bullying Prevention Hub:

A day in the life of father and son

Insights

Methodology

- Focus groups with teens, educators, parents
- In depth summit with nonprofits
- Data from social resolution flows

Learnings

- Awareness of bullying, but low comprehension of what to do
- All stakeholders want guidance
- It's about bullying intervention and prevention wherever it occurs – focus on the behavior, not the location or platform

Our goal was to develop *emotionally intelligent* bullying prevention resources

Resources for all stakeholders:

- Parents, educators, and teens
- Bullies, targets, and bystanders

Knowledge and skills content:

- Resources which build self-awareness, self-regulation, problem solving, and healthy communication.

Safety is a Conversation

- Provide the right advice to the right user at the right time.
- Expand our bullying prevention campaign and the Family Safety Center.
- Help on the other side of the reporting button.
- Showcase resources from dozens of organizations

Stop Bullying

Tools, tips and programs that help people stand up for each other.

Introducing the Bullying Prevention Hub

Resources for teens, parents and educators

We're sorry you're having this experience

No one should spread rumors about you. Learn More about how to handle situations like this, or try to resolve this with one of the options below.

Get Help
Ask for help or discuss this with someone you trust.

Message Isabella to resolve this
Ask Isabella to take it down.

Unfriend Isabella Anderson
Remove Isabella as a friend.

This has been submitted to Facebook for review
Based on what you've told us, this seems serious and Facebook will review this post. You can check the review status on your Support Dashboard.

Back

You're Temporarily Blocked

You're temporarily blocked from posting on Facebook for the next 12 hours. Please review our Community Standards so you can understand what's allowed on Facebook and keep your account in good standing:

Bullying and Harassment

Facebook does not tolerate bullying or harassment. We allow users to speak freely on matters and people of public interest, but take action on all reports of abusive behavior directed at private individuals. Repeatedly targeting other users with unwanted friend requests or messages is a form of harassment.

Learn more about how to recognize and prevent bullying on Facebook.

OK

Let's go back to the role play

Charlie was accused (and is guilty) of posting something inappropriate – a photo that was mean and hurtful – about his classmate. It was a picture of his classmate Jamie at a sleepover party. The photo showed her drinking a beer.

Marc, his father, got the call about this from the school principal.

STEP 1

Set yourself up for a successful conversation with your child.

- *Find the best space to have the conversation.*
- *Check in with and manage your own feelings (before)*
- *Remember, you are the role model.*
- *Support and listen.*

STEP 2

Talk with your child about the problem.

- *Find out what happened.*
- *Communicate your family's values (e.g., respect, kindness).*
- *Use a calm and steady voice; avoid making empty promises.*

Sample Conversation Starter:

"I got a call from your teacher today who told me that you have been posted a offensive photo of Jamie. I need to know what happened so we can decide what action needs to be taken."

STEP 3

Work with your child on an action plan.

- *Solve the problem together.*
- *Ask fact-finding/open-ended questions to help your child generate solutions*
- *Decide on an appropriate action plan (e.g., apologize)*

Sample Conversation Starter:

"What do believe are some appropriate ways to handle this situation?"

STEPS 4 & 5

Be clear about consequences, follow through, and follow-up

- *Be firm and consistent, taking into consideration your values and severity of incident.*

More opportunities to help your child…

- *Pay closer attention to your child's Internet and cell phone activity.*
- *Advocate for an evidence-based social and emotional learning program for your child's school.*
- *Consider counseling for your child and/or family.*

Back to Charlie and Marc

- **Set yourself up for a successful conversation with your child.**
- **Talk with your child about the problem.**
- **Work with your child on an action plan.**
- **Be clear about consequences, follow through, and follow-up**
- **Explore more opportunities to help your child**

The future of emotionally intelligent bullying prevention

- **Social Resolution Tools**
  - Examine role of gender and age in more detail
  - Conduct qualitative analysis of posts and photos
  - Run longitudinal studies on teens online behavior, including follow-up survyes
  - Begin cultural adaptations
  - Share findings in peer-reviewed journals

- **Bullying Prevention Hub**
  - Study the use and impact of hub
  - Create more interactive tools for all stakeholders (e.g., videos)
  - Build bully education center

Emotions Without Borders

Supporting Teens Across the World on Facebook

Marc Brackett, Mrinalini Rao,
Robin Stern, & Zorana Ivcevic
Yale Center for Emotional Intelligence

Facebook's Protect and Care Team

Yale *Center for Emotional Intelligence*

Vision

To use the power of emotional intelligence to create a more healthy, effective, and compassionate society.

Mission

To conduct rigorous research and develop innovative educational approaches to empower people of all ages with the emotional intelligence skills they need to succeed.

