People Disagree Content Seen by Teens Reporting Different Levels of Body Dissatisfaction After Viewing Content on IG

Directional estimated group differences from content analysis with wave 1 MYST data

do not distribute internally or externally without permission; legal approval (POC: [redacted] received 7/24/24.

Context

This report presents findings from a qualitative review of content samples derived from Meta and Youth Social Emotional Trends (MYST), a longitudinal study that follows a cohort of 1,149 U.S. teens and their parents over the course of the 2023-2024 academic school year, linking their survey data with teens' behavioral data, in an attempt to learn more about teens' experiences on FB and IG. See here for additional background and analysis plan.

TL;dr

- Teens who reported frequent body dissatisfaction after viewing posts on IG (see items here) at wave 1 saw about 3 times more body-focused/ED-adjacent content than other teens, over a subsequent 3-month period.
    - Note: It is not possible, nor the intention of this analysis, to examine the causal relationship between reporting frequent body dissatisfaction and seeing people disagree content (see definition here). Rather, the goal is to understand descriptively what type of content teens who report body dissatisfaction are seeing while using IG.
- Of the reviewed content that met labeling guidelines for people disagree content, only 1.5% was captured by the specific set of reviewed integrity classifiers at scores above 0.03; none of the content labeled as body-focused was captured by the set of reviewed integrity classifiers at thresholds above 0.1.
- Findings provide qualitative, directional evidence suggesting that there is a swath of content on IG that:

    ○ Is not captured by the subset of reviewed existing integrity classifiers even at lower thresholds than are currently used in production (note: they may be captured at much lower thresholds, and by classifiers not used in production in H12024, when this data was gathered and analyzed)

    ○ May be seen disproportionately by teens who report frequently experiencing body dissatisfaction after viewing content on IG; and ○ Is identified using body-focused

content labeling guidelines and that overlaps substantially with eating disorder-adjacent content, that external advisors have expressed support for limiting for teens.

Background

Prior non-causal internal research (see here, here, and here) has demonstrated an association between frequency of reporting feeling worse about one's body after viewing others' posts on IG (AKA, engaging in appearance comparison) and viewing more of a specific set of FIT topics, primarily related to fashion and beauty, on IG. In recent qualitative and quantitative research, teens and parents report that content that emphasizes bodies and appearance, or that promotes changing one's appearance, may be sensitive, or at least somewhat inappropriate for some teens. Additionally, a common theme we have heard in multiple engagements with external stakeholders (i.e., pediatric professional organizations; external advisors with expertise in eating disorders, adolescent mental health, and digital well-being) is concern about teens viewing content that focuses on bodies and appearance, especially in high concentrations, because this content may be detrimental to teen well-being, specifically by precipitating or exacerbating feelings of body dissatisfaction. Recently, members of Safety Policy's Eating Disorder & Body Image Advisory Council expressed support for expanding IG's search intervention suite to help limit the "eating disorder-adjacent" content that is seen by users, especially teens; this feedback was used to help develop and refine the labeling guidelines for body-focused content, and thus these two types of content overlap (see working definition for body-focused content and framework for eating disorder-adjacent content here).

Recently, labeling guidelines have been developed to, for the first time, define and objectively identify body-focused content attributes, triangulating across signals from teens, parents, external stakeholders, and the prior appearance comparison literature. In order to be maximally useful to support classifier development, these guidelines operationalize body-focused content by honing in on specific types of content related to bodies and faces (i.e., content depicting judgment/comparison or change in bodies/faces and content that prominently displays ideal or sexualized bodies; conceptually, this content overlaps substantially with eating disorder-adjacent content as well). While this specificity and objectivity supports classifier development, it is unclear whether the content captured by these guidelines is associated with the kinds of well-being measures that prior research has focused on (i.e., appearance comparison) and that parents, teens, and external stakeholders are concerned about (i.e., "erodes sense of self", and body dissatisfaction, respectively). To begin to close this gap in understanding, the goal of the analysis outlined in this doc is to use the research-derived people disagree content framework (formerly known as teen sensitive content), including objective labeling guidelines for body-focused content, to provide a directional assessment of whether teens in the MYST study who report frequently experiencing body dissatisfaction after viewing others' posts on IG may also see more body-focused content and more eating disorder-adjacent content than other teens in the study.

