A screenshot from the Pokémon game showing a character with a Rattata. A speech bubble contains the text: "Hi! Do you want to trade your Rattata for my Rattata?".

Hi! Do you want to trade your Rattata for my Rattata?

Integrity Tradeoffs

MONDAY, JANUARY 6, 2020 · READING TIME: 5 MINUTES

As we wrap up 2019H2, I want to publish one last bit on Integrity tradeoffs (for some other work on this, see [The incremental impact of Integrity demotions](#)). The goal of this note is to use simulations to map out the relationship between the strength of Integrity demotions and the impact on various engagement metrics.

Credit to [REDACTED] for the idea.

TLDR

Chats

REDACTED FOR CONGRESS

Integrity Tradeoffs

MONDAY, JANUARY 6, 2020 · READING TIME: 5 MINUTES

As we wrap up 2019H2, I want to publish one last bit on Integrity tradeoffs (for some other work on this, see [The incremental impact of Integrity demotions](#)). The goal of this note is to use simulations to map out the relationship between the strength of Integrity demotions and the impact on various engagement metrics.

Credit to [REDACTED] for the idea.

TLDR

- I ran a number of OVM simulations where I tweaked the strength of Integrity demotions and compared to a ranking config where Integrity rules were not enforced on.
- **Currently, Integrity demotions are responsible for ~10% of the VPVs seen, i.e. 10% of VPVs change as a result of Integrity demotions. Our maximal impact is ~17%.**
- For WYT, the stronger the Integrity demotion, the greater the WYT!
- Comments decline significantly as Integrity demotions increase (likely due to EB demotions), while MSI actually peaks at the status quo, which suggests MSI is heavily tuned right now given Integrity demotions.
- Integrity is not terribly correlated with engagement or WYT and we're not making significant tradeoffs with these metrics, at least not on average. It's possible we want to limit the amount of times where Integrity makes WYT or MSI substantial

Chats

REDACTED FOR CONGRESS

demotions), while MSI actually peaks at the status quo, which suggests MSI is heavily tuned right now given Integrity demotions.

- Integrity is not terribly correlated with engagement or WYT and we're not making significant tradeoffs with these metrics, at least not on average. It's possible we want to limit the amount of times where Integrity makes WYT or MSI substantially worse and that we should tune accordingly, but the average impact is small. These types of measures don't really give us much of a meaningful tradeoff, at least right now, and suggest that we might need to focus more on producers or other measures of what users want to see.

Methodology behind the OVM Simulations

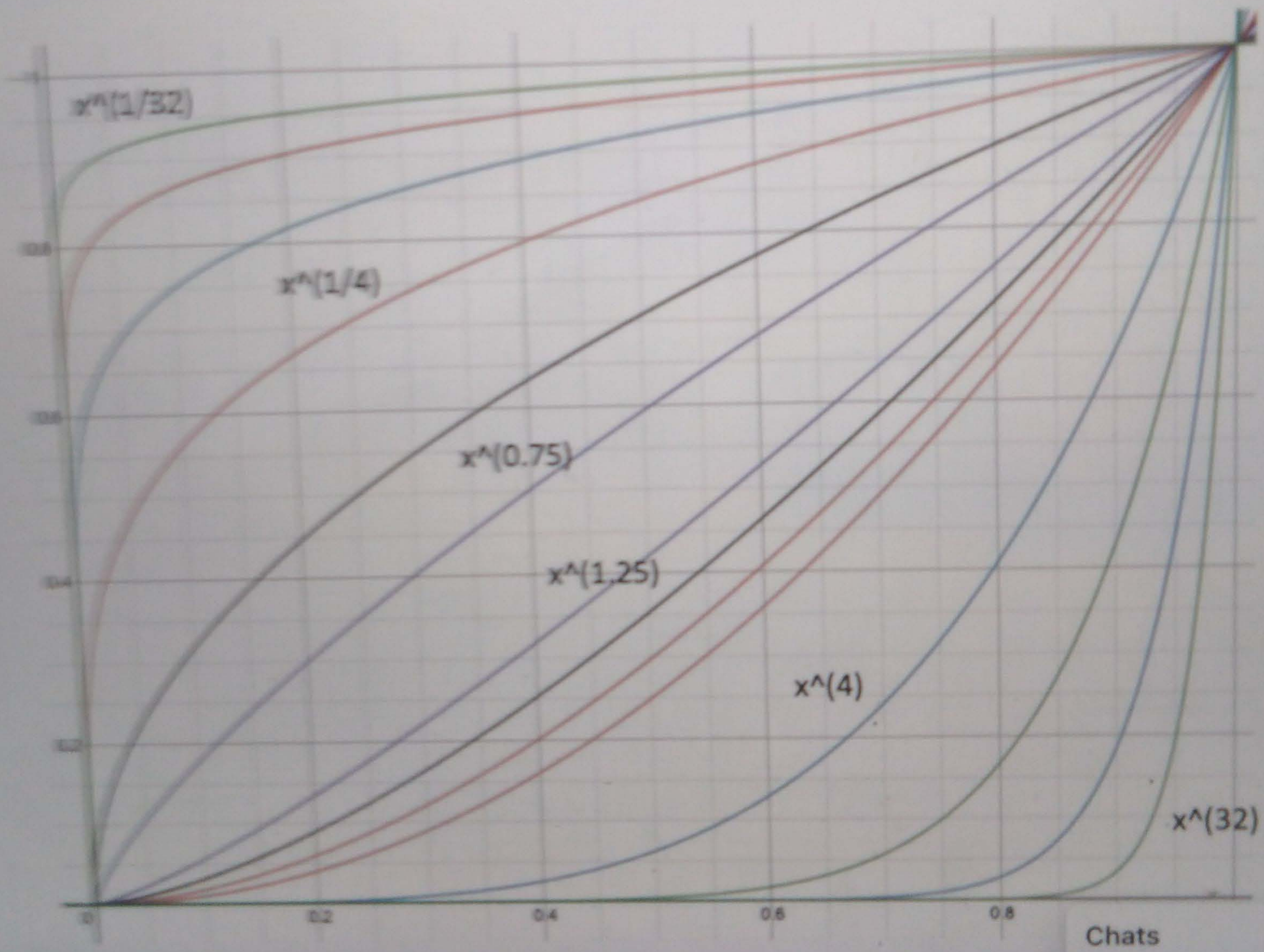
How would MSI or WYT change if we increased the strength of our Integrity demotions, relative to a ranking in which no Integrity rules are applied? That is the question I am attempting to answer with the OVM simulations that I ran.

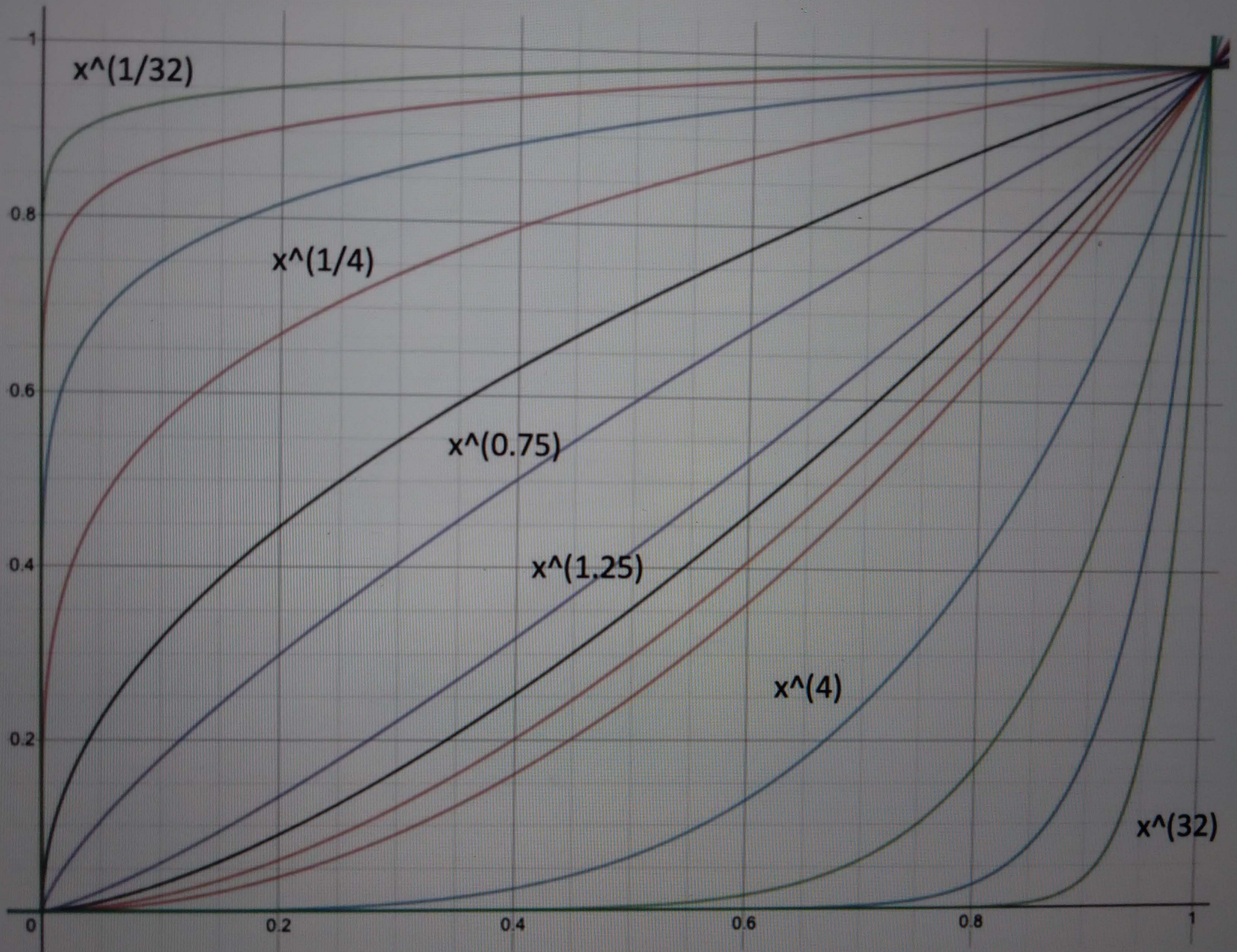
Relative to a ranking config where all Integrity rules are removed, I adjusted the strength of the Integrity demotions from $(1/32)x$ of the current size to $32x$ the current size (and lots of values in between). For the multiplicative demotion, this means that when the demotion is:

- $(1/32)x$ the current size: all demotions are scaled back to $\sim 0\%$ impact, e.g. even if content would have been demoted by 80% in the current production config, it will now only be demoted by $\sim 5\%$
- $32x$ the current size: all demotions are scaled to $\sim 100\%$ impact, e.g. even if content was just going to receive a 20% demotion in the current production config, it will now be demoted by 100%

→ y_{22} : the current size: all demotions are scaled to -100% impact, e.g. even if content was just going to receive a 20% demotion in the current production config, it will now be demoted by 100%

I roughly traced out various curves that approximate different amounts of scaling back / scaling up.





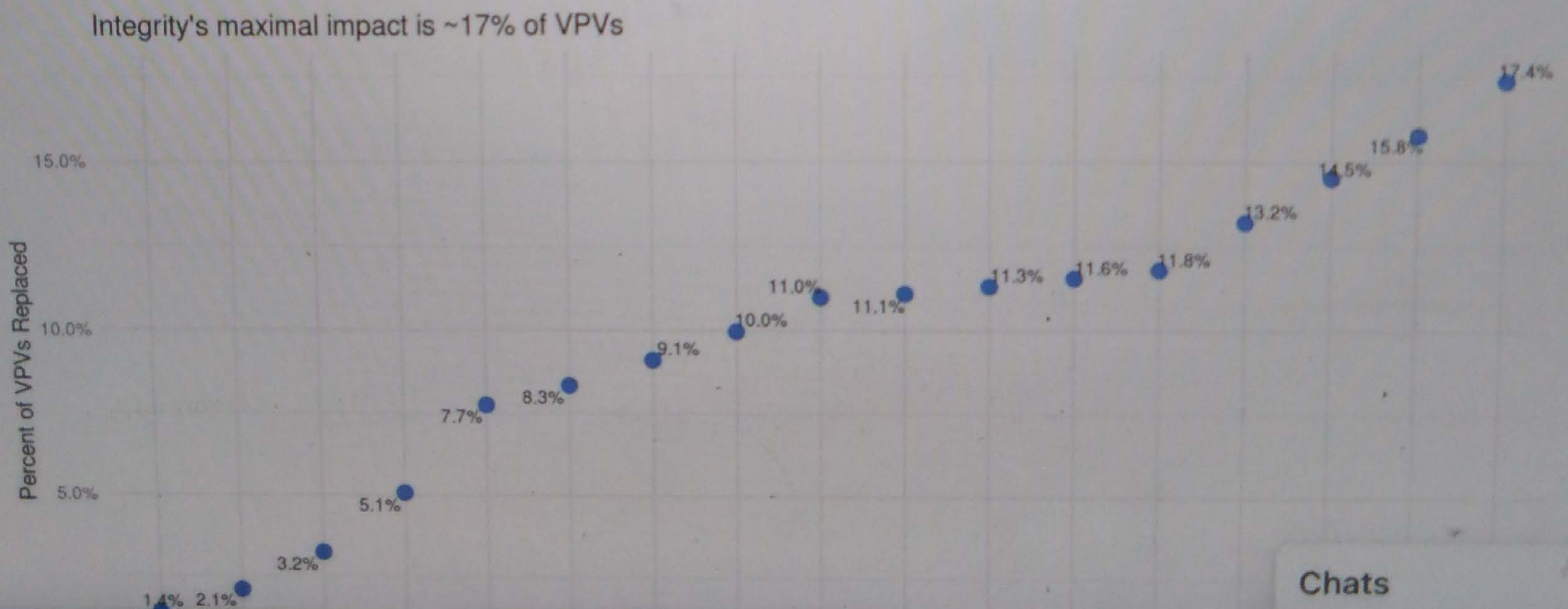
REDACTED FOR CONGRESS

In addition to the multiplicative scaling, I also scaled the viewed component by simply multiplying the viewed demotion by $(1/32)$ if the scaling is $(1/32)x$, 32 if the scaling is $32x$, and so forth.

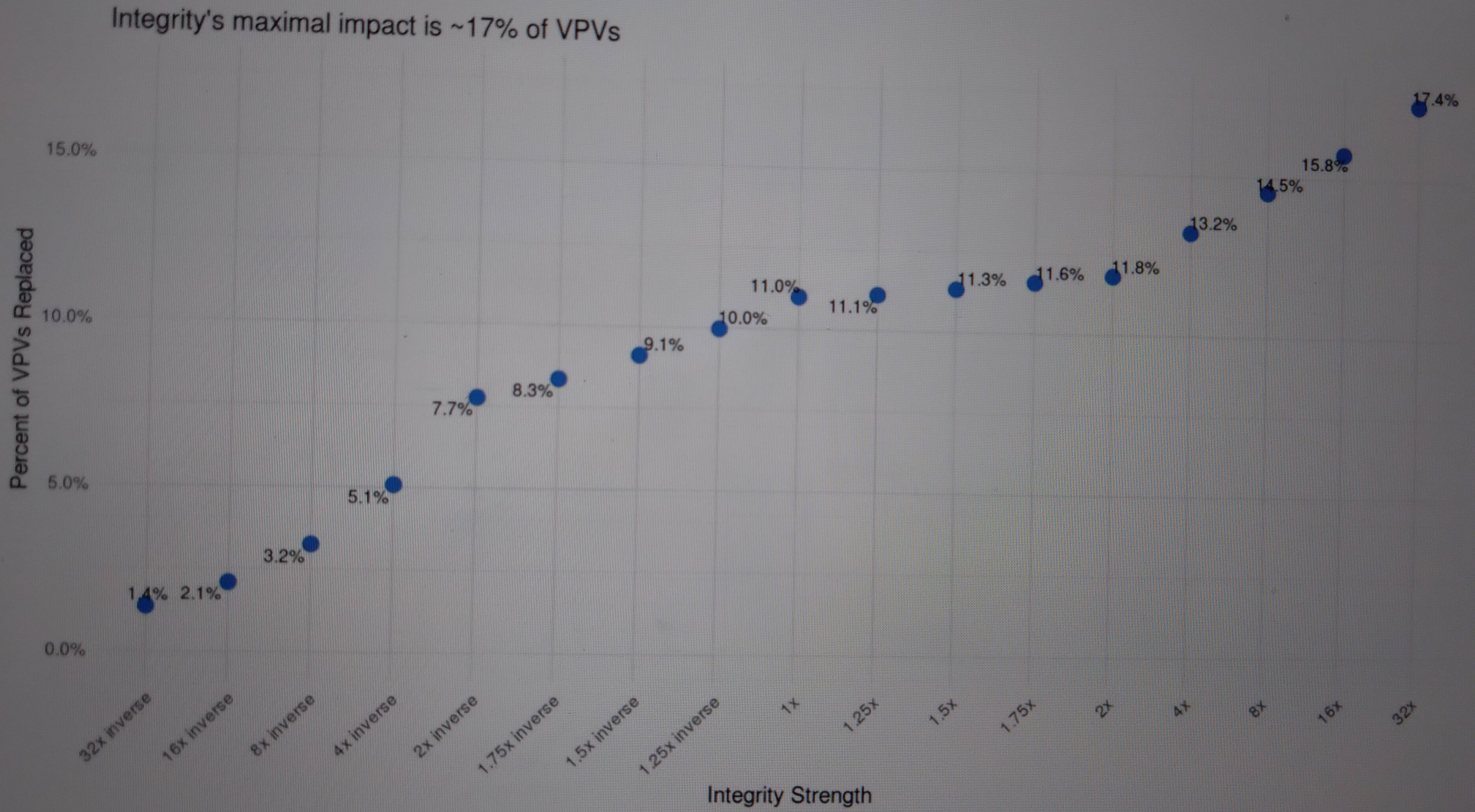
Note that the "1x" strength represents current Integrity production.

How many VPVs could Integrity impact?

Currently, Integrity demotions are responsible for ~10% of the VPVs seen, i.e. 10% of VPVs change as a result of Integrity demotions. If we increased the size of Integrity demotions substantially so that any non-zero demotion meant the content was almost blacklisted (or demoted by close to 100%), we would still only be responsible for ~17% of VPVs. This basically just says that most content does not have any meaningful Integrity demotion attached to it.