Emotions Matter

A rollercoaster of emotions

Emotions Matter

Emotions drive:

- Attention, memory, and learning
- Decision making and judgment
- Relationship quality
- Physical and mental health
- Everyday effectiveness

Emotions Matter for Teenagers

To many, adolescents appear to be illogical, irrational, and invincible, but…

- Puberty introduces hormonal changes
- Emotion and cognitive systems are not harmonized
- Separation and individuation from parents is
- Peers have a strong influence

Emotions Matter for Teenagers

- Seeking easier means to gain rewards
- Increased risk taking (e.g., driving, risky sexual behavior)
- Delinquent behaviors
- Substance abuse
- Psychiatric diagnoses
- Suicidality

Emotional Intelligence

- **R**ecognizing emotions
- **U**nderstanding emotions
- **L**abeling emotions
- **E**xpressing emotions
- **R**egulating emotions

How Emotional Intelligence Develops

What was your emotion education like?

Developing Emotional Intelligence

"Between stimulus and response, there is a space. In that space lies our freedom and power to choose our response. In our response lies our growth and freedom."

VIKTOR E. FRANKL

Developing Emotional Intelligence

Moving from automatic to intentional ways of behaving

- Yelling to deep breathing
- Negative self-talk to positive self-talk
- Impulsivity to reframing
- Rumination to positive visualizations
- Avoidance to finding support from others

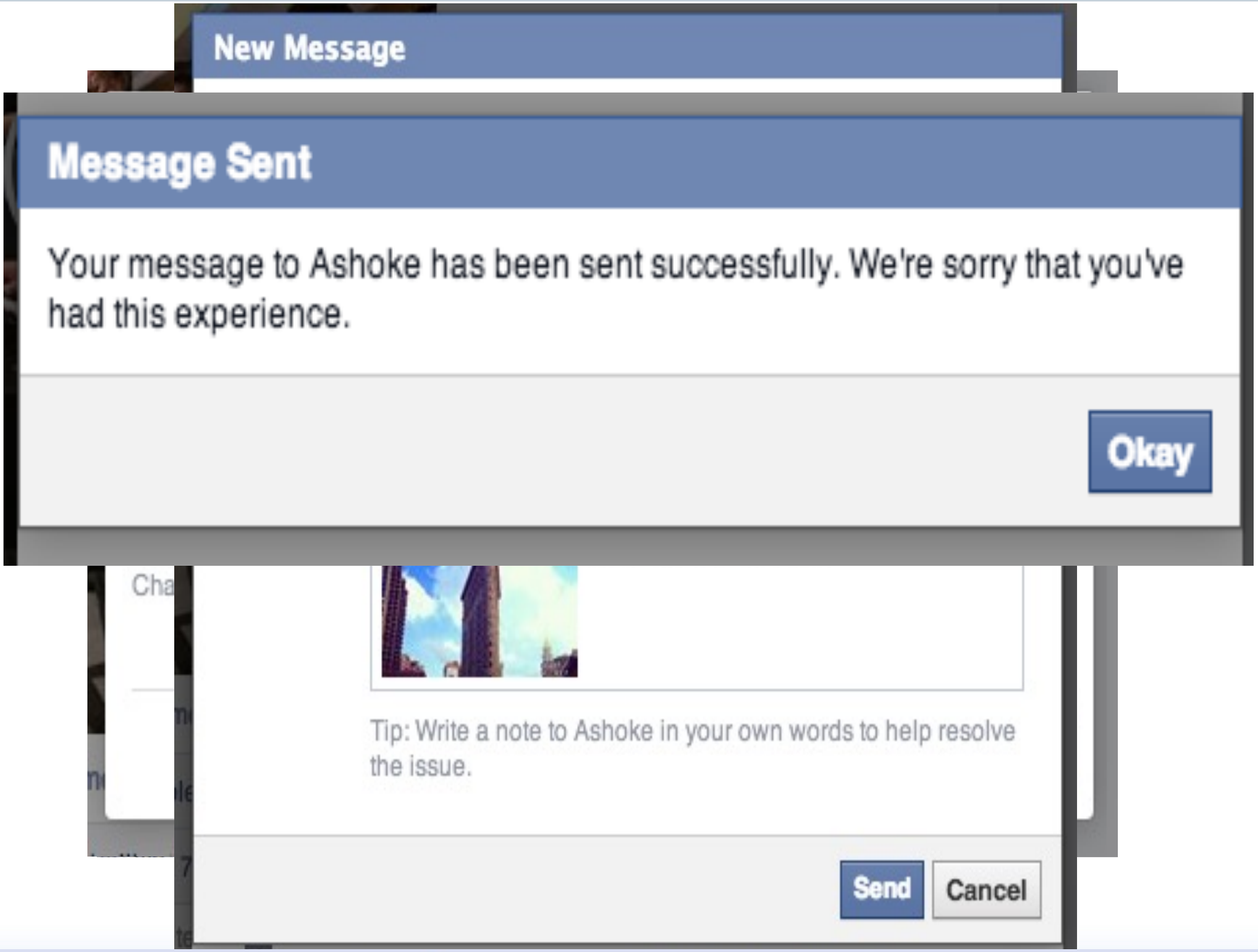
Facebook – Yale Collaboration

How can social media *incorporate design* that integrates emotional intelligence and developmental science to promote more positive online behavior among adolescents?