Method

As part of the Meta & Youth Social Emotional Trends Study, between Sept-Dec 2023, a group of n=1,149 US teens completed a comprehensive survey that included questions about their

experiences, emotions, and perceptions. Their survey responses were paired with server log data of which content they viewed. Teens completed several questions that assess the extent to which and frequency with which they experience body dissatisfaction after viewing others' posts on IG: how often they compare themselves to others on IG, and how often they feel worse about their own bodies after engaging in such comparison (see item text and responses here). Assessing the extent to which teens experience body dissatisfaction after viewing others' posts on IG enables evaluation of body dissatisfaction that is maximally proximal to teens' experiences with content on IG, as opposed to their sense of body dissatisfaction in life more generally (note: in this study, teen experience of body dissatisfaction while using IG is also correlated with more general body dissatisfaction, alpha=0.39, though the moderate correlation suggest these measures are capturing different aspects of teens' experiences). Teens' responses to IG-specific body dissatisfaction items were combined to form a scale, which has been used in prior work (scale alpha= 0.88).

For the purpose of this analysis, teens in the sample were grouped into two groups: 1) teens who scored high on the IG content-specific body dissatisfaction scale (i.e., reported that they often compared themselves to others on IG, and frequently felt worse about their bodies after doing so; n=223 teens), and 2) all other teens in the sample (n=795 teens). We then generated a random, VPV-weighted sample of 500 pieces of content viewed by teens in each group over the 3 month period ending 5/12/24 (see detailed description of this process here).

A well-being subject matter expert researcher systematically labeled all available pieces of sampled content for each group, applying the people disagree content framework (formerly known as teen sensitive content), the body-focused content labeling guidelines, and the ED-adjacent content framework. Counts, and percentages, of each people disagree content topic, including body-focused content, were calculated for both groups; differences in group estimates for each topic were then tested statistically using two-tailed t-tests. These results represent directional, qualitative, estimated prevalence of teen sensitive content topics for US teens in this sample who report frequent body dissatisfaction after viewing other people's posts. These estimates may not be generalizable to the IG teen-using population.

Results

Three-quarters (74%) of teens who reported frequent IG content-specific body dissatisfaction at wave 1 identified as female, compared to 50% of the teens who reported none to occasional IG content-specific body dissatisfaction. Roughly half of the teens in each group were early teens (ages 13-15 years old; 54% in the high group and 50% in the low-to-moderate group).

On average, teens who report frequent IG content-specific body dissatisfaction at wave 1 may have seen almost twice as much People Disagree content compared to other teens, over a 3 month period. Specifically, teens who report high IG content-specific body dissatisfaction at wave 1 may have seen almost three times as much body-focused/ED-adjacent content compared to other teens (see content examples here), with the group difference driven by significantly more content that meets criteria for the prominent body display sub-theme of body-focused content. It is not possible to establish the causal direction of these findings (e.g., teens who report high IG content-specific body dissatisfaction may also be more likely to seek out these types of

content, or a 3rd factor may explain the observed patterns)**.** There were no categories of People Disagree content that were viewed more frequently by teens who did not report high IG content-specific body dissatisfaction. Of all the content reviewed across both groups, only three pieces of People Disagree content received an existing integrity classifier score above 0.03 (i.e., these pieces of content were labeled as Mature Themes, specifically the subtopic Explicit Language); the integrity classifier scores received by these pieces of content was approximately 0.99). This is not necessarily surprising, given that People Disagree content captures unrestricted content that is not covered by Meta's current content policies. In H124, a team was assembled to enable measurement of the People Disagree content space; this team is developing a new classifier to identify this content (currently on its 5th iteration). In future analyses, the MYST study team will complement the current label-based analysis with analyses that utilize classifier-based People Disagree content.

Estimated prevalence of people disagree content seen by teens with high IG content-specific body dissatisfaction compared to teens with none to occasional IG content-specific body dissatisfaction

(n=500 pieces of random, VPV-weighted content in each group)

| **People Disagree content topic** | **Estimated prevalence of sampled content seen by teens in group 1:** | **Estimated prevalence of sampled content seen by teens in group 2:** |

[Breakdown by sub-theme here]	Frequent IG content-specific body dissatisfaction (n=223 teens)*	All other teens (n=795 teens)*
Mature Themes	n=42 (8.4%)	n=32 (6.6%)
Body-focused Content/ED-adjacent content**	**n=52 (10.5%)**	**n=16 (3.3%)**
Risky Behavior	n=18 (3.6%)	n=11 (2.2%)
Harm & Cruelty	n=10 (2.0%)	n=4 (0.8%)
Suffering	**n=13 (2.6%)**	**n=2 (0.4%)**
Total People Disagree Content	**n=135 (27.3%)**	**n=65 (13.6%)**

