Tennessee v. Meta
Case No. 23-1364-IV
PX-00762

Document Provided in Native Format

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

Slide 1

1

@[redacted]@fb.com - If you have a "Safe" definition to add next to "Supportive" it would help ground this dimension as well for "reach", since this framework can cover bad experiences for both.

Feel free to add directly or as a comment here.
Assigned to otaru@fb.com
Sam[redacted] 11/19/2020 3:11:13 AM

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.

CONTEXT · **WHY BAD EXPERIENCES**

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

Slide 1

1 Is this true across problems or predominantly on bullying and hate speech where context matters a lot?
Miki Rothschild; 11/19/2020 5:42:46 PM

2 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).

Perception is broad.

People who take the effort to report content or otherwise self-remediate are in the middle.
Sam [redacted]; 11/19/2020 6:23:43 AM

1 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)
Shilpa Mody; 11/19/2020 5:42:46 PM

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

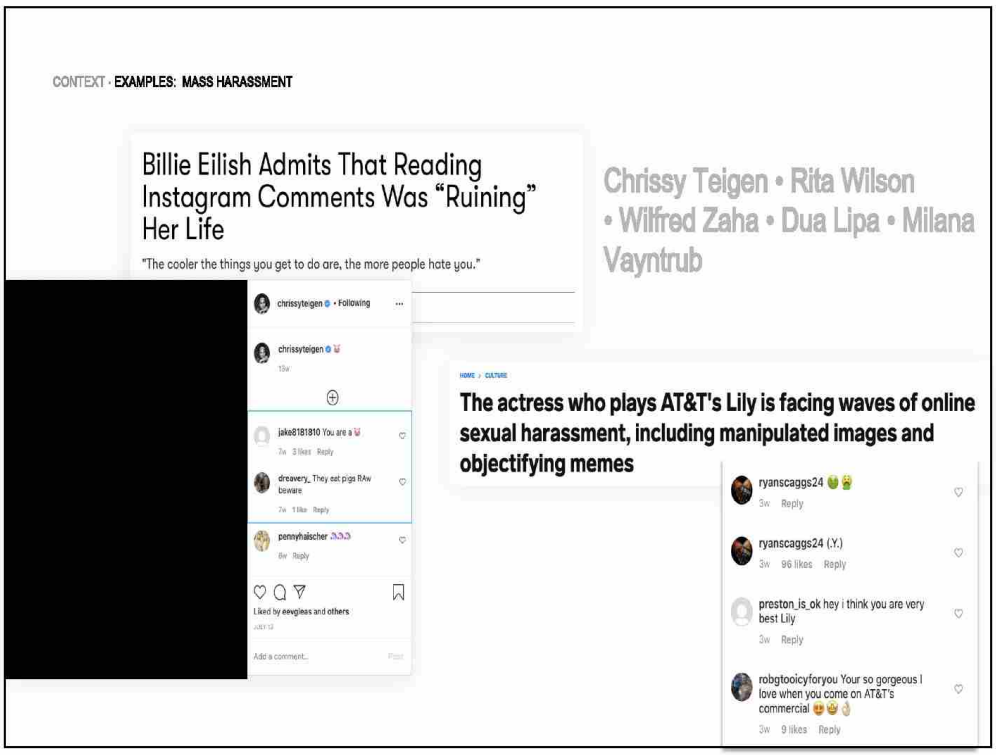
Very intense, low reach. Disproportionately affects creators. Portion isn't CS policy violating but still felt as intense.

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.



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

CONTEXT · **EXAMPLES: THREATS**

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

Comment



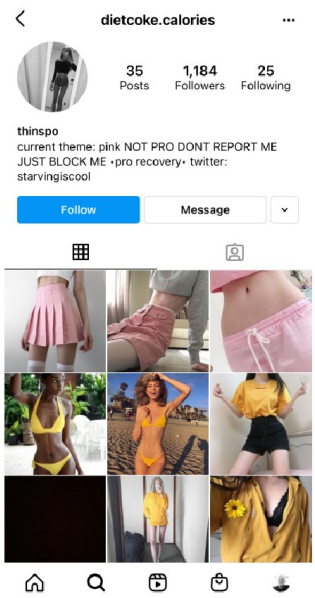
FIRST PERSON
Nicola Thorp Friday 24 Jul 2020 3:52 pm

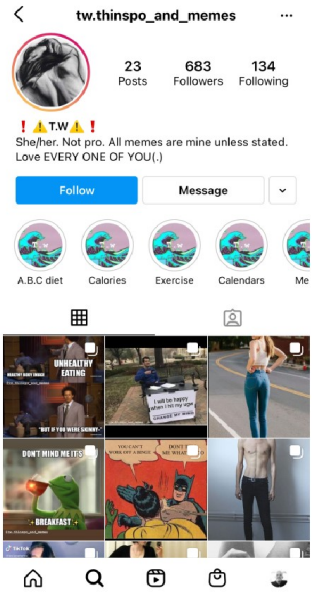
I've been on the receiving end of an unsolicited dick pic or two in my time. I laughed them off at first.