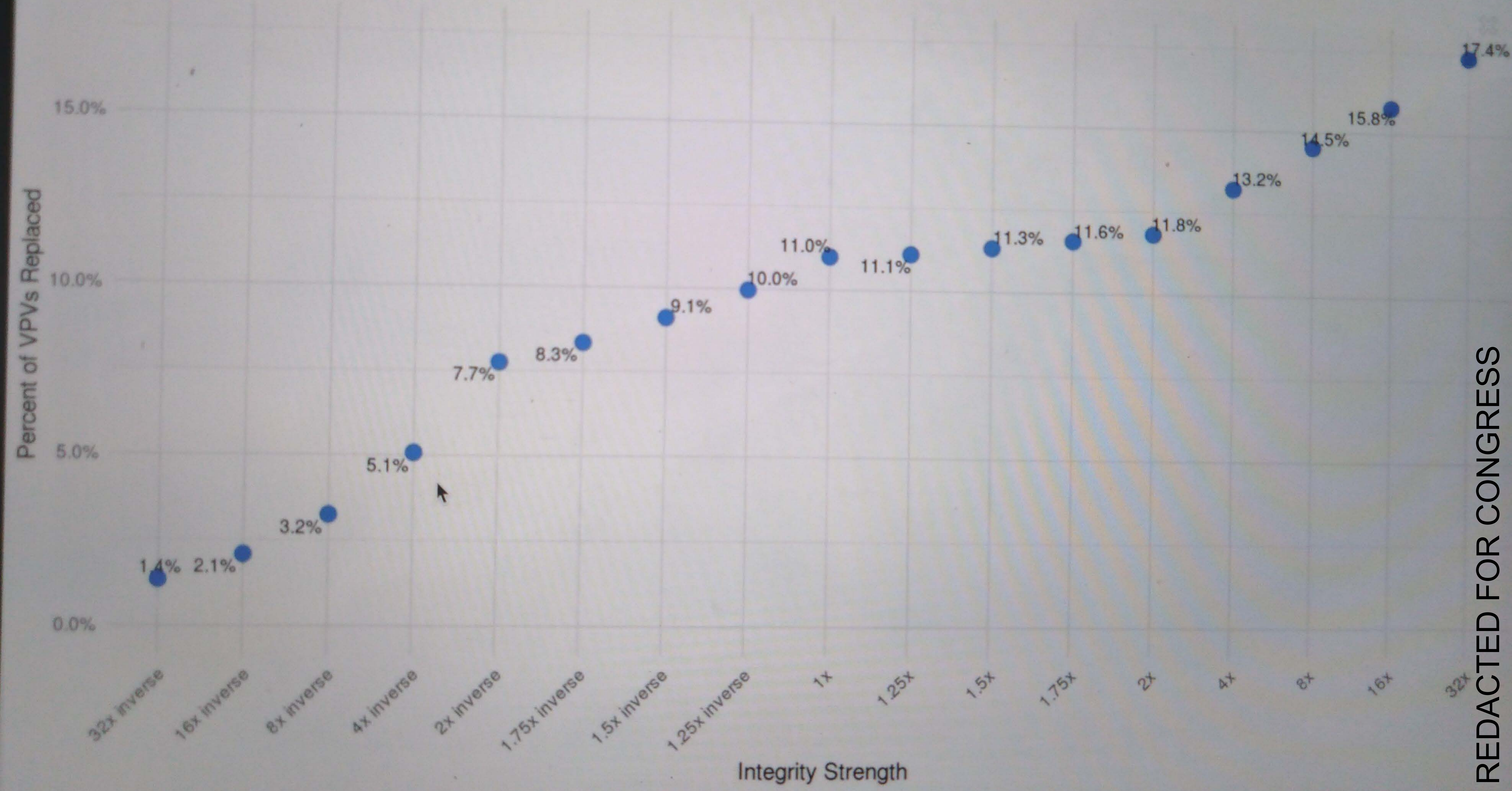
blacklisted (or demoted by close to 100%), **we would still only be responsible for ~17% of VPVs.** This basically just says that most content does not have any meaningful Integrity demotion attached to it.



How does Integrity impact engagement and WVT?

REDACTED FOR CONGRESS

Integrity's maximal impact is ~17% of VPVs




REDACTED FOR CONGRESS

How does Integrity impact engagement and WYT?

So we can see that the stronger Integrity demotions are, the more VPVs we impact, but how does that impact MSI, WYT, or other engagement metrics?

The graph below shows a few interesting trends:

- For WYT, the stronger the Integrity demotion, the greater the WYT!
- For MSI and Likes, as Integrity demotions increase up to the current status quo, these metrics actually go up and then decline. To me, this suggests that our ranking config is heavily tuned to optimize for MSI so any change causes a drop.
- For Shares, Comments, and Significant Comments, removing Integrity doesn't really change things much but if we were to increase Integrity, all of these metrics would be hit quite hard. 

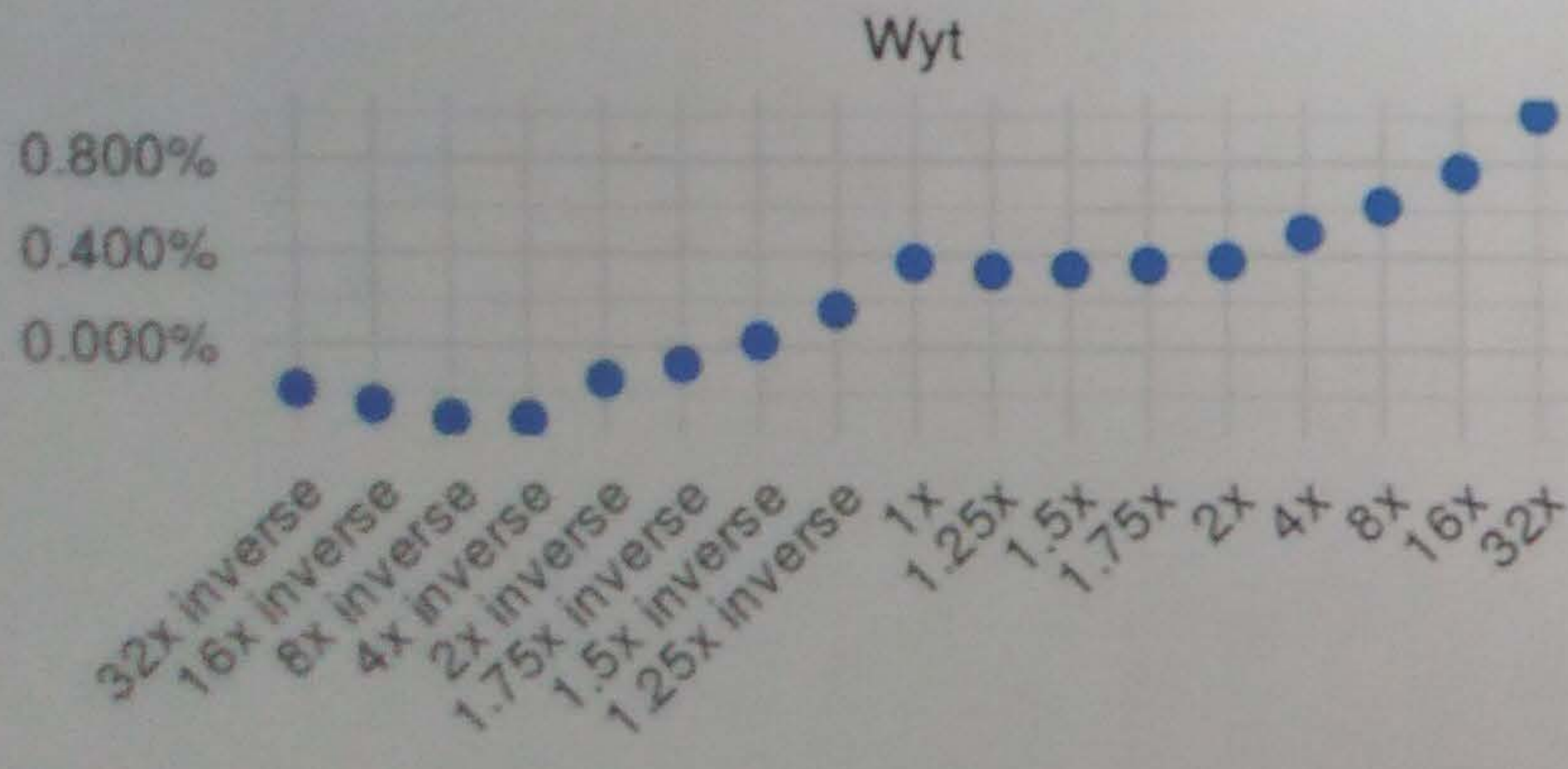
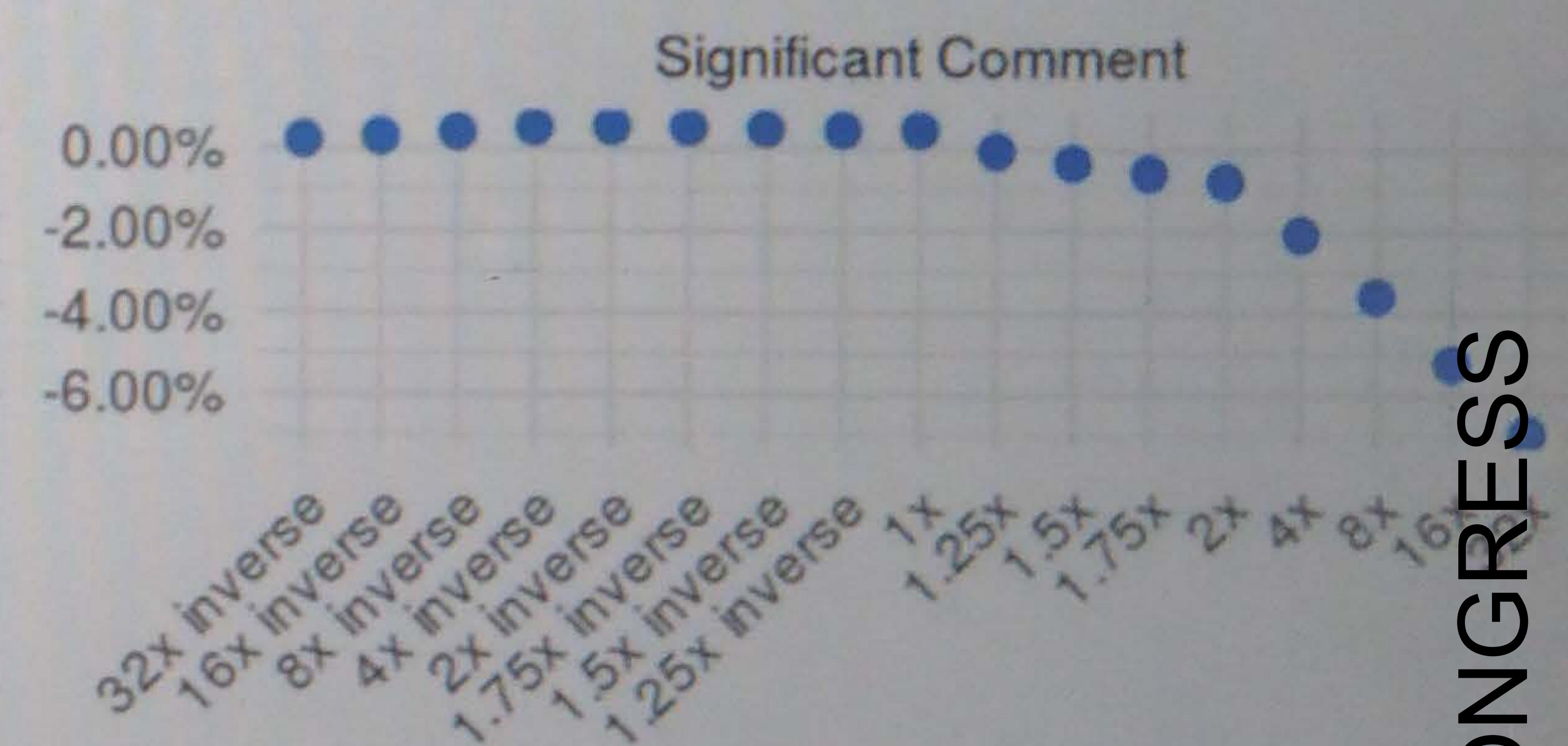
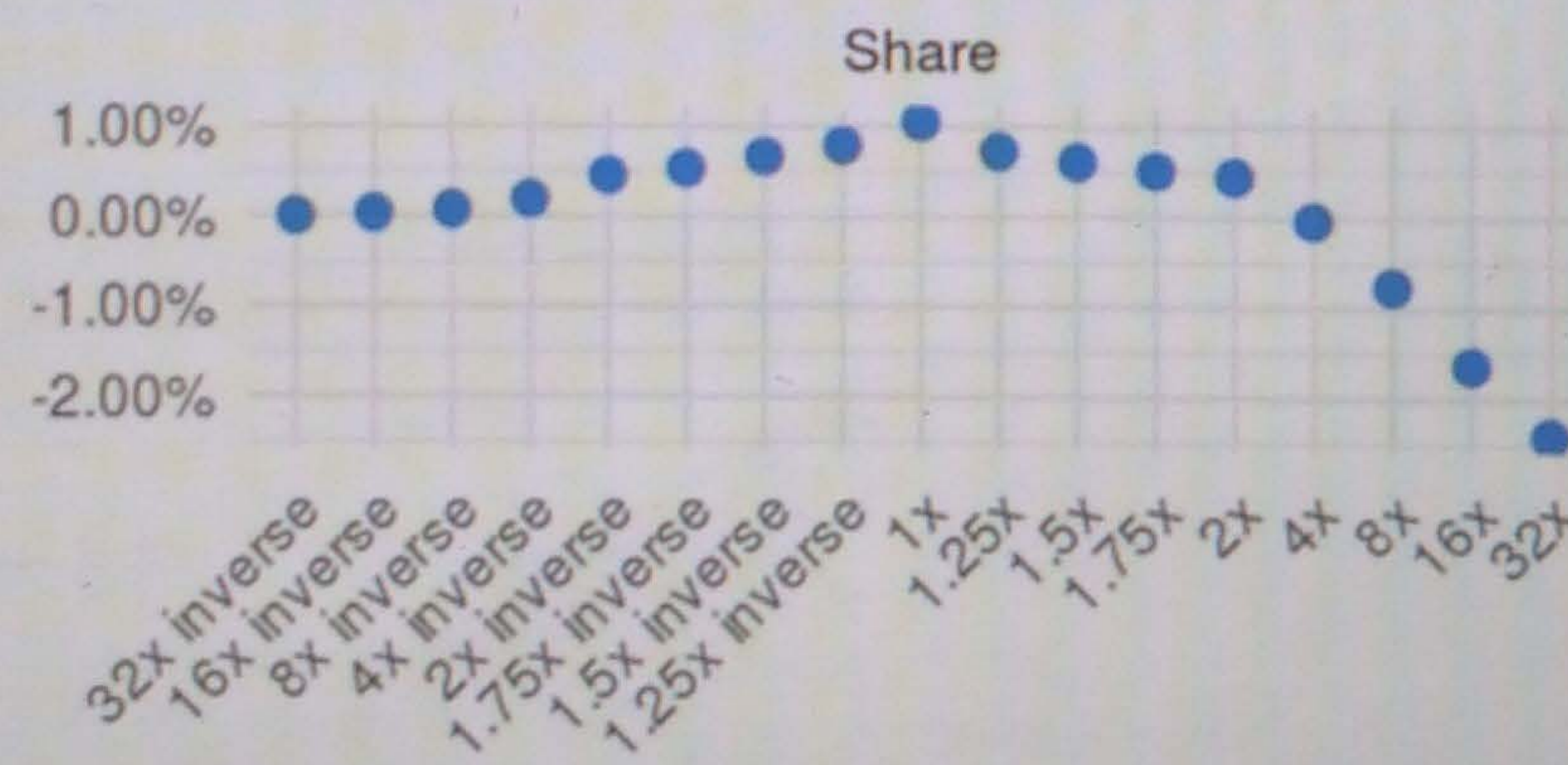
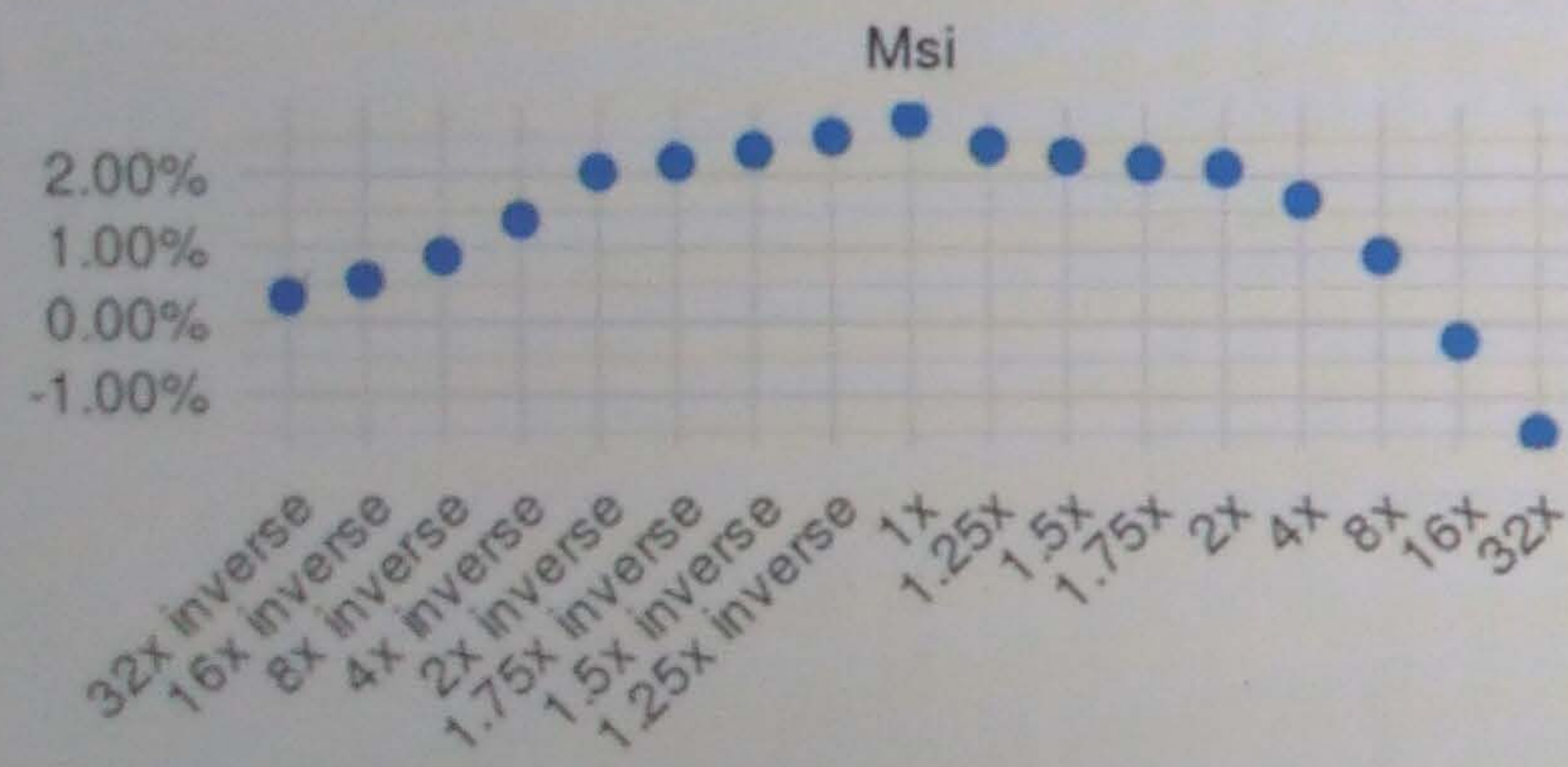
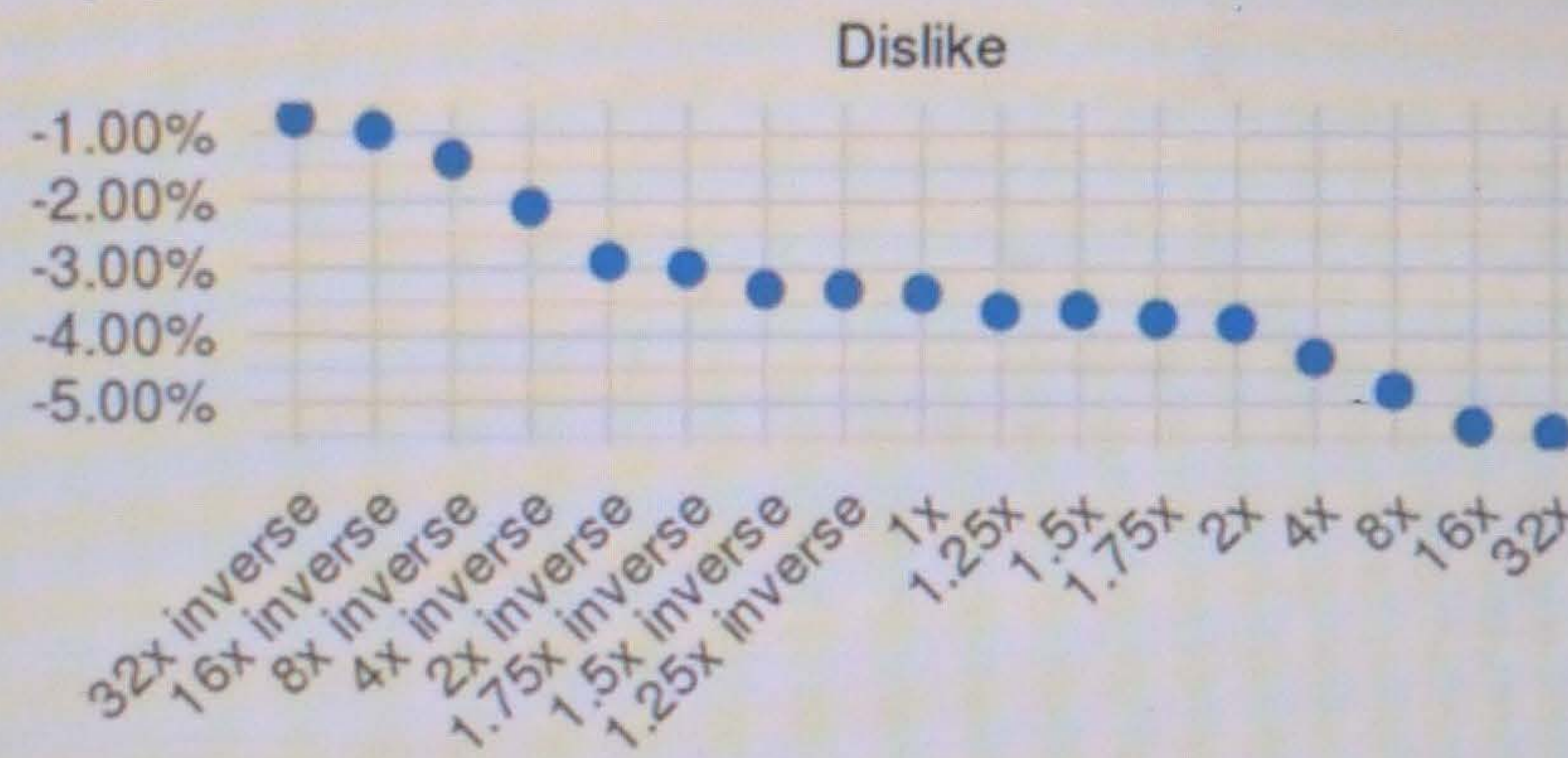
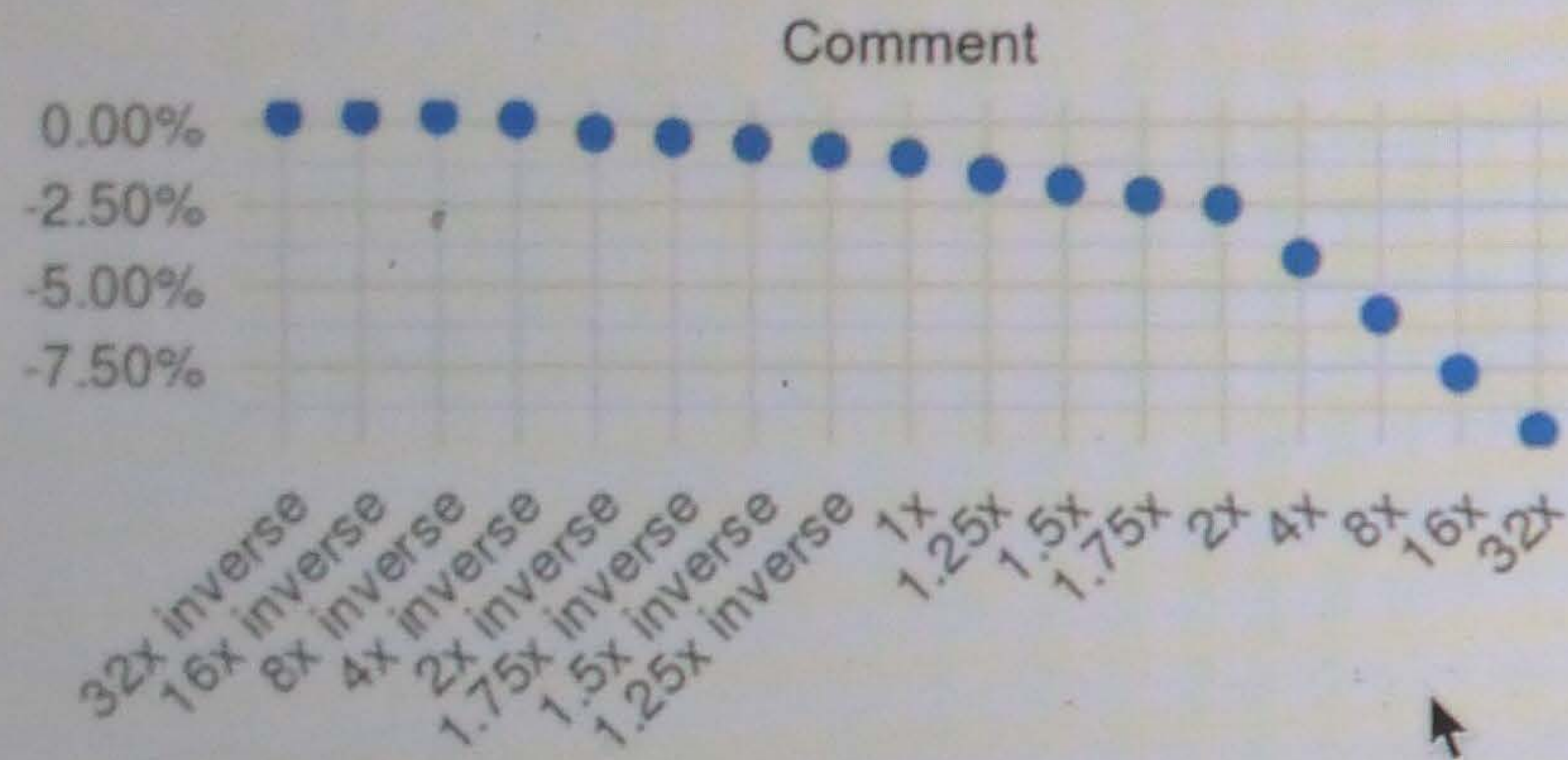
How do engagement metrics change as Integrity strength changes?