**Due to content availability, 495 and 485 pieces of sampled content were substantively reviewed in group 1 and group 2, respectively.*

***In 2-tailed z-test, between group differences were statistically significant for Body-focused Content and Suffering topics, as well as total People Disagree Content overall ($p<0.00001$, $p=0.0048$, and $p<0.00001$, respectively).*

****ED-adjacent content is a superset of body-focused content; all body-focused content was also labeled as ED-adjacent.*

Implications

Results from this analysis provide qualitative, directional evidence suggesting that there is a swath of content, that: 1) is not captured by our existing integrity classifiers even at low thresholds; 2) may be seen disproportionately by teens who report frequent IG content-specific body dissatisfaction; and 3) is identified using the people disagree content framework in general, and body-focused content labeling guidelines/ "eating disorder-adjacent" content in specific (which external advisors have expressed support for limiting, especially for teens).

Next Steps

In follow-up analysis, we plan to: 1) examine potential quantitative associations between the People Disagree content classifier and teen well-being measures, including IG content-specific body dissatisfaction, over time; and, 2) extend this analysis to look at any potential group differences for teens who report frequent/high vs. low/none to occasional levels across other well-being measures (i.e., emotional well-being, social well-being, potential problematic use, and positive app value).

Appendix

IG Content-specific body dissatisfaction items:

IG content-specific body dissatisfaction (also known as appearance comparison) is a scale that takes the mean across responses to two items that assess how frequently teens compare themselves to others on IG, and how frequently they experience body dissatisfaction after viewing other people's posts on IG.

Item 1: How often do you compare your appearance to the appearance of people on the following social media app(s)?

Item 2: How often do you feel worse about your appearance after seeing posts on the following social media app(s)?

Response options for both items were Never (0), Rarely (1), Sometimes (2), Often (3), Extremely Often (4).

In this analysis, teens were bucketed into group 1 if they responded "Often" or "Extremely Often" on both items, and into group 2 if they responded "Never", "Rarely", or "Sometimes" to either item.

People Disagree Content (formerly known as Teen Sensitive Content):

"People Disagree content" is unrestricted content that some parents (as well as teens and experts) are not aligned with teens seeing, especially in high quantities. Alignment on the appropriateness

of this content for teens can vary across individuals and cultures. People disagree content captures content that is NOT part of our current content policies.

Working Definitions/frameworks [WIP]:

Body-focused content [details here]: Body-focused content includes any content in which there is a prominent display of body shapes or sexualized body parts (specifically, chest, buttocks, or thighs). It also includes content in which there are explicit judgments or comparisons of body shapes and compositions, or in which products or behaviors are promoted for their ability to change faces, body shapes, or compositions (including weight loss).

Eating Disorder-Adjacent content [details here]: Content related to topics that do not explicitly reference eating disorders (ED), and are thus not classified as violating or borderline, but which may be related to, and/or may be triggering to someone experiencing disordered eating (i.e., body-focused, weight loss, dieting, health-related). This may include content related to disordered eating and/or negative body image, or ED topics that fall below the threshold for violating or borderline content policies due to the scope of these policies (e.g., pica, compulsive overeating) or the specified strength of signaling ("goal weight" without mention of "current weight").

[WIP] Eating Disorder-Adjacent Topics
ED Recovery
Negative body image. Referencing a thin or muscular ideal (e.g., "#skinnygirl"), fat phobia (e.g., "#fatwhale"), body-shaming (e.g., "#fatarms"), body surveillance (e.g., "body check"), body alteration (e.g., before and after images/videos).
Disordered eating beyond restrictive eating & binge/purge types. Compulsive overeating, night eating, food preoccupation, food addiction, sugar addiction
Extreme health, nutrition, and fitness advice, including diet supplements, juice cleanses, losing weight fast, or fitness programs that are unsustainable or overly focused on body image (e.g., "bikini body"), fitspo, healthspo, "what I eat in a day", "what [person] eats in a day", dieting, fasting/starving, use of weight loss medications/diet pills, muscle-building supplements, "flat tummy [product]", nutrition or fitness content in the context of weight loss, rules-based eating (e.g., intermittent fasting), named diets ("keto diet", "clean eating").