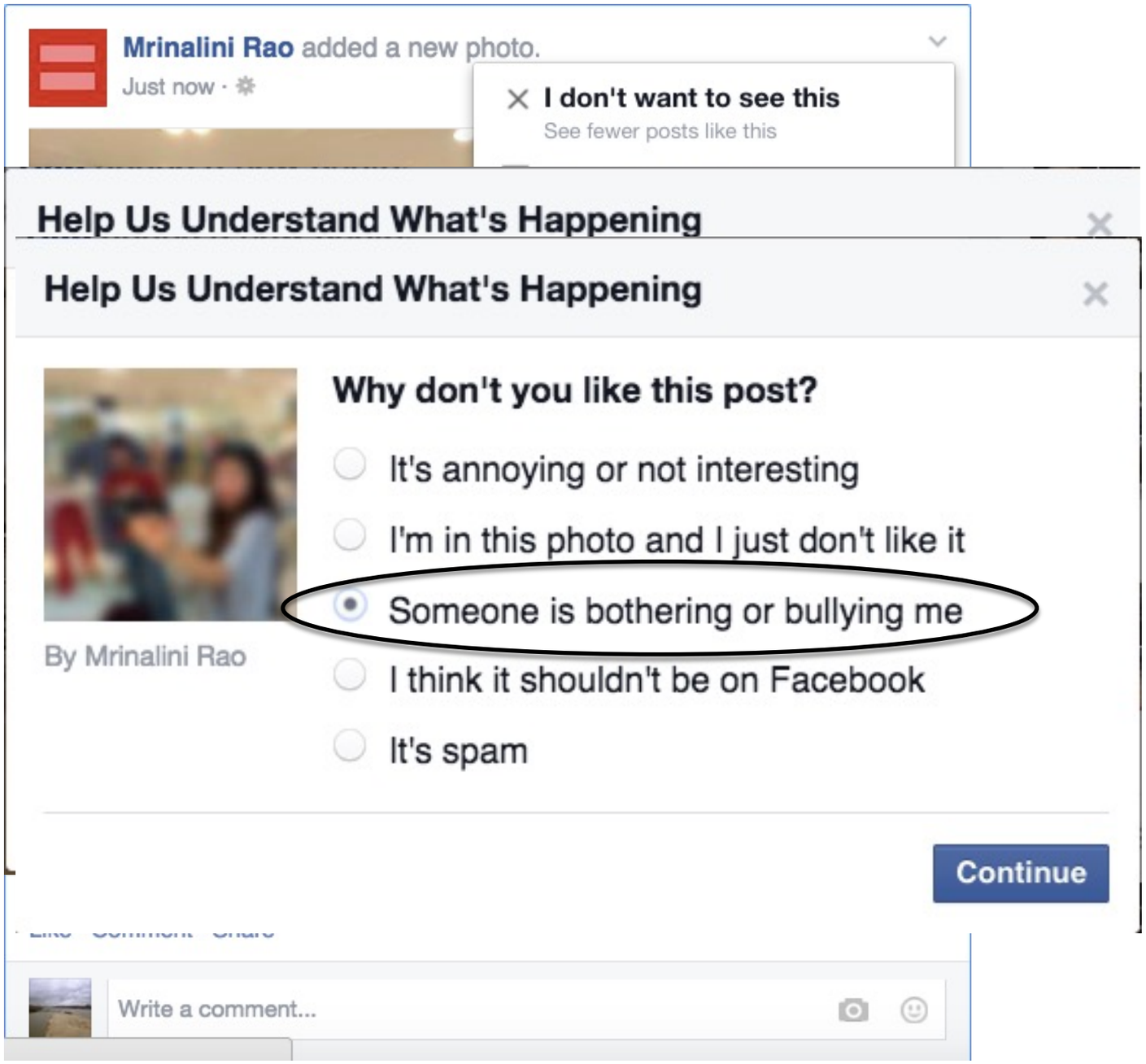
Applying EI to Facebook

- Initial focus was on building social resolution tools - Helping youth manage unpleasant experiences
- Began working with teens (13-18) in the U.S.
  - Consulted with teens and other stake holders
  - Used a developmental framework
  - Infused age-appropriate language
  - Incorporated emotion science