Comment

Dislike

Like

How do engagement metrics change as Integrity strength changes?



Integrity Strength

REDACTED FOR CONGRESS

How do engagement metrics change as Integrity streng

Comment

0.00%
-2.50%
-5.00%
-7.50%

-1.00%
-2.00%
-3.00%
-4.00%
-5.00%

32x inverse
16x inverse
8x inverse
4x inverse
2x inverse
1.75x inverse
1.5x inverse
1.25x inverse
1x
1.25x
1.5x
1.75x
2x
4x
8x
16x
32x

32x inverse
16x inverse
8x inverse
4x inverse
2x inverse
1.75x inverse
1.5x inverse
1.25x inverse



Msi

Percent Change

2.00%
1.00%
0.00%
-1.00%

1.00%
0.00%
-1.00%
-2.00%

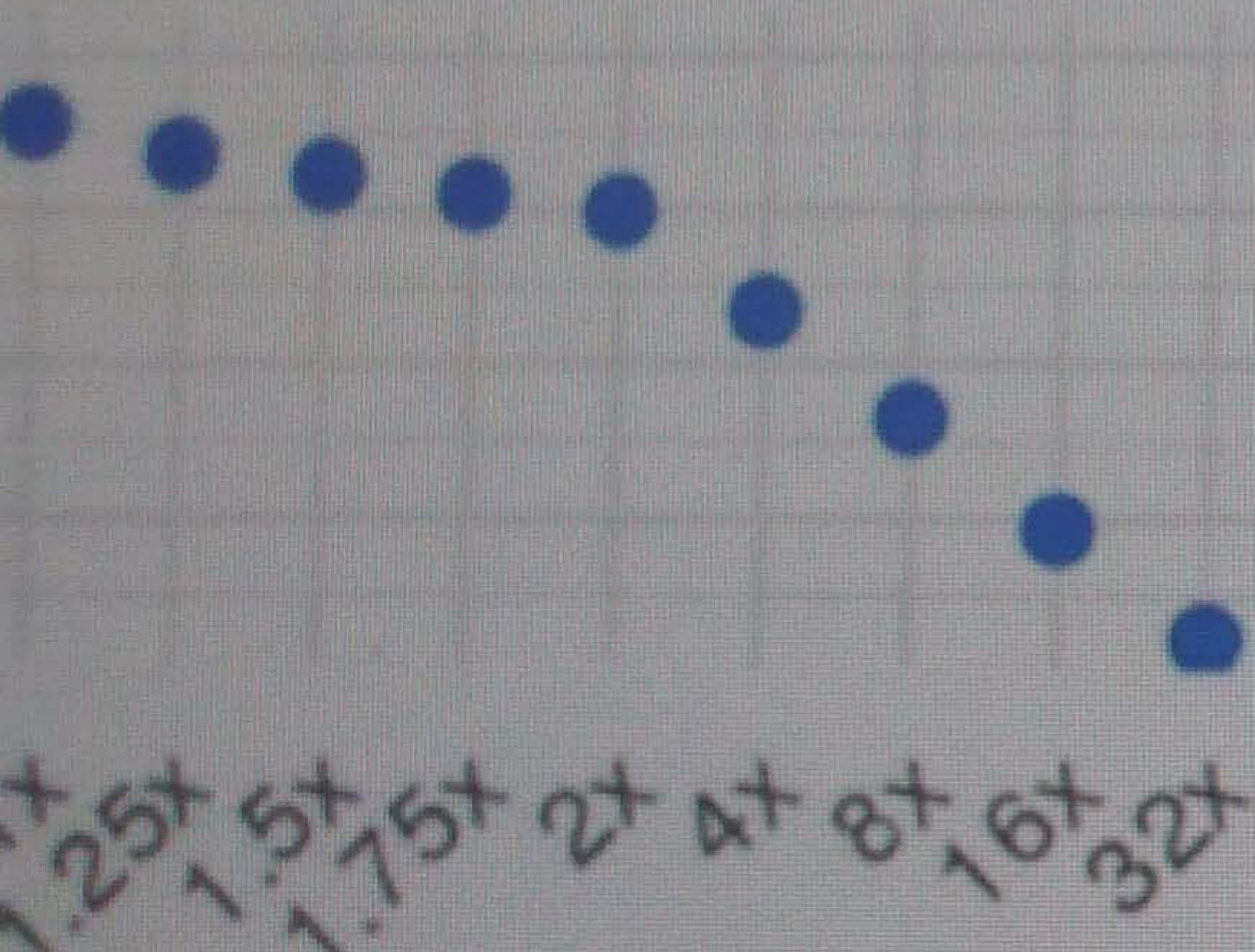
32x inverse
16x inverse
8x inverse
4x inverse
2x inverse
1.75x inverse
1.5x inverse
1.25x inverse
1x
1.25x
1.5x
1.75x
2x
4x
8x
16x
32x

32x inverse
16x inverse
8x inverse
4x inverse
2x inverse
1.75x inverse
1.5x inverse
1.25x inverse

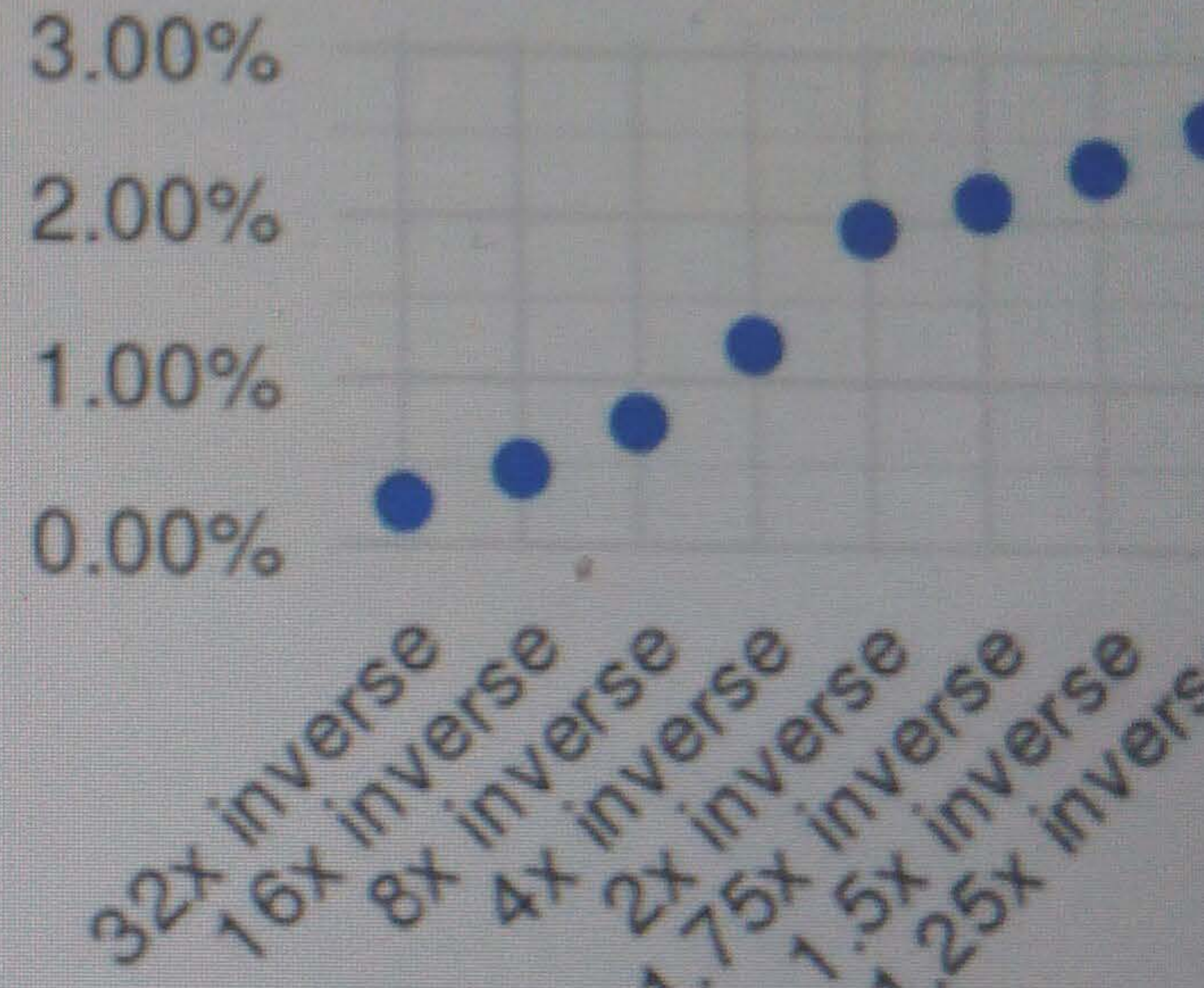
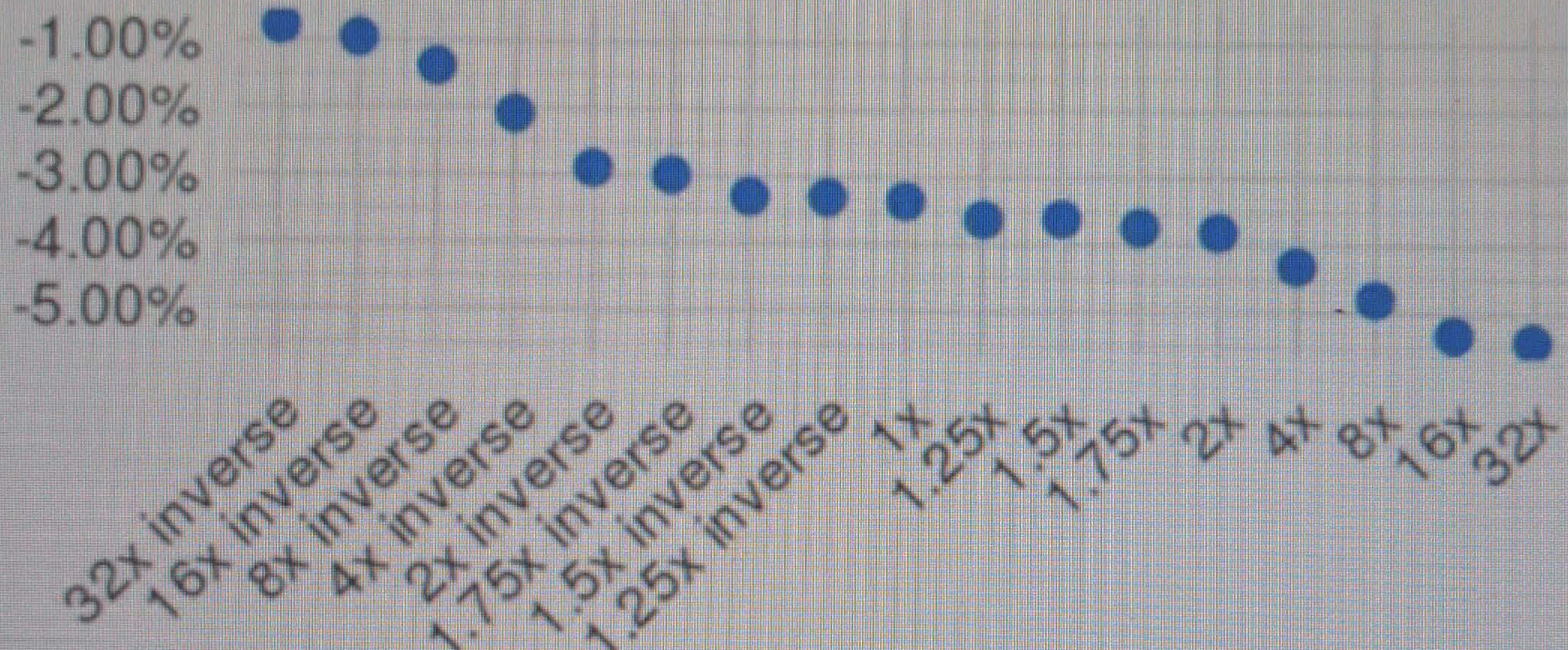
REDACTED FOR CONGRESS

ment metrics change as Integrity strength changes?