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.

CONTEXT - **EXAMPLES: THINSPIRATION**

No Image





"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

Slide 1

2 Why not intensity? Seems like it's as important as reach if we think working on something like mass harassment should be incentivised
Miki Rothschild; 11/19/2020 3:10:39 PM

3 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
Sam [redacted]; 11/19/2020 3:10:39 PM

4 Common question: Shouldn't we consider intensity and/or severity?

Both of these can fit inside reach to a degree. For example, how FB XI handles:

1- severity buckets for combining problems into "high harm" and "low harm" for TRIPS

2- "intensity-weighted reach" for TRIPS, where the binary "Yes, I saw in the last 7 days" is combined with the 5-point intensity scale to create metric that weights individual responses by intensity.

Definitions:

Intensity = felt experience by users

Severity = combines risk to business and risk of real world harm for people with felt intensity research
Sam [redacted]; 11/19/2020 3:14:36 PM

5 cc @[redacted]@fb.com
Sam [redacted] 11/19/2020 6:24:05 AM

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.

Slide 1

1 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?
Andrew [redacted] 11/19/2020 6:52:12 PM

6 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. There was some foundational analysis by CDS in H1. Probably harder to do attribution for but maybe our recent attribution work could help here.

cc @[redacted]@fb.com @[redacted]@fb.com
Sam [redacted]; 11/19/2020 6:20:39 PM

1 the recent switch is a good idea. We only look at whether it is currently private as a feature
Carlos [redacted] 11/19/2020 6:42:51 PM

2 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 how to get help
Andrew [redacted] 11/19/2020 6:52:12 PM

1 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?
Lori [redacted] 11/19/2020 7:59:52 AM

7 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).
Sam [redacted] 11/19/2020 7:59:52 AM

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)

Slide 1

3 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
Miki Rothschild, 11/19/2020 6:42:02 AM

8 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 scope of this review.

Here just trying to share some high level updates since we last discussed.
Sam [redacted] 11/19/2020 6:42:02 AM

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.

17

Slide 1

4 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
Miki Rothschild; 11/19/2020 6:55:13 AM

9 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.

This is just a light progress update on H2 progress this group may not be aware of in order to inform this 2021 topic.
Sam [redacted] 11/19/2020 6:55:13 AM

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.

Slide 1

5 I believe the bullying p0 would be oriented towards creators. Does that changes things or same measure?
Miki Rothschild; 11/19/2020 3:07:39 PM

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

For bullying measurement strategy, I think we'll want to consider:

1- balance of focus between "experiencing bullying" (targets) and "seeing" (witnesses)

2- how measurement execution details vary between creators, teens, and gen-pop

3- how much product strategy mirrors measurement strategy

For #2, there are going to be some measures that will have CIs that are too large if we directly filter to creators. But they can still be useful when triangulating with other measures with good granularity for creators.

For #3, from what we've seen with teens, we can inspire the product strategy from measurement of the problem and effectively develop products with a teen focus, then use more gen-pop measurement during launch experiment review as part of launching for everyone. So less of a full mirror between product and measurement, with more push to see the focus come through in the product strategy than the measurement.
Sam [redacted] 11/19/2020 3:07:39 PM

Discussion

Feedback on ambition and examples?

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

Feedback on approach?