Social Resolution Tools



Research and Evaluation

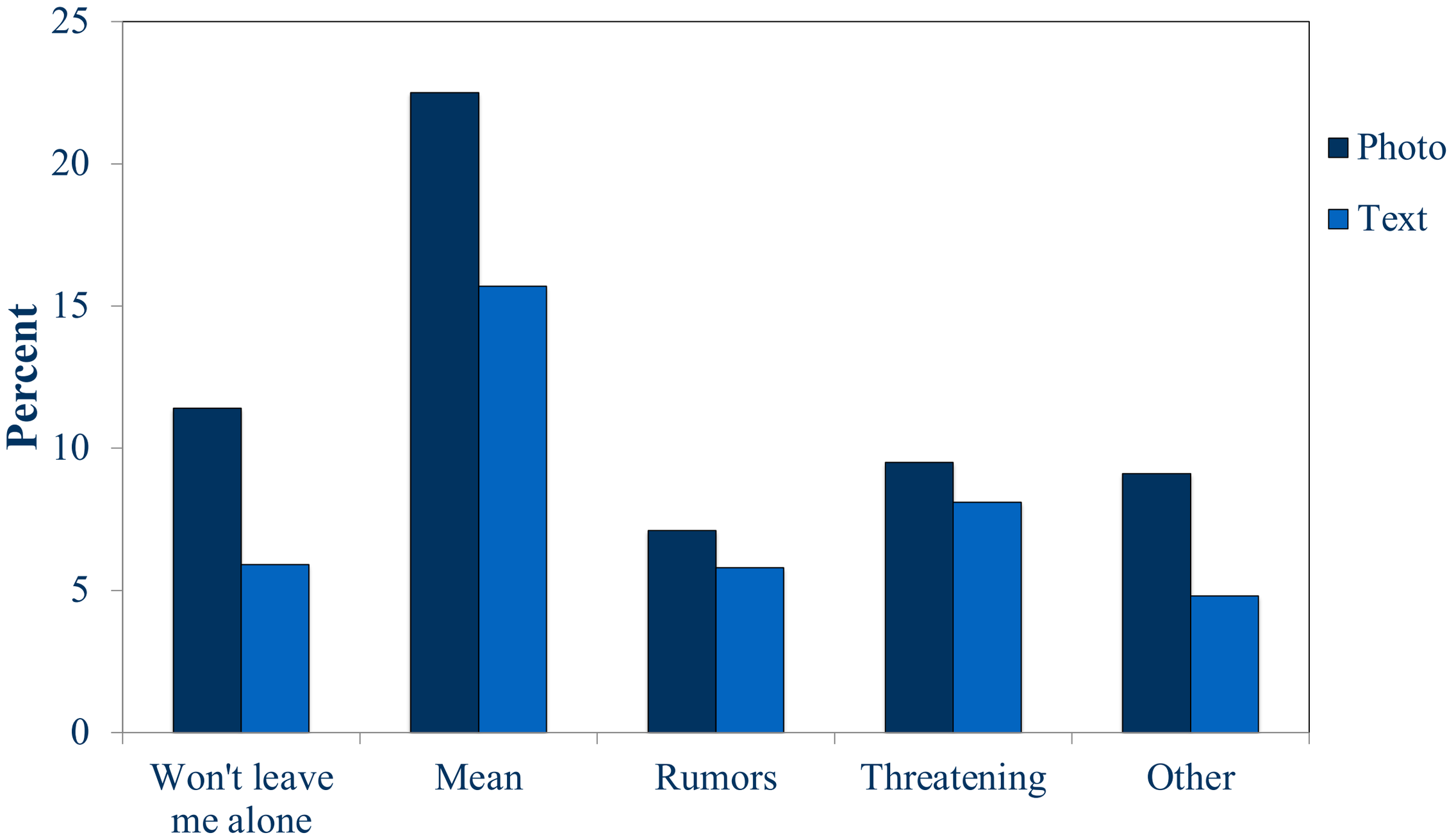


17.2 million events
(50-day period)

904,000 teens (5.2%)

59,311(6.6% or 0.3%)