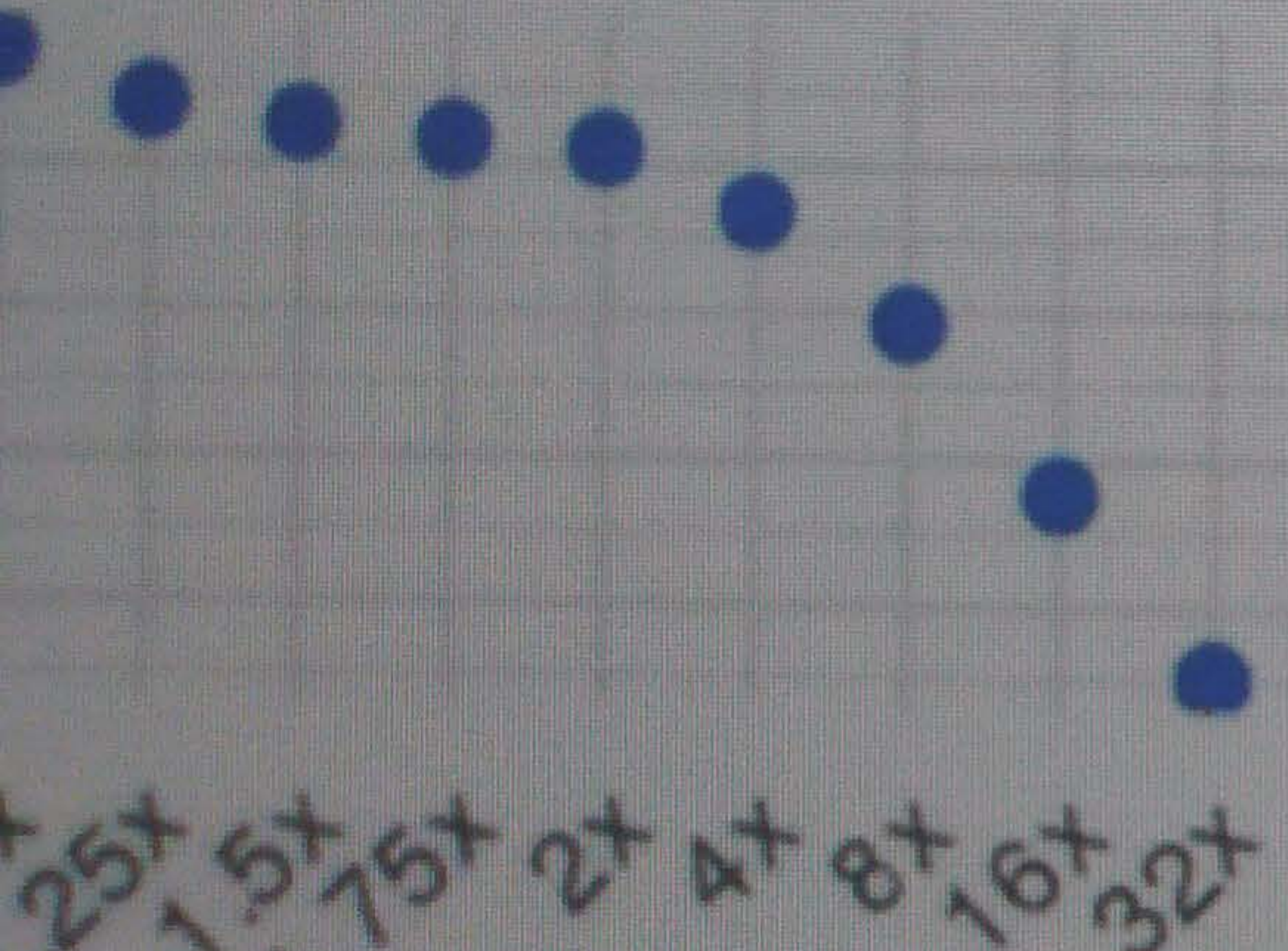
ment



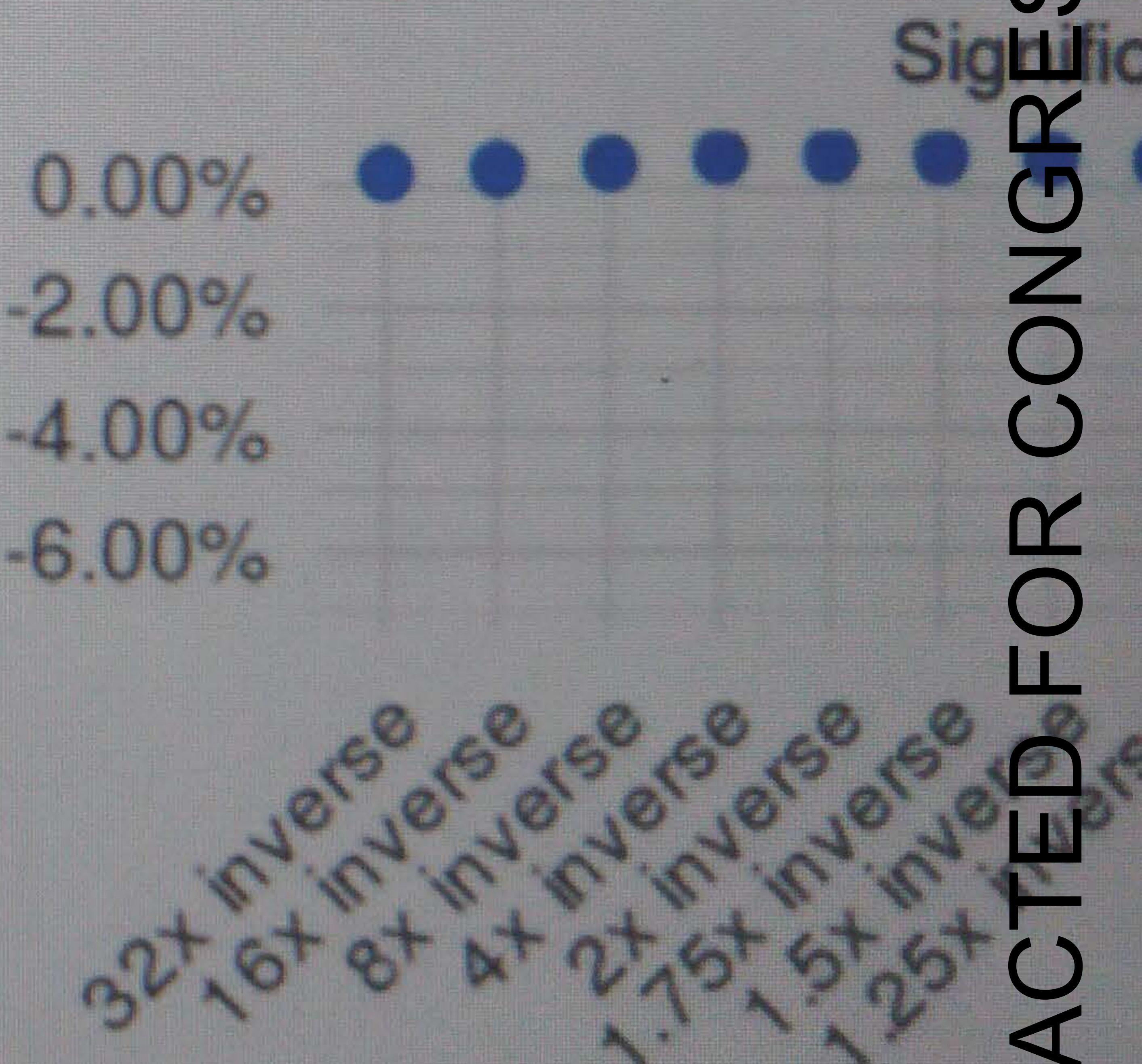
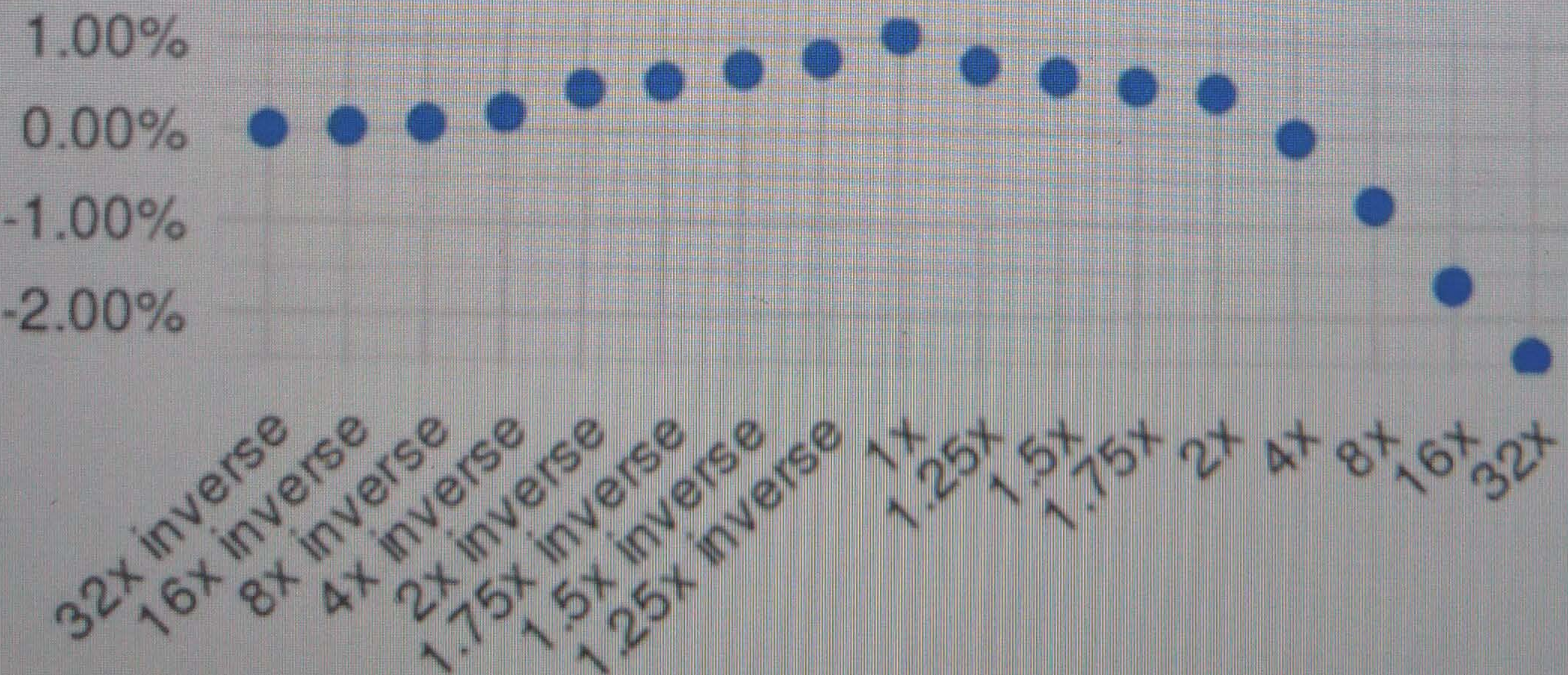
Dislike