Breakdown by People Disagree Content Topic Sub-theme

	Estimated prevalence of sampled content seen by teens in group 1: Frequent ID content-specific body dissatisfaction (n=223 teens)*		Estimated prevalence of sampled content seen by teens in group 2: All other teens (n=785 teens)*	
	n	%	n	%
Mature Themes	**42**	**8.40%**	**32**	**6.60%**
Explicit Language	16	3.20%	15	3.00%
Incidental Explicit Language	0	0.00%	0	0.00%
Sexual Themes	20	4.00%	10	2.10%
Political/Controversial	6	1.20%	8	1.70%
Adult use of substances	0	0.00%	0	0.00%
Body-Focused Content/ED-adjacent Content	**52**	**10.50%**	**16**	**3.30%**
Prominent Body Display	37	7.50%	8	1.70%
Chest or Buttocks Focus	10	2.00%	5	1.00%
Judgment or Change	5	1.00%	3	0.60%
Rules-Based Eating	0	0.00%	0	0.00%
Risky Behavior	**18**	**3.60%**	**11**	**2.20%**
Stunts/pranks	3	0.60%	2	0.41%
Substance Use	9	1.80%	5	0.10%
Reckless Driving	4	0.81%	1	0.21%
Tattoos	0	0.00%	0	0.00%
Challenges/dares	0	0.00%	0	0.00%
Gambling	2	0.40%	0	0.00%
Criminal Behavior	0	0.00%	0	0.00%
Harm & Cruelty	**9**	**2.00%**	**4**	**0.80%**
Mocking	0	0.00%	0	0.00%
Violence	0	0.00%	0	0.00%
Weapons	0	0.00%	0	0.00%
Mistreatment/Disrespect	9	1.80%	3	0.62%
Toxic Narratives	1	0.20%	1	0.21%

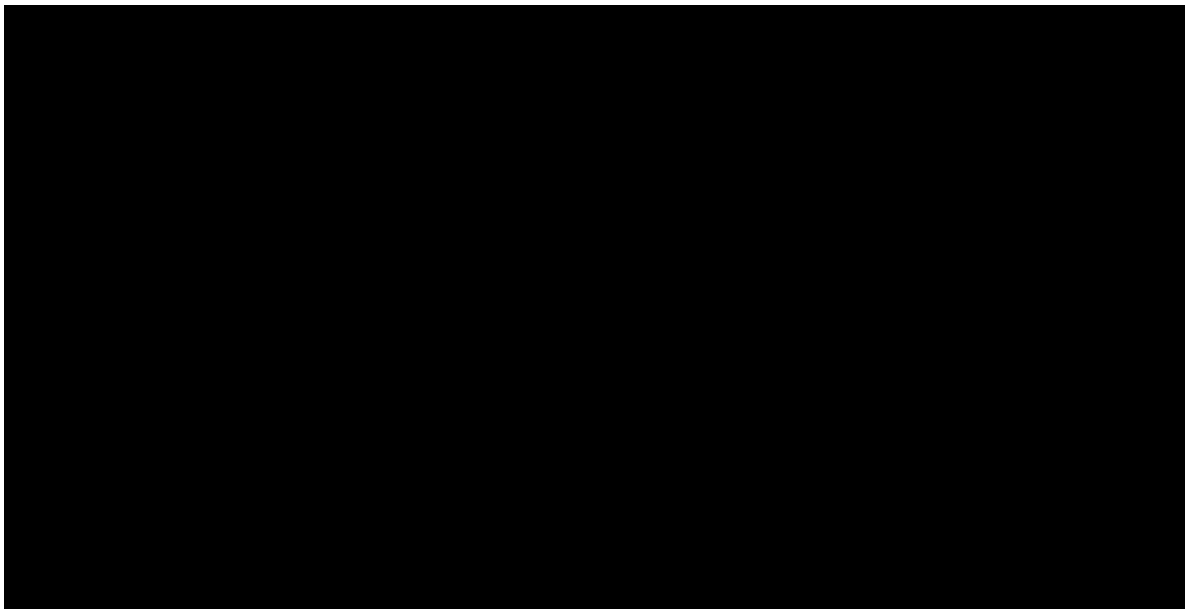
Suffering	**13**	**2.60%**	**3**	**0.40%**
Mental Health Challenges	2	0.40%	1	0.21%
Hardship/harsh realities	2	0.40%	0	0.00%
Death, pain, illness, injury	9	1.80%	2	0.41%
Total Unaligned Content	**135**	**27.30%**	**66**	**13.60%**

*Note: Due to content availability, 495 and 485 pieces of sampled content were substantively reviewed in group 1 and group 2, respectively.

This table shows the detailed breakdown of all People Disagree content categories and subcategories, with the specific counts (n) and percentages (%) for each group. The bolded rows indicate the main categories, while non-bolded rows show the subcategories within each main category.

Content Examples for Body-focused and Suffering topics [Warning: Sensitive Content]

Body-focused content



Suffering