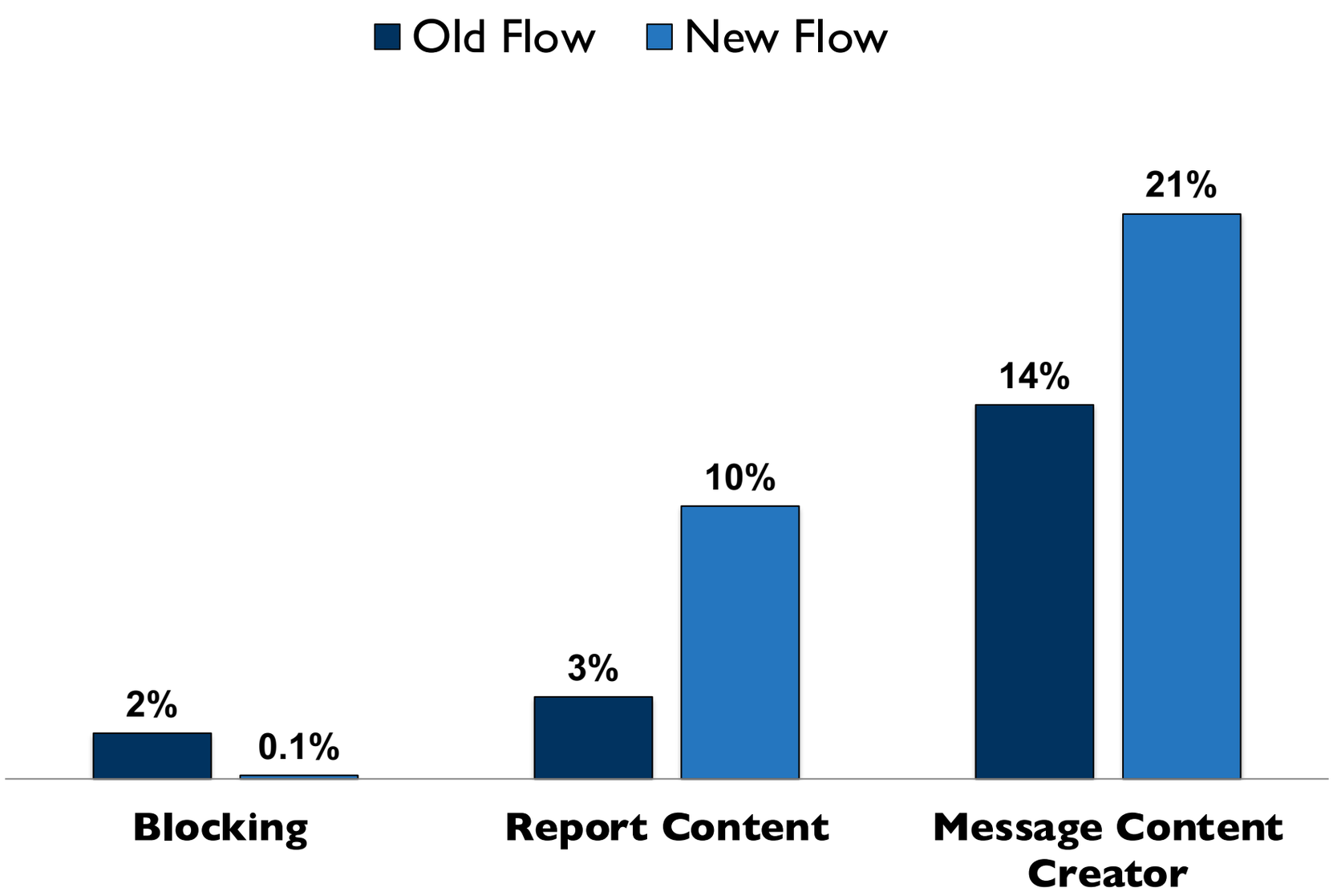
Bullying Behaviors



Emotional Reactions to Feeling Bullied

Emotional Reactions to Feeling Bullied

Influence on Behavior



*These data are from earlier pilot data, comparing adult flows to the revised flows

Facebook – A Global Company

How can we support teens in a developmentally and culturally responsive manner?