si



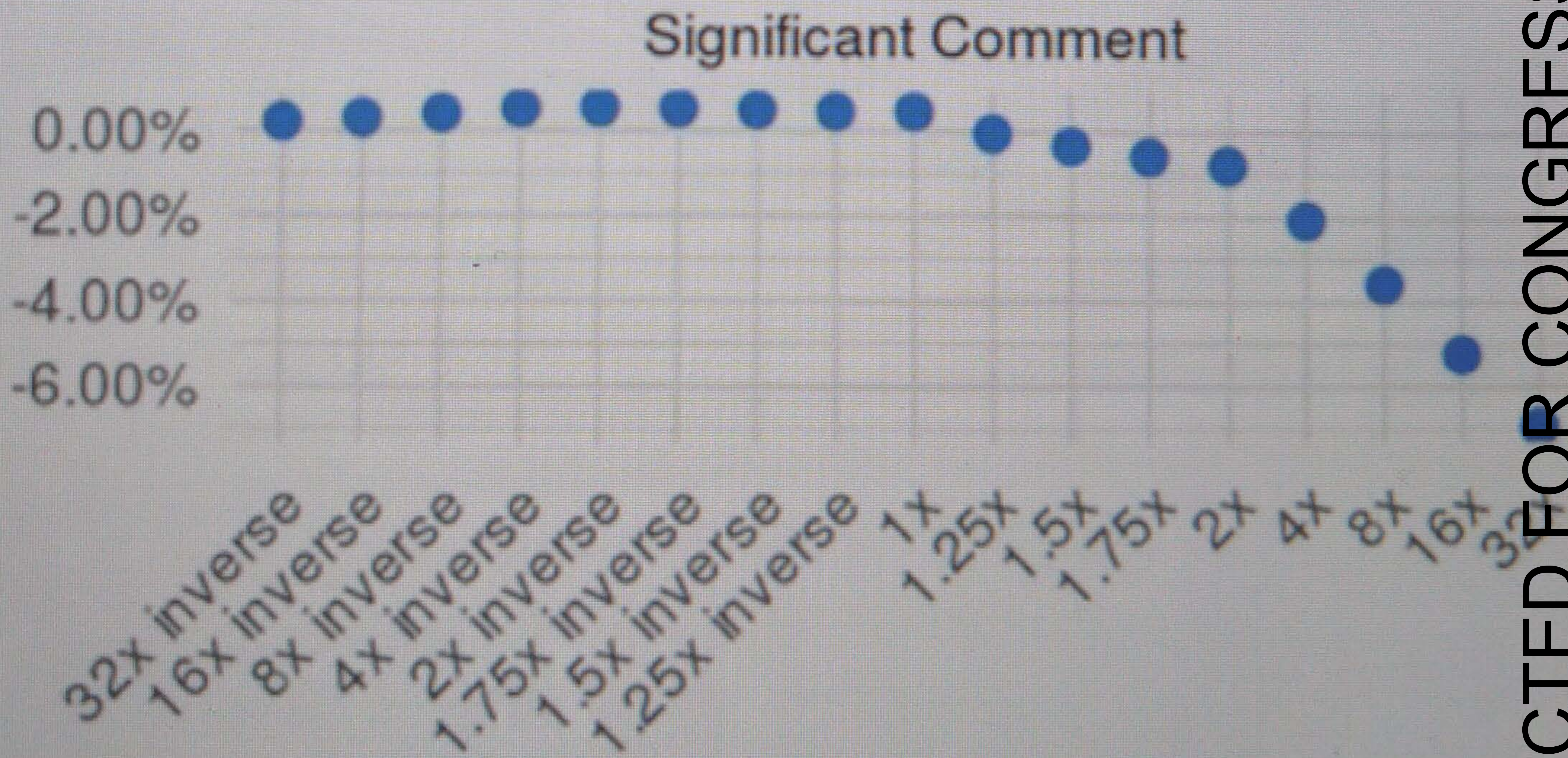
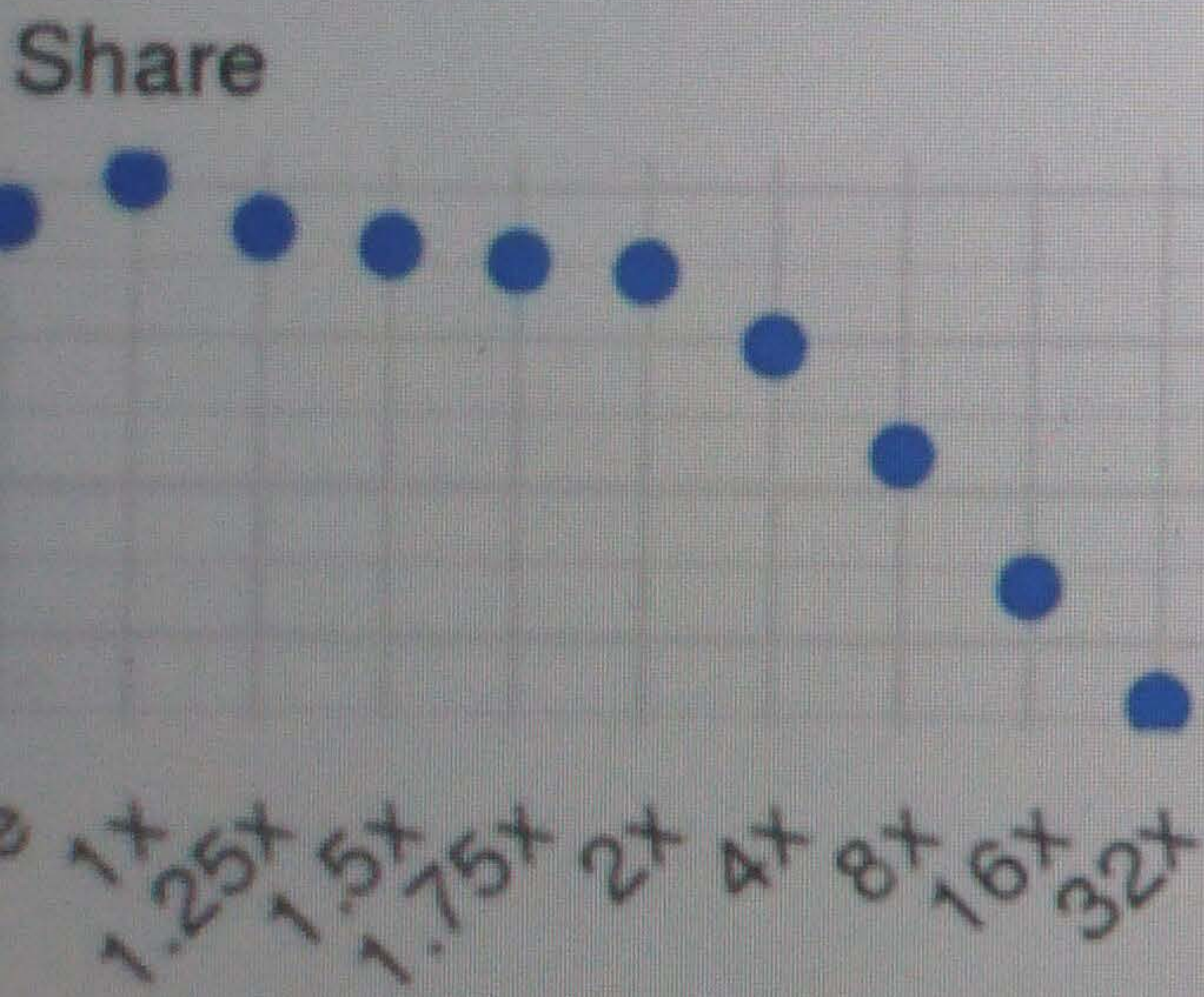
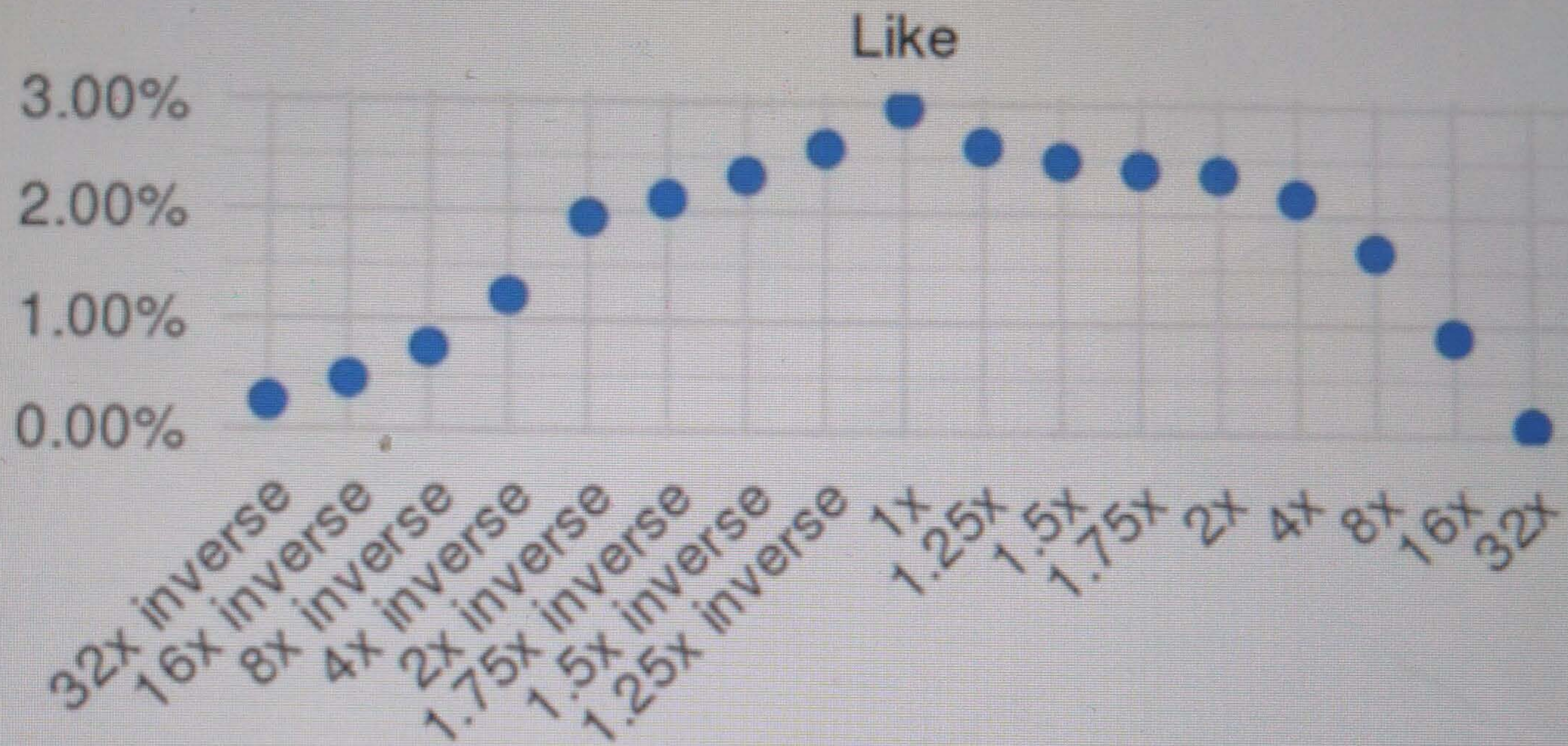
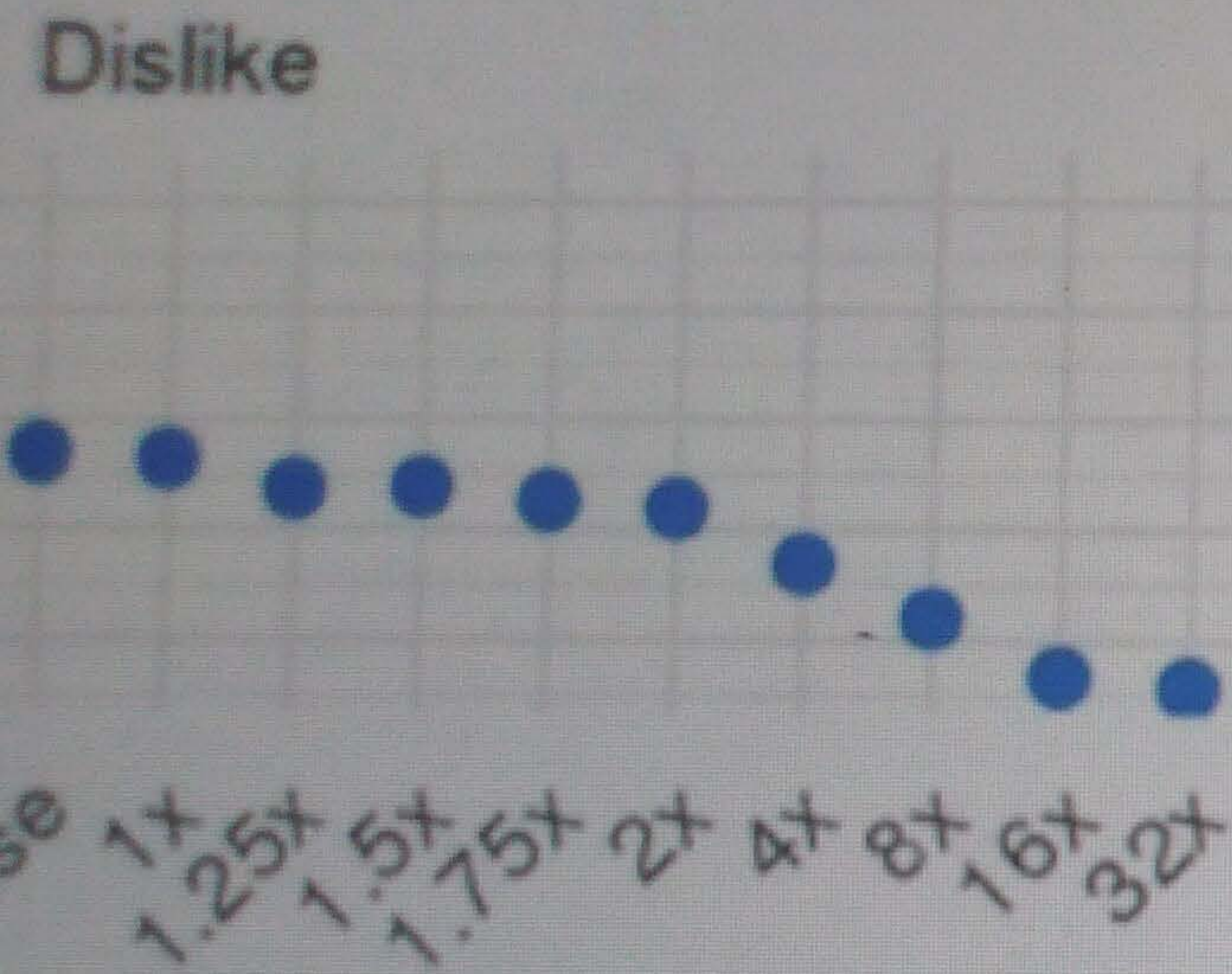
Share



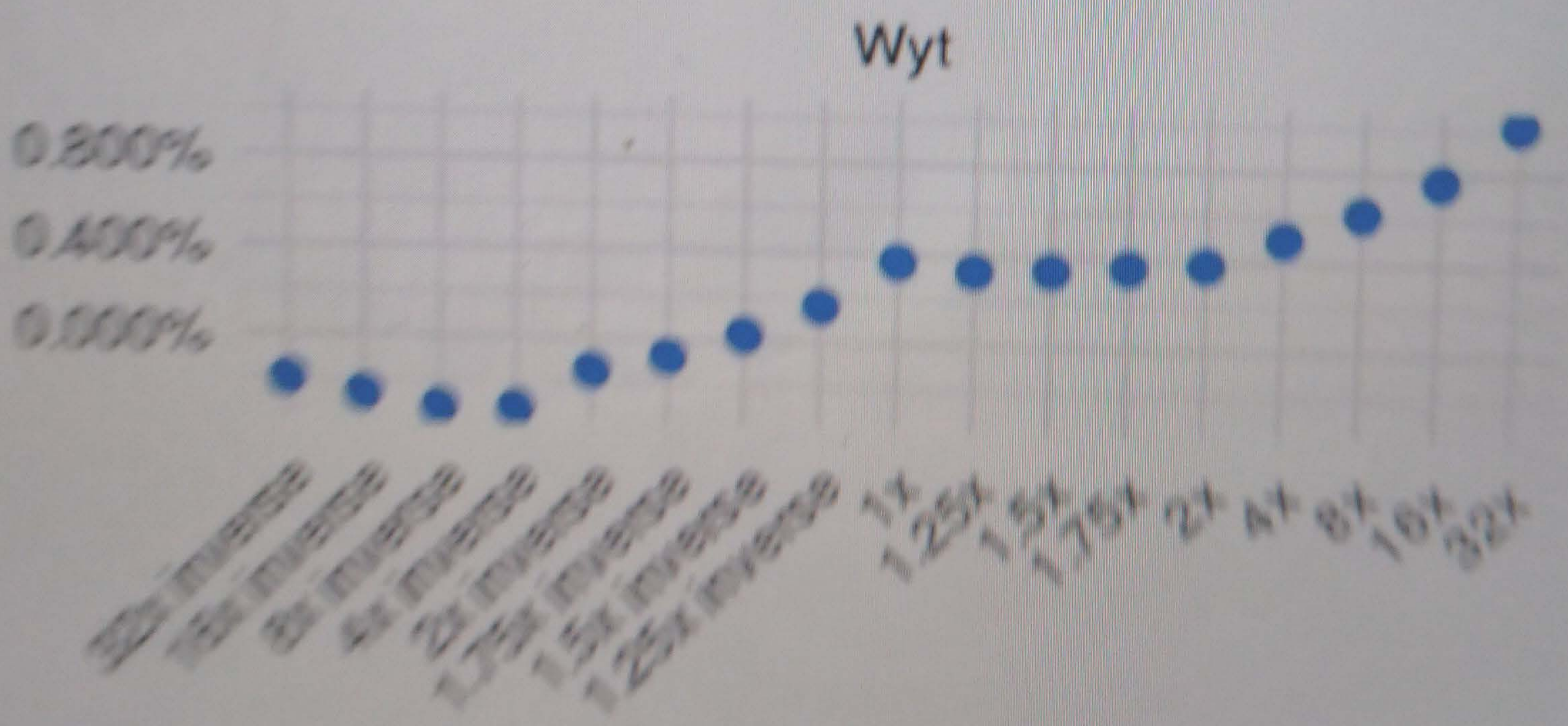
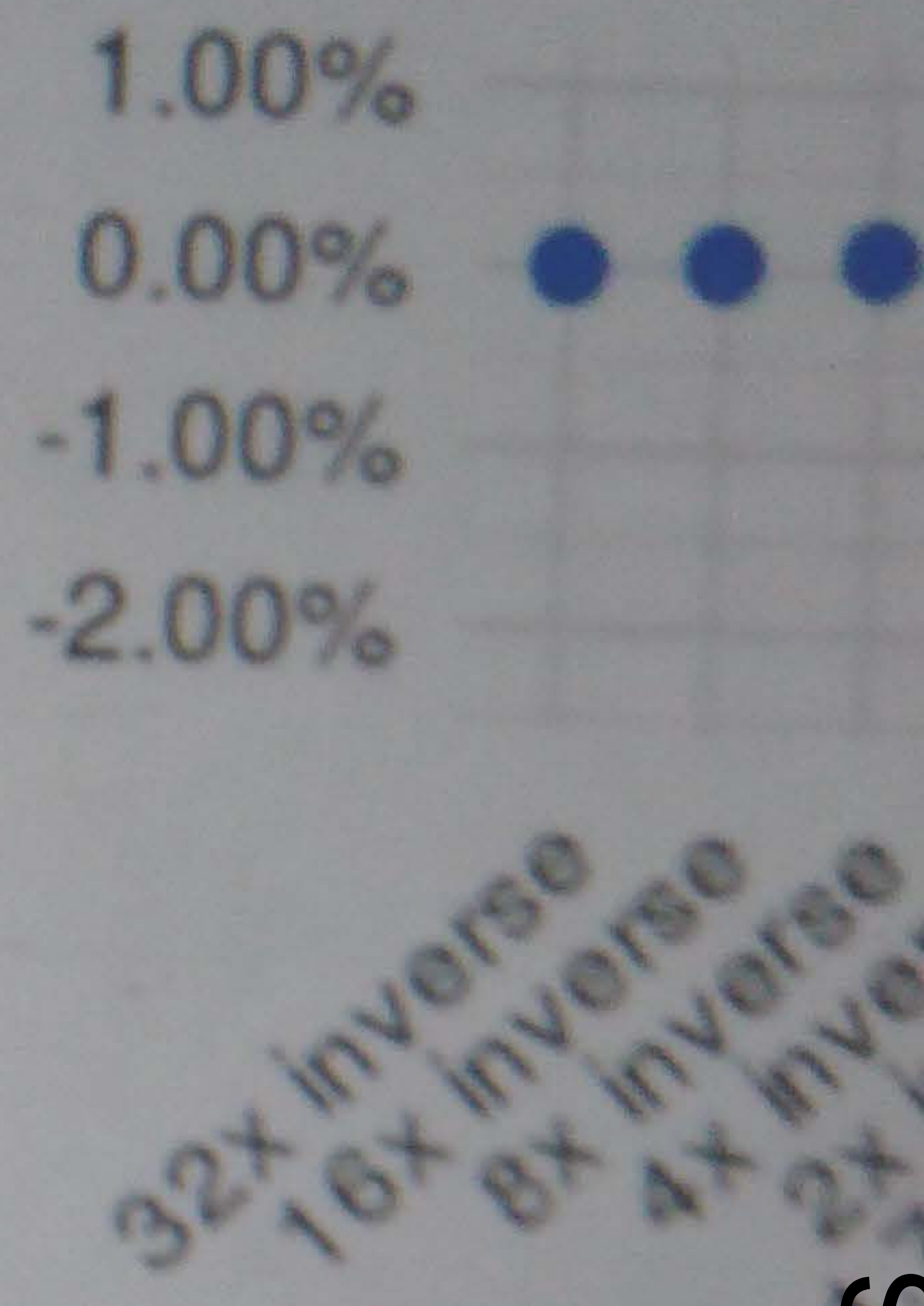
Signific

REDACTED FOR CONGRESS

What changes?



Percent Change



REDACTED FOR CONGRESS

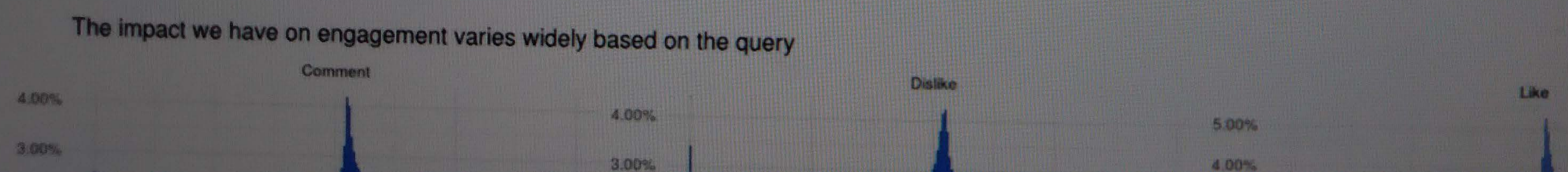
32x
16x
8x
4x
2x
1.75x
1.5x
1.25x

Distributions of impact

I want to focus in on the current production config (i.e. '1x' above) just to emphasize that while the mean impact of Integrity is generally small, there are queries where Integrity demotions radically changes the predicted MSI, WYT, Comments, etc. and thus has a big impact on the user's Feed. This is expected of any change, i.e. there will be a distribution of impact, but it has been suggested that perhaps we should limit negative impact from Integrity and it's worth getting a sense of what these distributions look like.