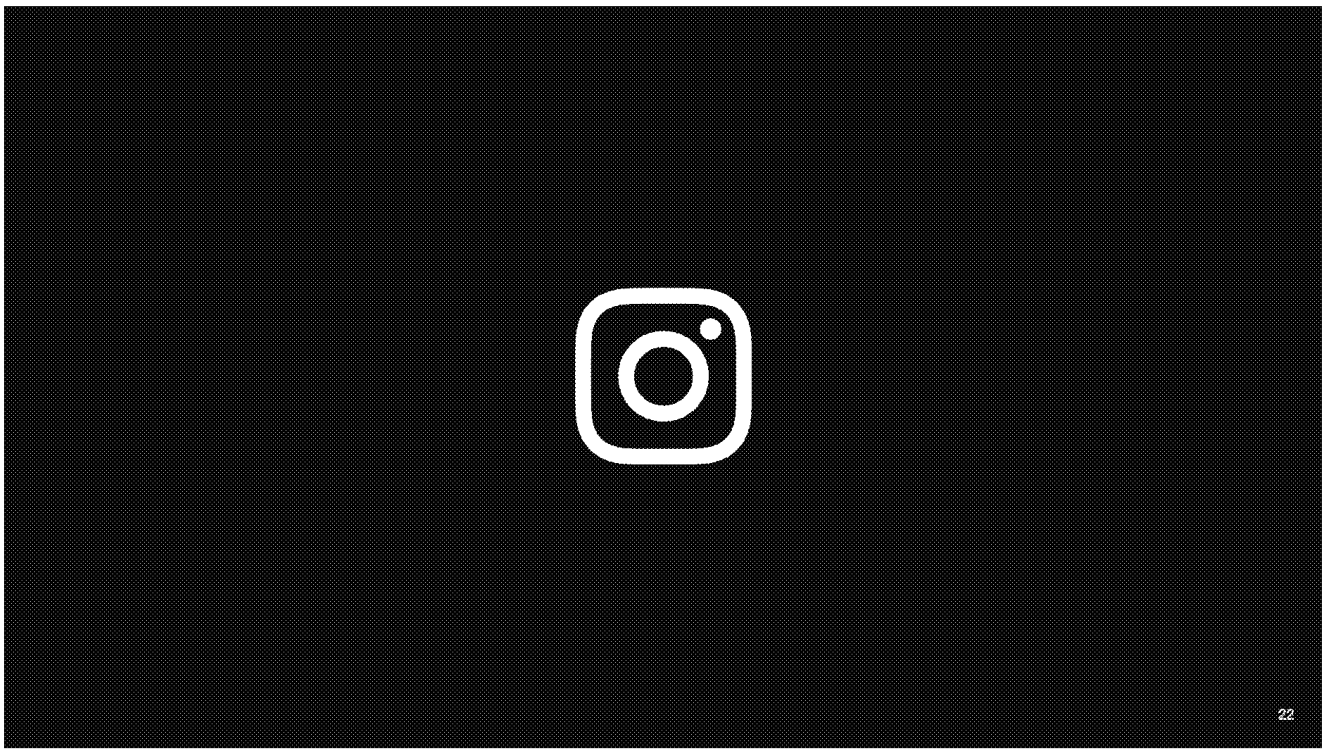
Is the framing intuitive? Leverage existing concepts and terminology for “extreme clarity”?

Feedback on next steps?

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

Slide 1

2	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? Lori [redacted], 11/19/2020 6:30:53 AM
6	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 and make a bet on one where we want to lead but not try to chase all of the failures with the rest of the pack because that'll be very inefficient. So the key question for me is: 1) which metrics are proven effective with great conviction that we should follow? 2) if we have to choose one metric that's so important to us that we're willing to make a bet on it, which one would that be? Miki Rothschild, 11/19/2020 6:49:04 AM
11	Yes, this feedback seems to create inputs for step #3 on slide 20 (prioritize based on what would create impact for our teams). Sam [redacted] 11/19/2020 6:49:04 AM


22

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

CONTEXT - **ASSESSING EXISTING METRICS**

12 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

Slide 1

12 Used as primary proxy metric for if problems were healthy during COVID surge and US2020.
Sam [redacted] 11/19/2020 5:20:43 PM

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
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

DOCUMENT SLIPSHEET

Bates Number: META3047MDL-019-00094148

BEGATTACH: META3047MDL-019-00094148

ENDATTACH: META3047MDL-019-00094148

Attachment Count: 0

PRODVOL: META3047MDL_VOL019

Custodian: ANDREWS_KYLE; BEJAR_ARTURO; CASTANEDA_DIEGO; [redacted]_CARLOS; [redacted]_HITOMI; [redacted]_LORI; [redacted]_SATISH; OTARU_SAYED; [redacted]_SAM; ROTHSCHILD_MIKI; [redacted]_BRIAN

File Path: /[redacted]_LORI; /ANDREWS_KYLE; /[redacted]_SATISH; /[redacted]_BRIAN; /[redacted]_HITOMI; /[redacted]_CARLOS; /[redacted]_SAM; /OTARU_SAYED; /CASTANEDA_DIEGO; /ROTHSCHILD_MIKI; /BEJAR_ARTURO

Confidentiality Designation: HIGHLY CONFIDENTIAL (COMPETITOR)

HASHVALUE: 805EB1FEB13775CA954AFCE575AE09B5

Author: [redacted]@META.COM; [redacted]@META.COM

Family Date: 2/1/2021 12:00 AM

Create Date: 11/18/2020 12:00 AM

Last Modified Date: 2/1/2021 12:00 AM

LAST MODIFIED BY:

TRACK CHANGES: N

COMMENTS: Y

HASHIDDENDATA: N

Filename: "BAD EXPERIENCES" MEASUREMENT - PLAN FOR A _1TSYGUNR6VVYFZRKGXGJ3BUOAUIICTNTT33OSPPK47T0.PPTX

Title: Â€ŒBAD EXPERIENCESÂ€ MEASUREMENT - PLAN FOR A PLAN - NOV 19 WB REVIEW,Â€ŒBAD EXPERIENCESÂ€ MEASUREMENT - PLAN FOR A PLAN - NOV 19 WB REVIEW

DOCEXT: PPTX

Email From:

Email To:

Email CC:

Email BCC:

Received Date: 12/31/9999 7:00 PM

Sent Date: 12/31/9999 7:00 PM

Subject:

THREADID:

REDACTION TYPE:

REDACTIONS: N