Culture and Emotions

- Individual differences
- Social norms
- Culture

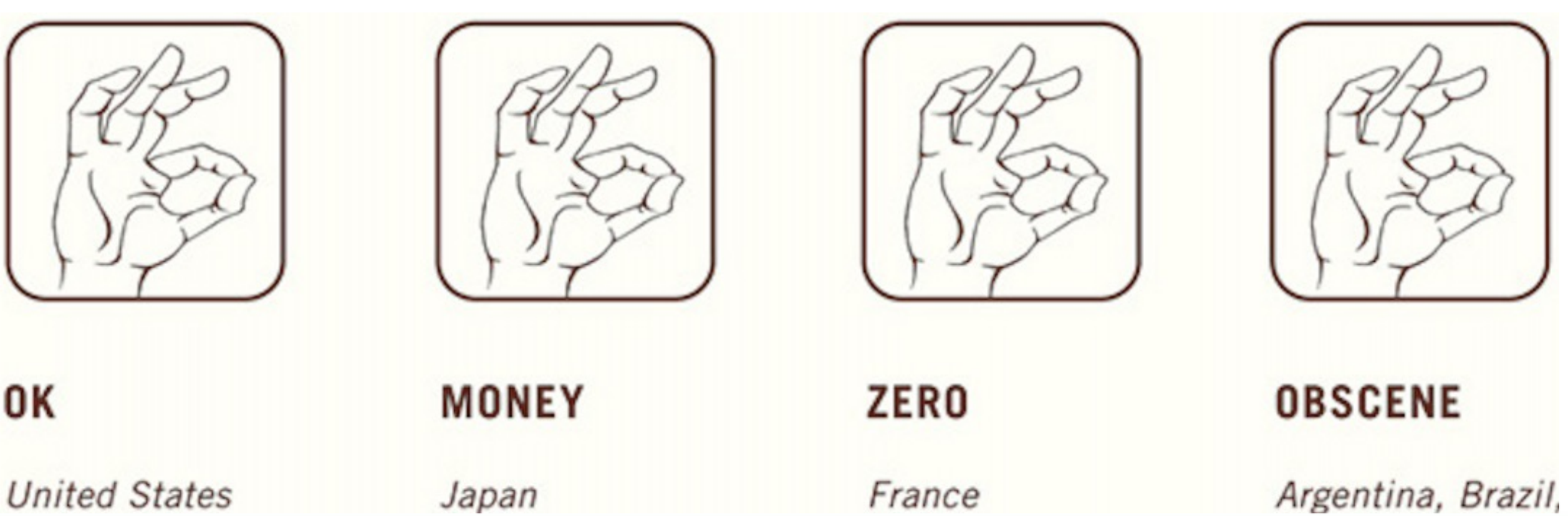
Cultural Display Rules

Culturally prescribed rules that govern how universal emotions can be expressed.

- Rules of social appropriateness
- Learned early in life
- Automatic practice by adulthood

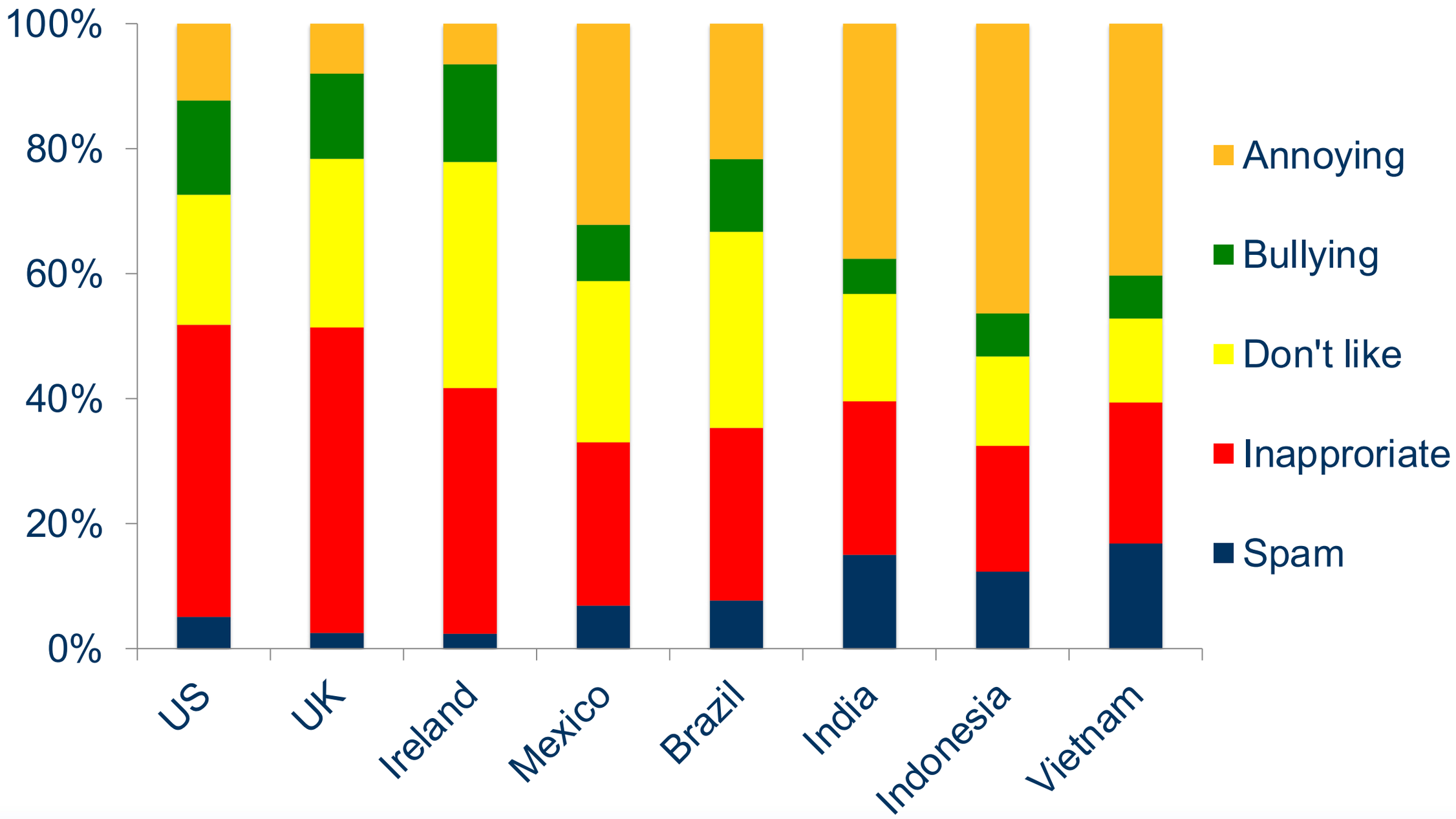
Cross-cultural differences

- Acceptable behavior
- Unwanted behavior
- Experience of emotion
- Social resolution