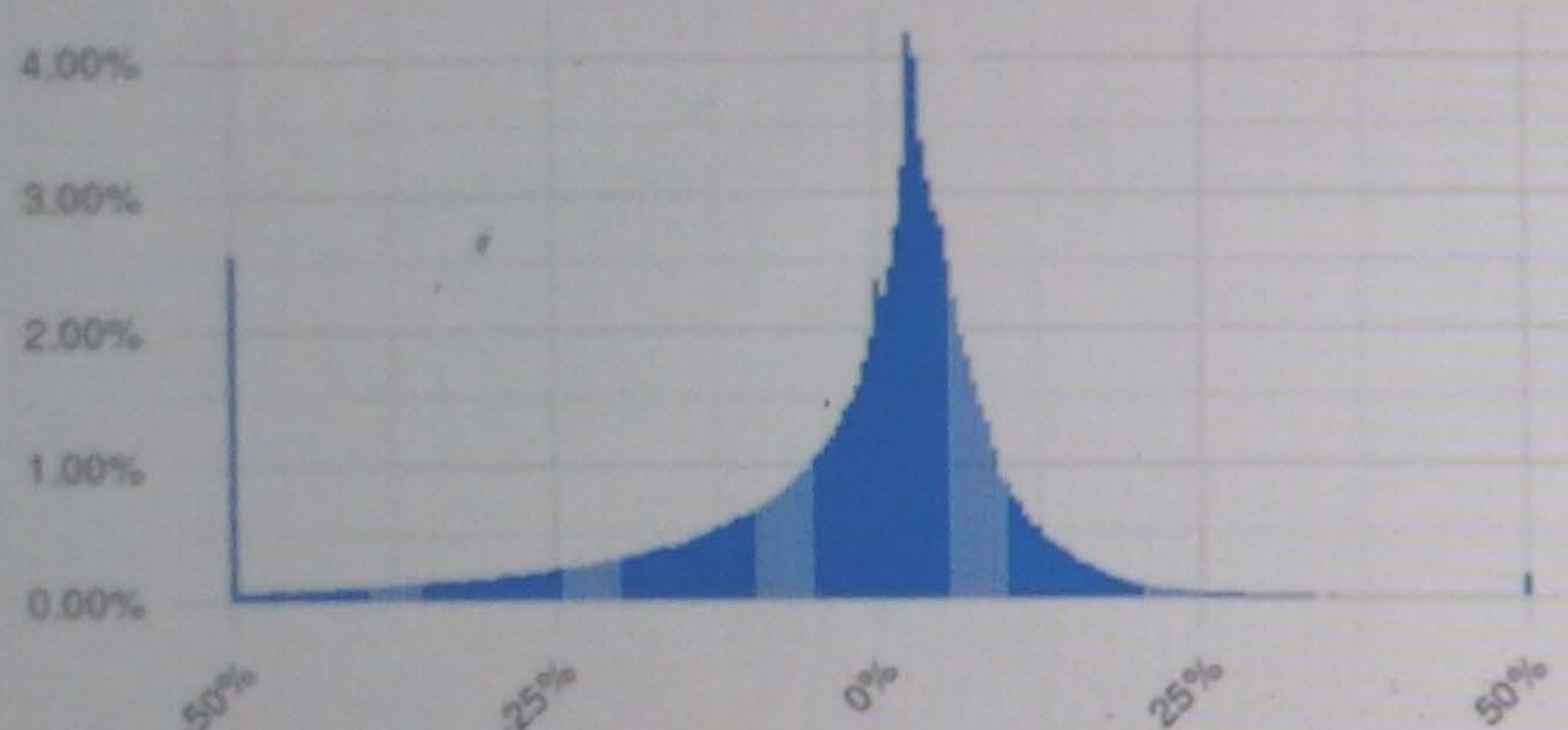
There's a fairly wide distribution of impact for most engagement metrics, much wider than for WYT. Comments and shares, as expected, are often negatively impacted by Integrity, while Likes and MSI tend to be more symmetric.