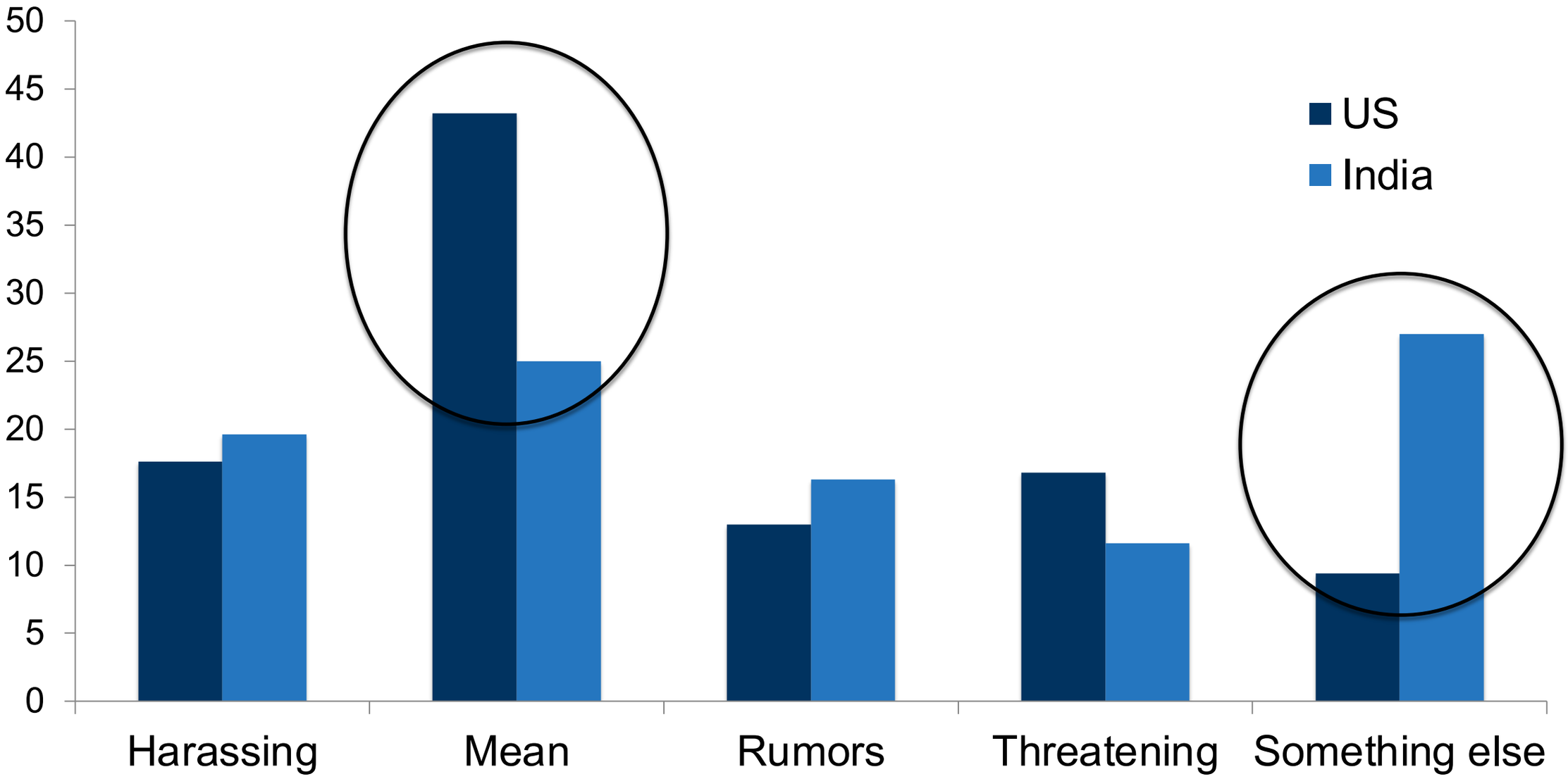


Understanding the Role of Culture and Language

Why don’t you want to see this?



What is happening in this post?



What is happening in this post?

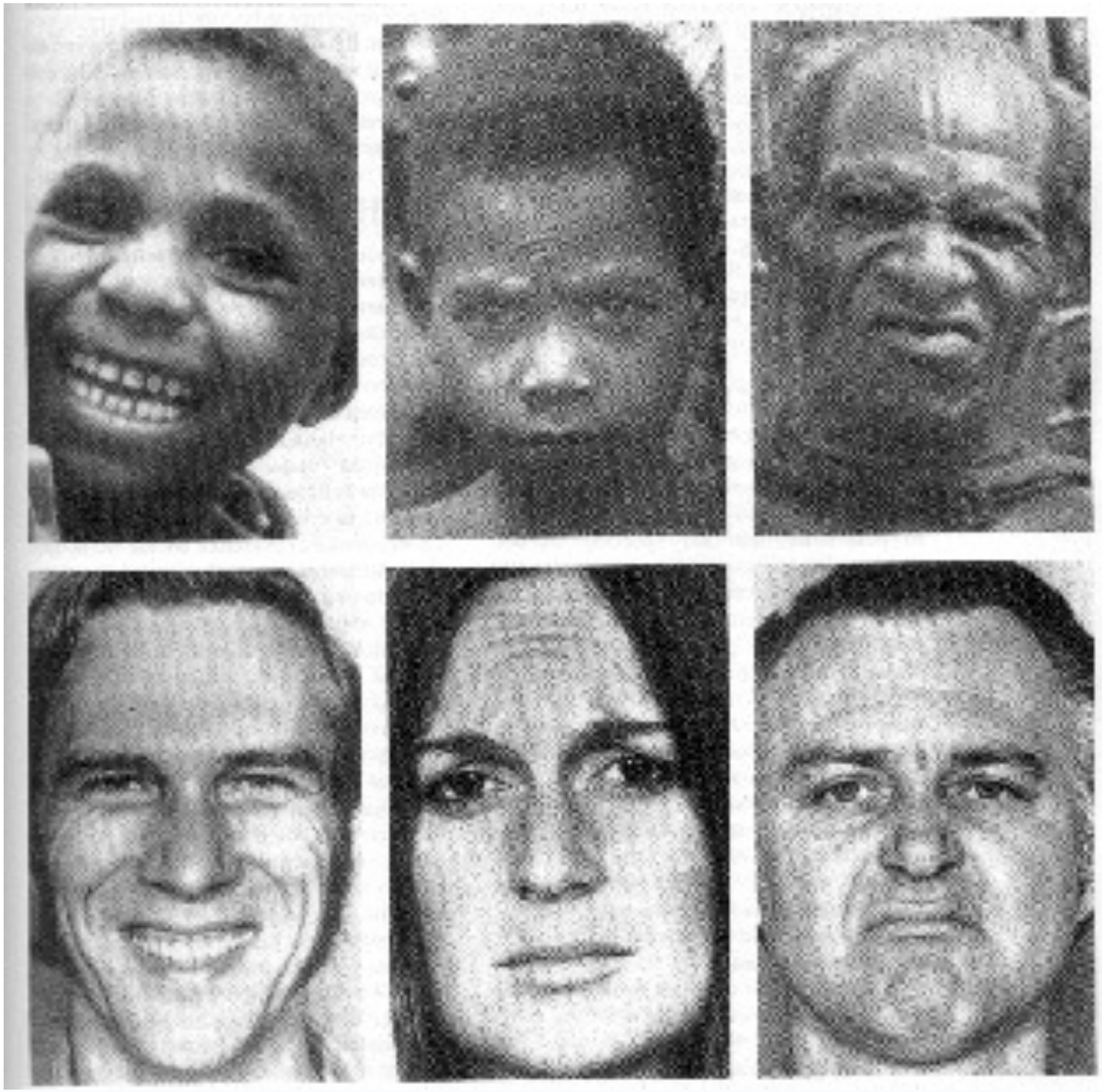
"Oh yes, we bully each other all the time. It's fun, it's drama."

"I bully my boyfriend's ex. Then I found out my boyfriend was still friends with her on Facebook, so I bully him about it too."

"Bullying is entertaining, I love it. I'm so "kepo" – I'll go their profile to see it"

- High school students in Indonesia

How does this make you feel?



Challenges interpreting cross-cultural findings

- What does bullying mean?
- How do behaviors like bullying impact the individual?
- What do people do offline when they are offended?
- What is the best way to facilitate resolution on Facebook?

Supporting Teens Across the World

- Aligning with teens' lived experiences.
- Learning from teens: What is going on?
- Offering online support that parallels offline cultural norms

In Conclusion

- Emotional experiences around “meanness” or bullying are universal. Behaviors that elicit the emotions vary culture to culture.

- There is a universal need to be seen, heard, and met

- The ways in which people desire to be seen, heard, and met vary as a function of culture

- Our goals are to investigate ways to promote both universal and culturally specific respectful, compassionate interactions online

Thank you

?