REDACTED FOR CONGRESS

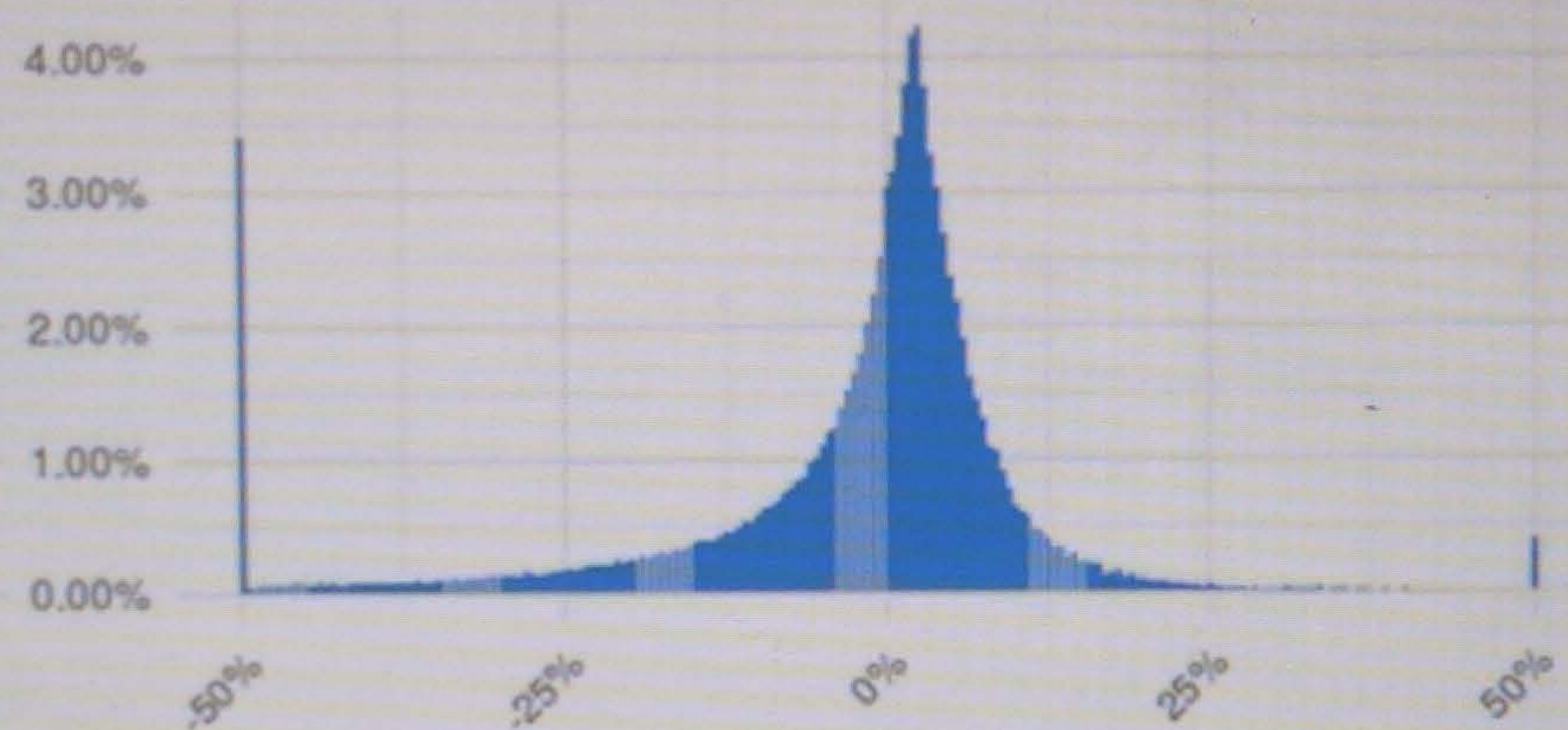


The impact we have on engagement varies widely based on the query

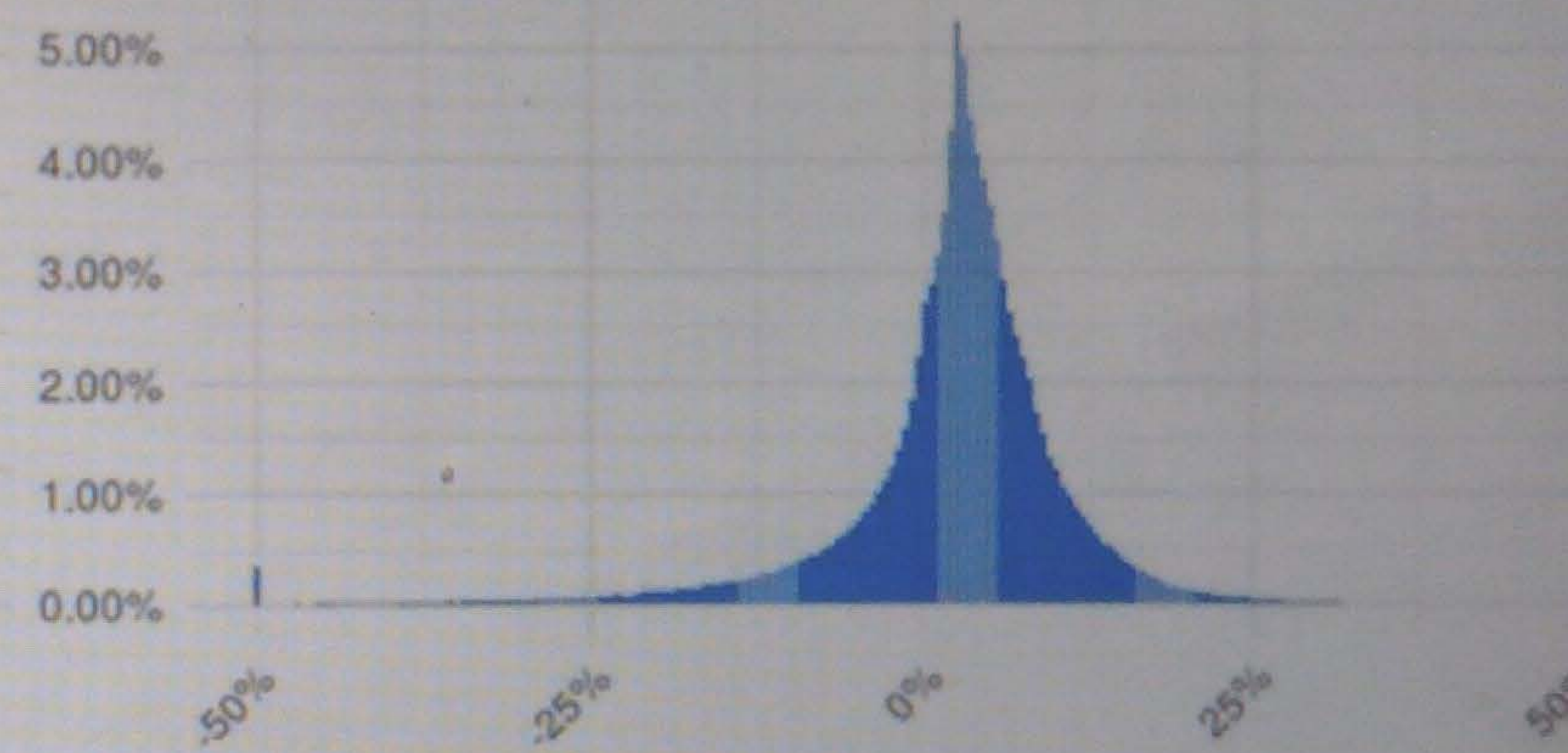
Comment



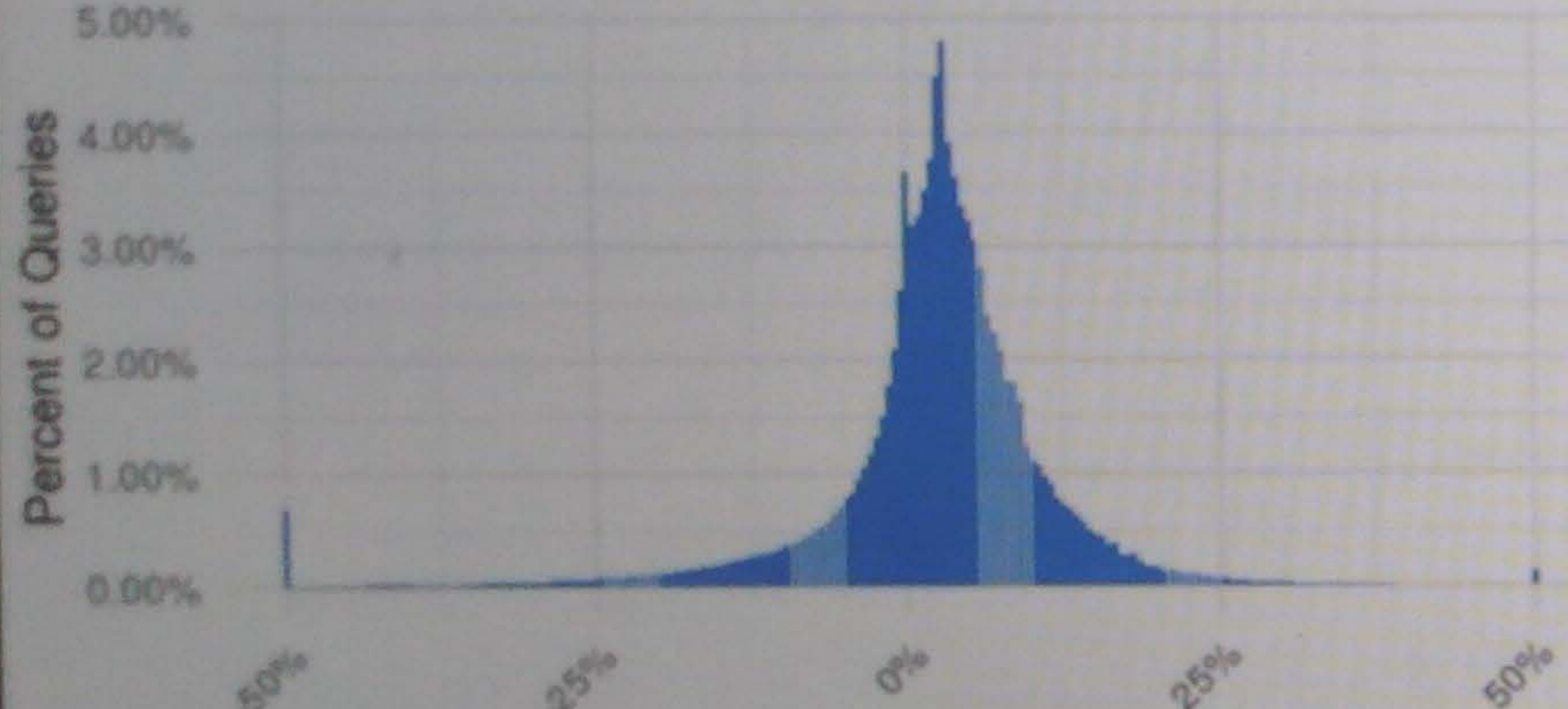
Dislike



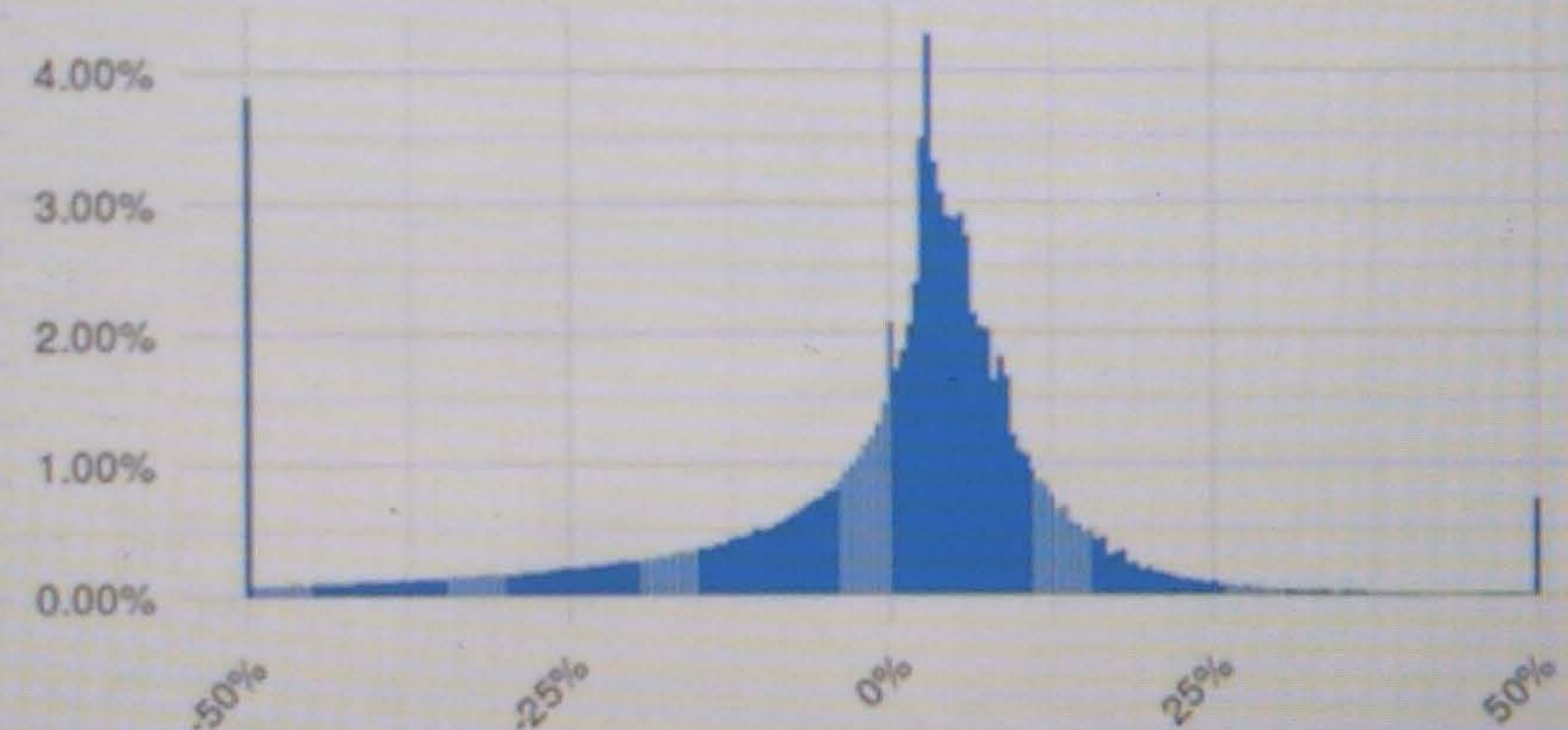
Like



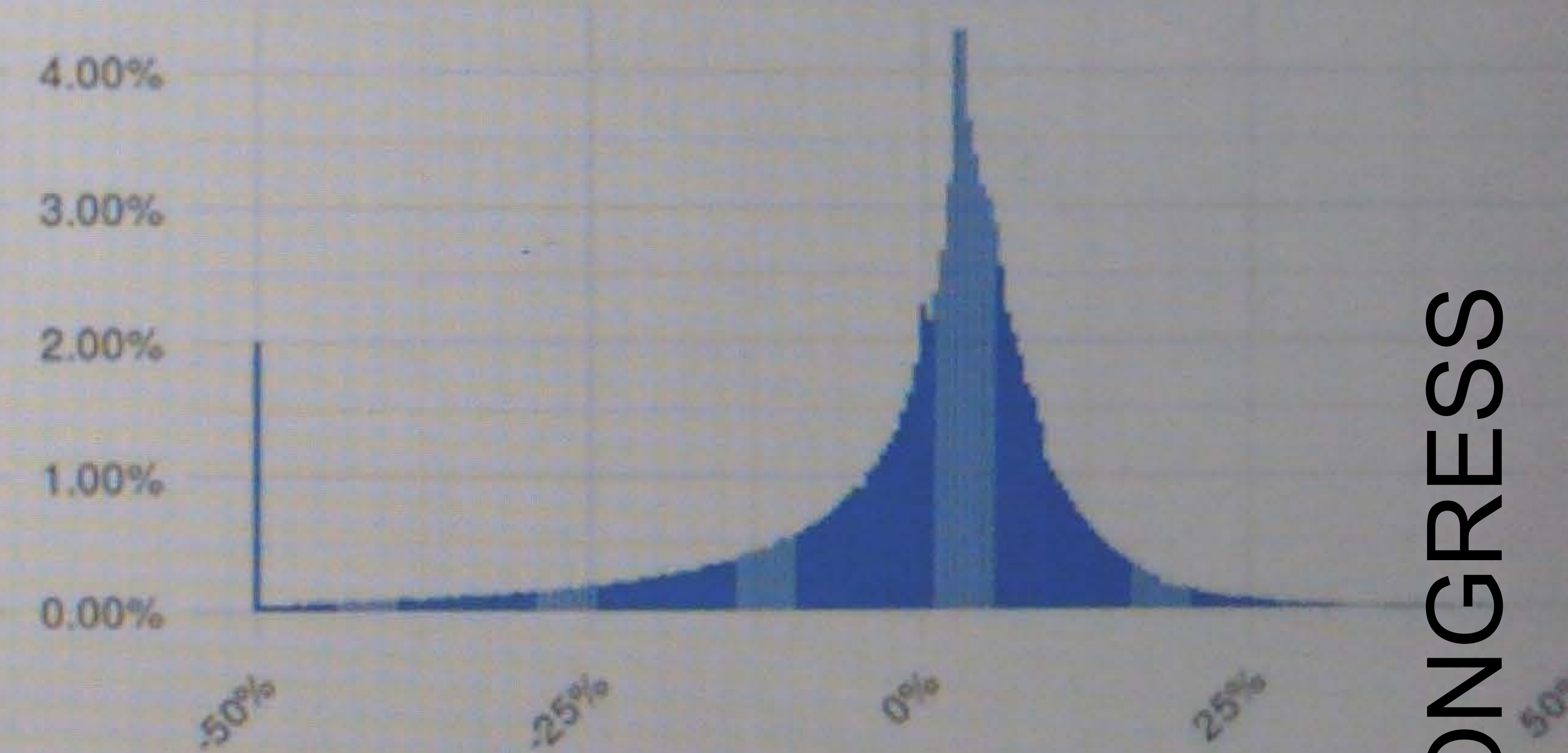
Msi



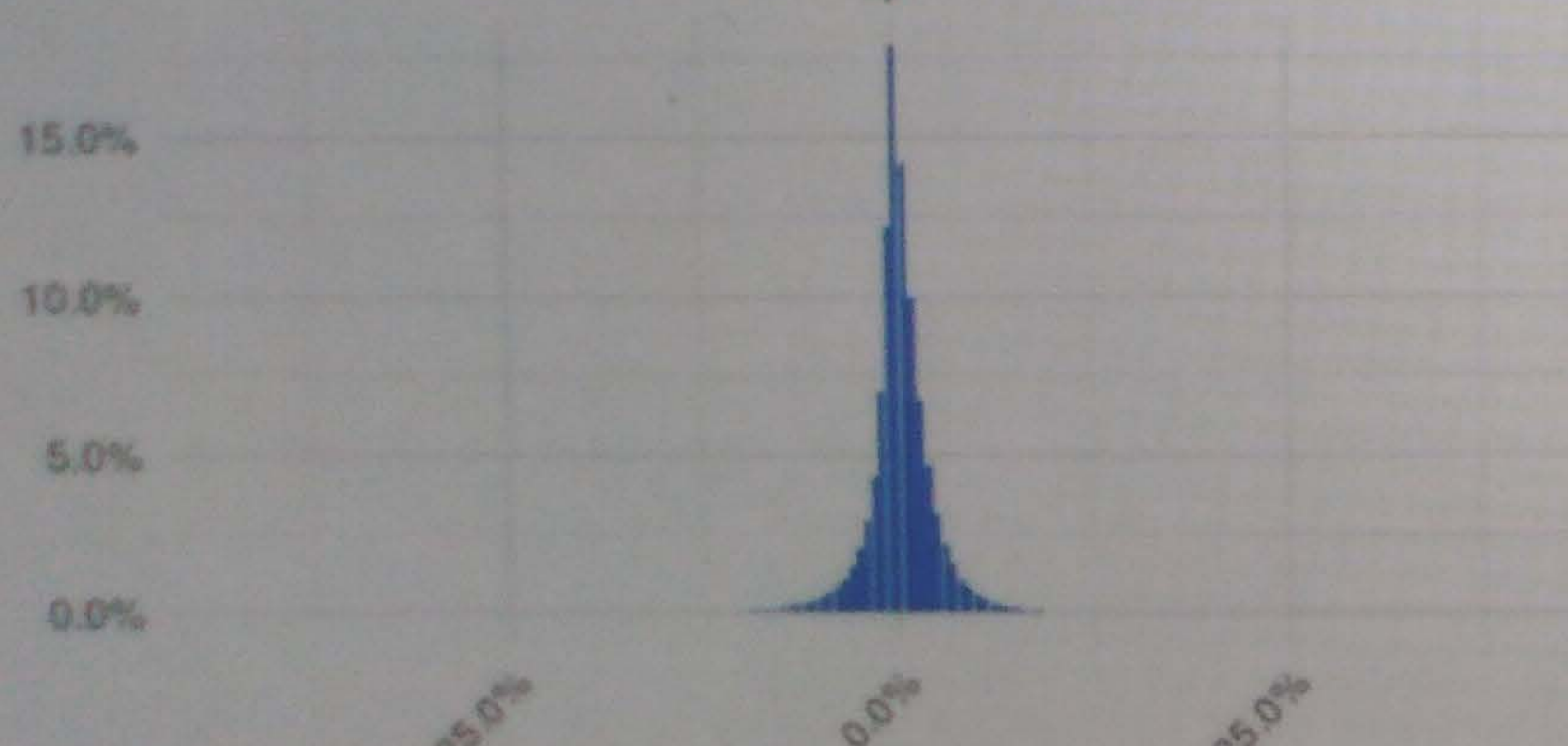
Share



Significant Comment



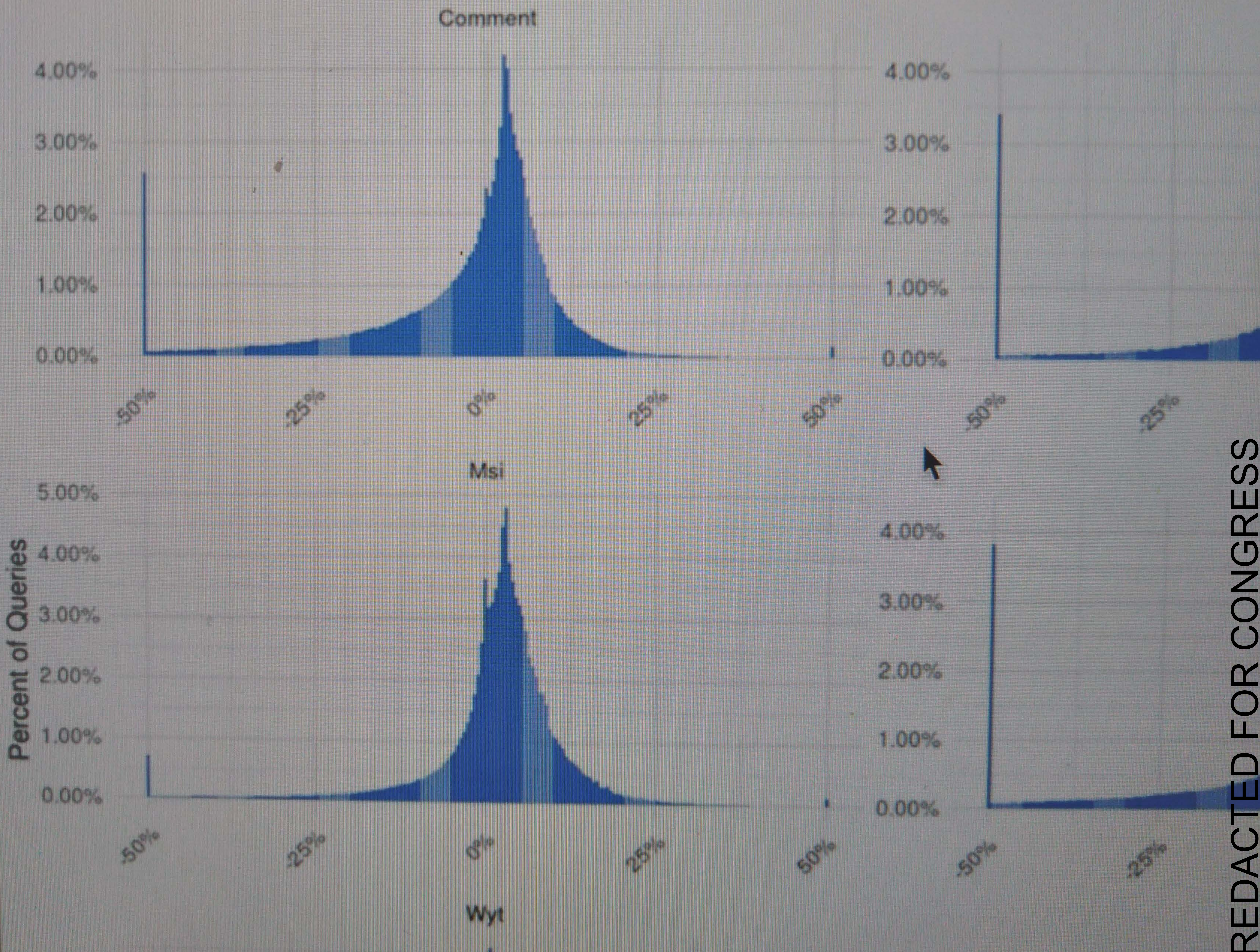
Wyt



Percent Change

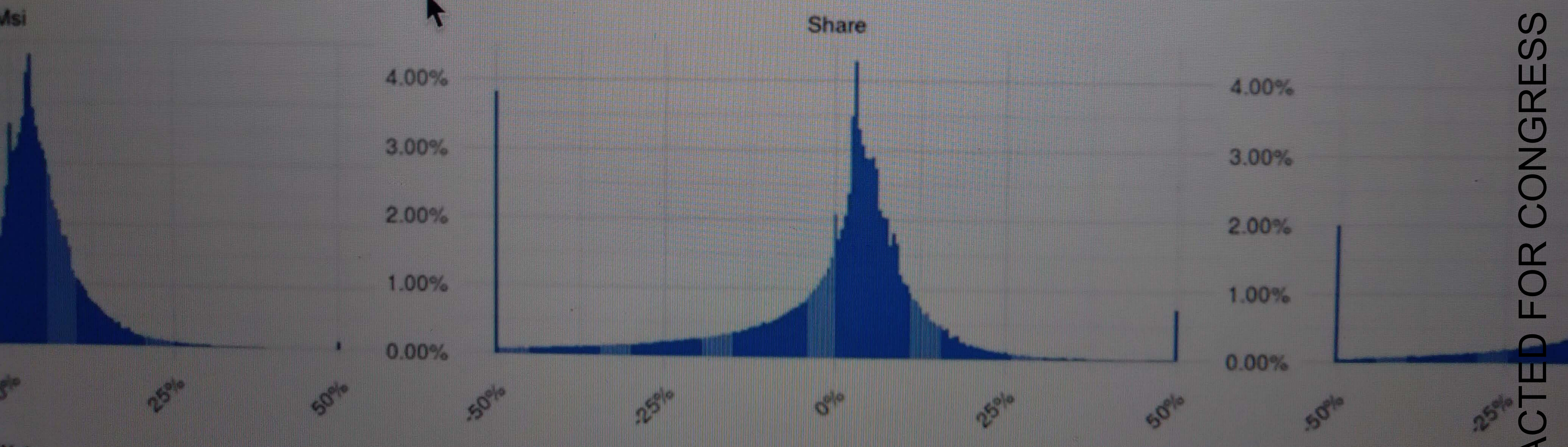
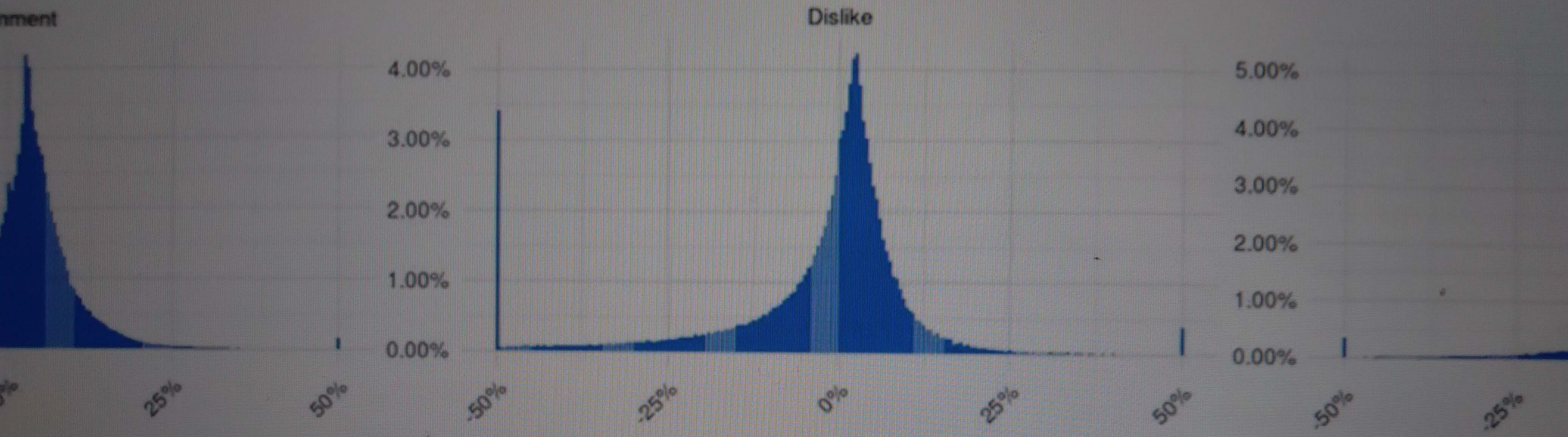
REDACTED FOR CONGRESS

The impact we have on engagement varies widely based on the query



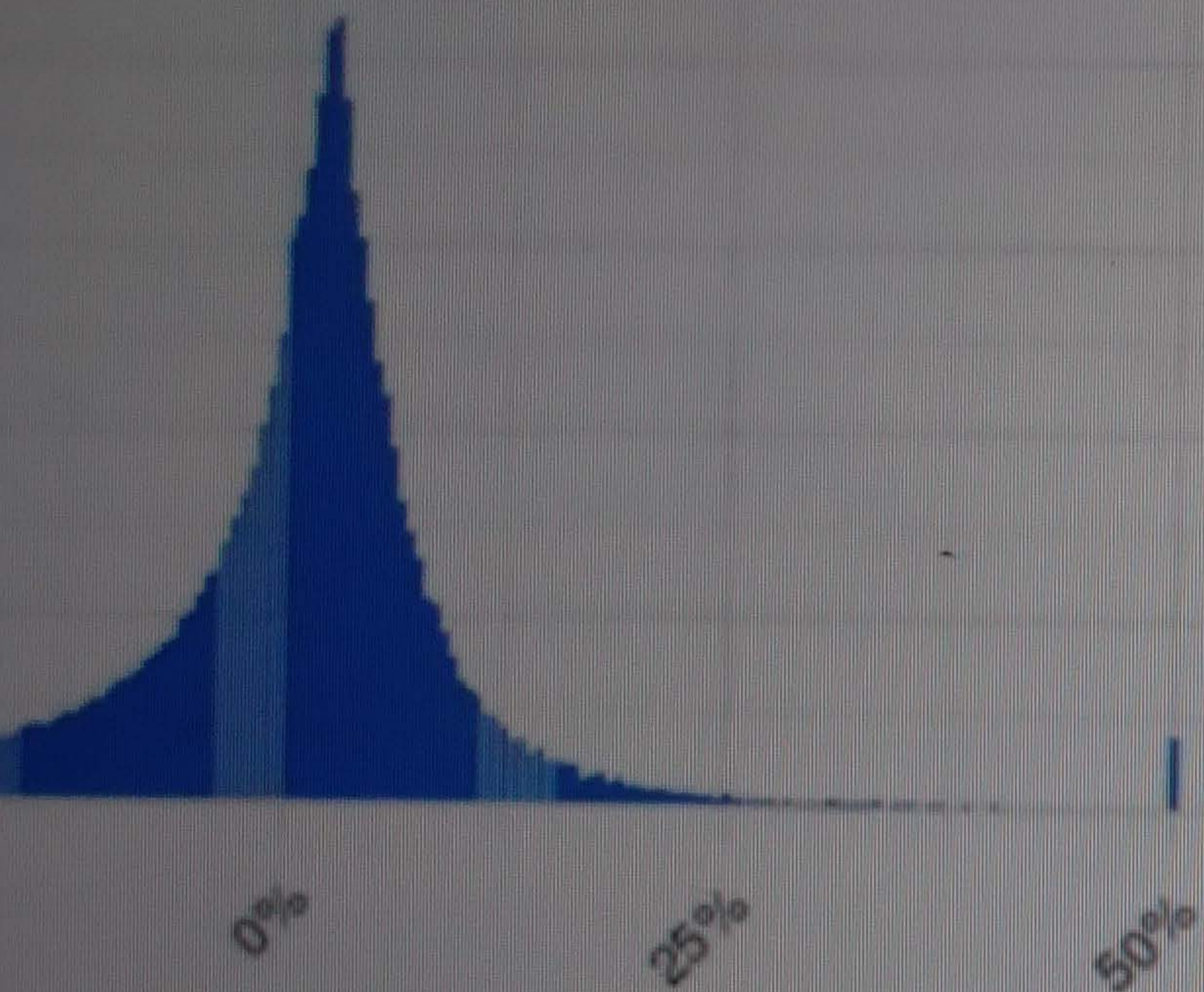
REDACTED FOR CONGRESS

Engagement varies widely based on the query



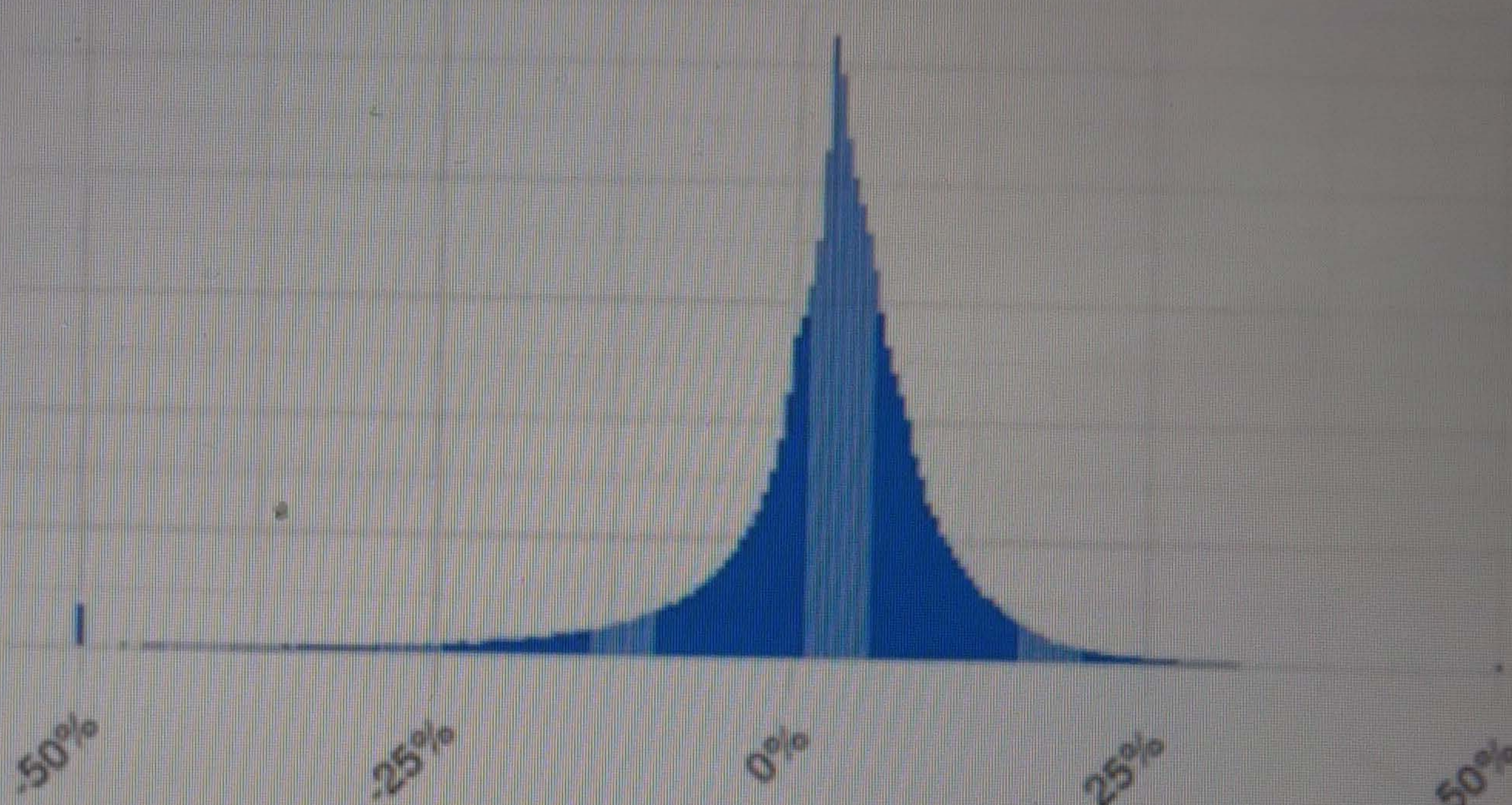
REDACTED FOR CONGRESS

Dislike

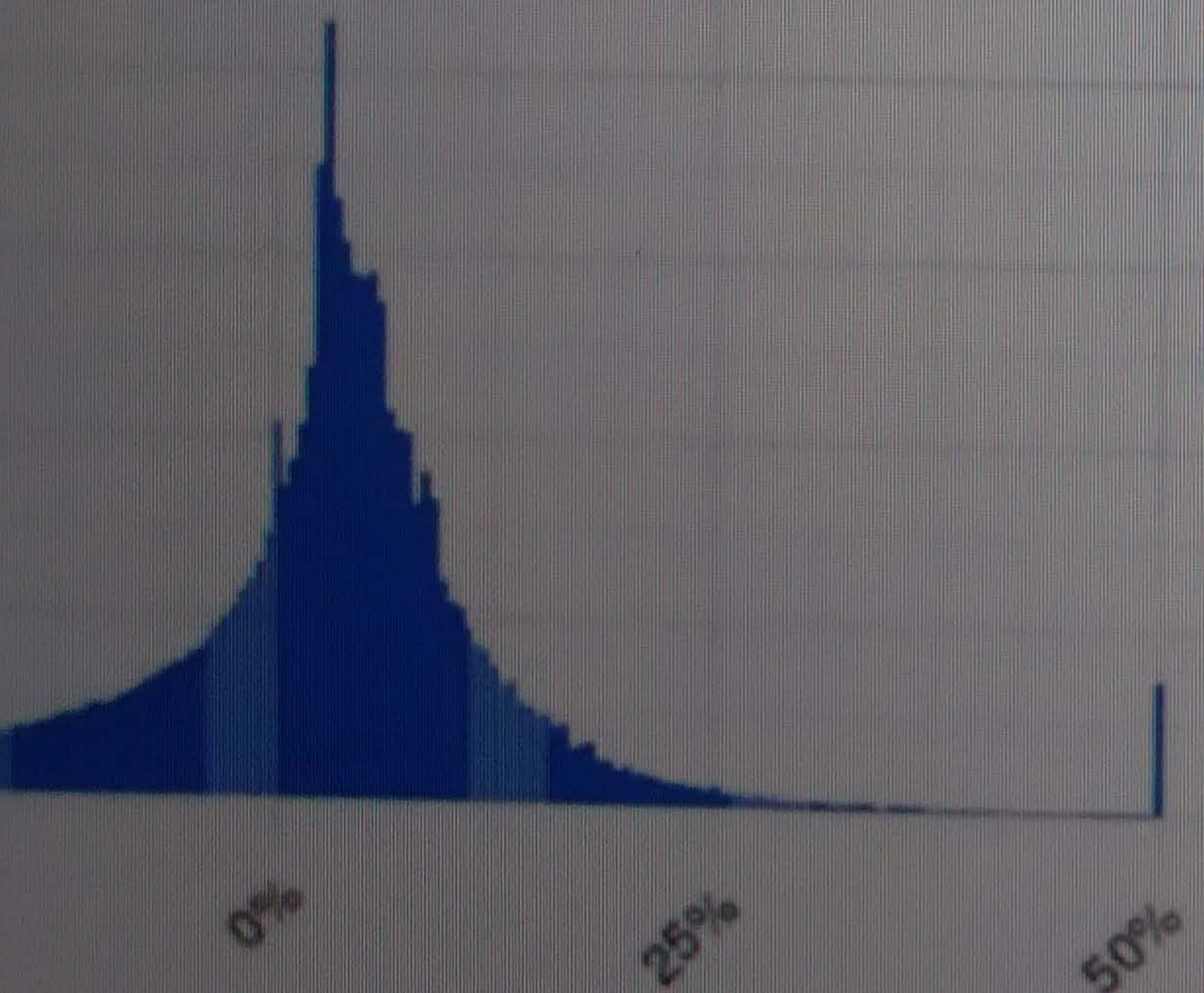


5.00%
4.00%
3.00%
2.00%
1.00%
0.00%

Like

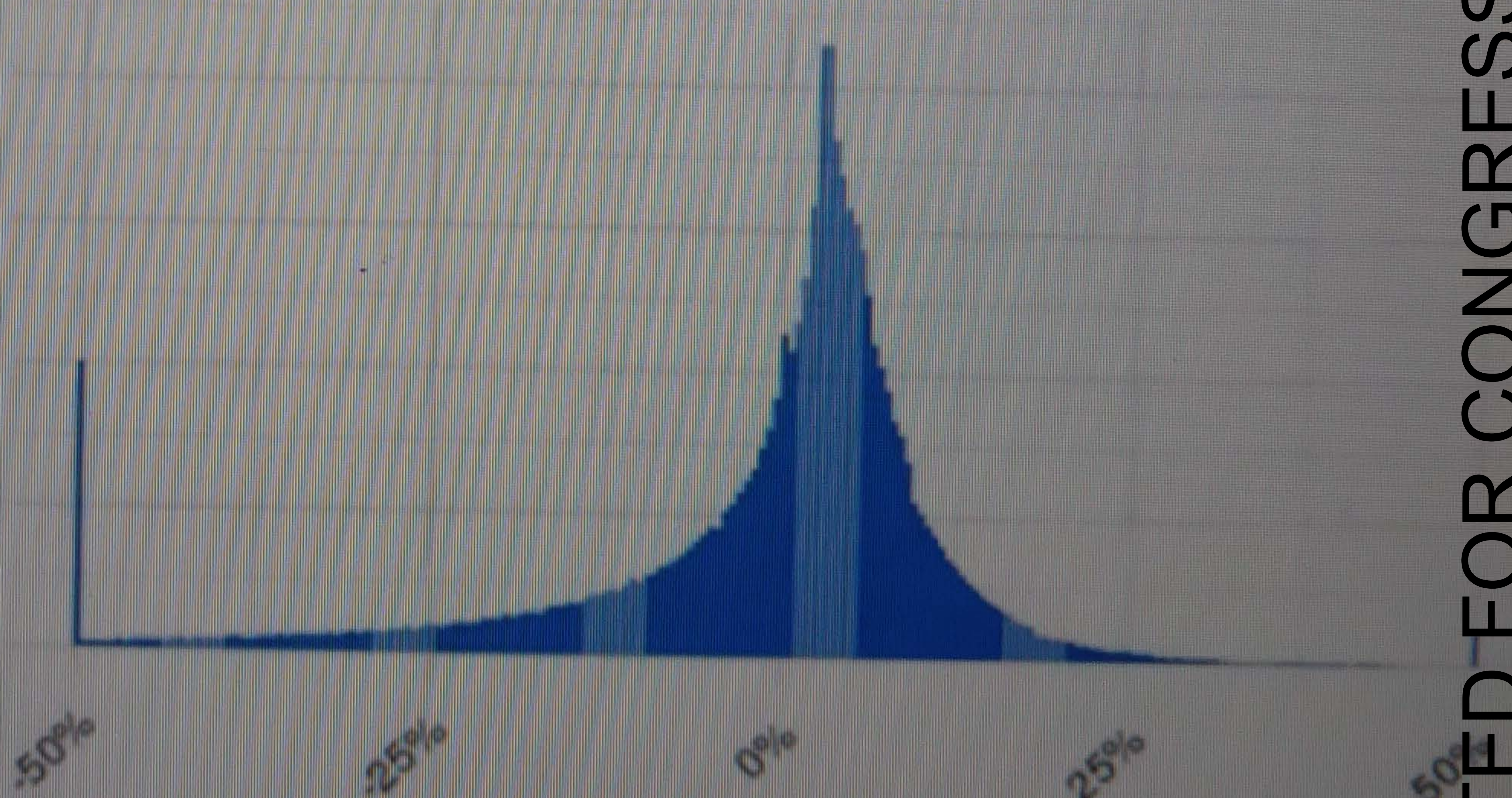


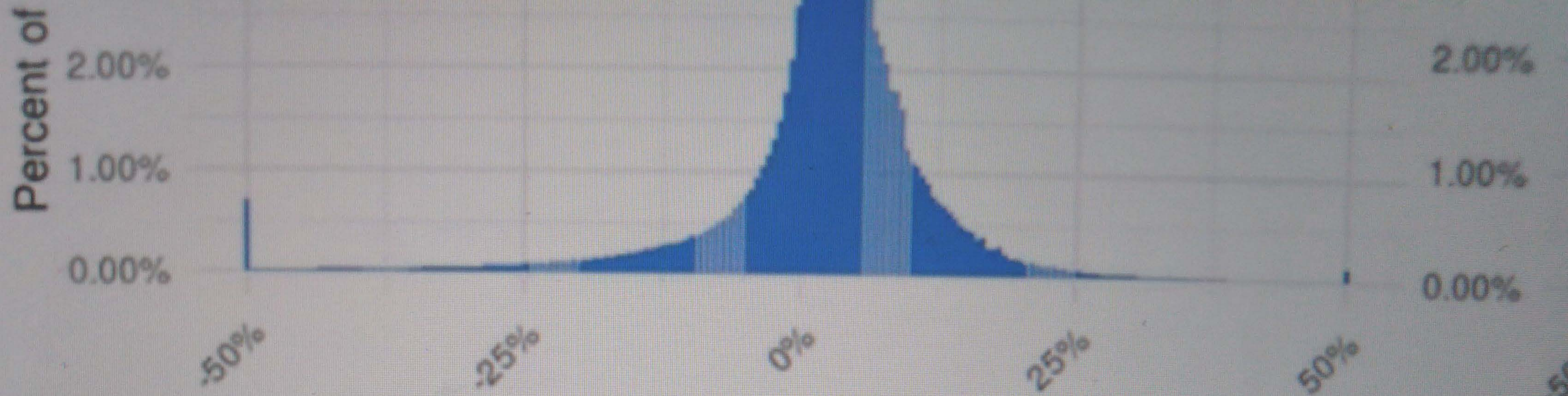
Share



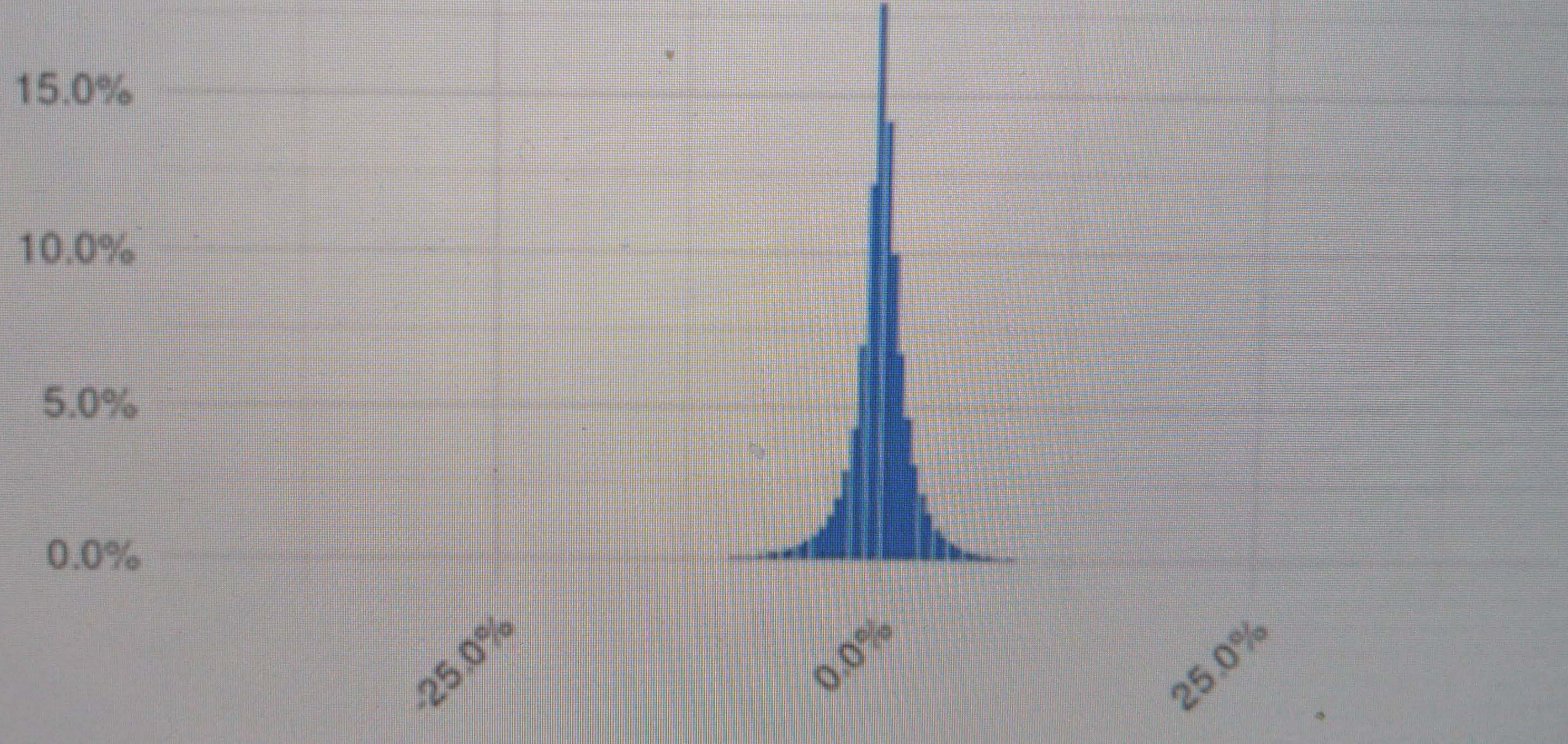
4.00%
3.00%
2.00%
1.00%
0.00%

Significant Comment





Wyt



REDACTED FOR CONGRESS

Final thoughts

Integrity is not terribly correlated with engagement and is actually positively correlated with WYT; we're not making significant tradeoffs with these metrics, at least not on average. It's possible we want to limit the amount of times where Integrity makes WYT or MSI substantially worse and that we should tune accordingly, but the average impact is generally small.

This leaves us with three options for collateral damage:

1. We focus on producers, i.e. voice
2. We limit the downsides of Integrity, perhaps by focusing on certain groups of users who are having a negative impact or certain types of content (also applicable to producers)
3. We develop other measures / proxies for what users want to see and make sure we respect those

All will continue to be explored in 2020 but hopefully this is a reasonable summary of what we've learned in 2019. Thanks for reading and Happy PSC!



17 Comments 3 Shares



Like



Comment



Share



Save

how do you square these two things: "For WYT, the stronger the Integrity demotion, the greater the WYT!" and "Integrity is not terribly correlated with ... WYT"? In the chart, it looks like Integrity is linearly correlated with Worth Your Time.



1

Like · Reply · 1y

The range on that one is small, so it does seem like WYT is less affected by cranking these demotions than other metrics. But it does provide a slight lift -- in general it's surprising to me that all these metrics are lifted by integrity demotions (at least at the 1x level). A ranking change that increases MSI seems like something we would ship because it makes the product better, anyway.



1

Like · Reply · 1y · Edited

This is all based on simulation, although I do wonder if we simply optimize for MSI so much that any deviation for current production (i.e. 1x Integrity) makes us slightly worse off in this respect. All this needs to be debiased with production experiments tbh.



1

Like · Reply · 1y · Formatted

To what extent do you think these results are influenced by our launch process? I'd guess integrity rules are a lot easier to get launched when they don't show much impact on engagement metrics, which would make it unsurprising if our current rules don't move them much. Might this look a lot different in a world where we shipped integrity rules based purely on their power to reduce harms?



7

Like · Reply · 1y

It's certainly possible but to be honest, I've never seen many launch docs with much top-line impact, although it's possible that people just don't bring them to launch at all. But generally I've only seen top-line move when we do fairly radical things.



1

Like · Reply · 1y

I would love to understand this a bit more. Do you have examples we could review?



1

Like · Reply · 1y

Yeah, people tend to not bring them to launch. 's "give more people a voice", 's Manufactured Virality, various CORGI experiments by , and the negotiation around weights in downstream engagement models by all come to mind.



3

Like · Reply · 1y

who probably has more examples of launches we've discouraged due to engagement effects

Chats

REDACTED FOR CONGRESS

...ean, people tend to not bring them to launch. Sanar's "give more people a voice", ...'s
Manufactured Virality, various CORGI experiments by ... and the negotiation around weights in downstream
engagement models by ... all come to mind.

Like · Reply · 1y



... who probably has more examples of launches we've discouraged due to engagement
effects

Like · Reply · 1y



Write a reply...



... To summarize the results another way, does that mean that 17% of VPVs in the ecosystem are problematic,
according to current definitions? My guess is that the 'current definitions' part is key, and that we're potentially missing a
bunch of other problems (or at least bad experiences) and could do more work to identify and measure more problems.

I also find those "How does Integrity impact engagement and WYT?" charts fascinating! My takeaway is that we should do
an experiment where we 32x a demotion and compare it to a 2x demotion, then actually see what's being affected more in
the 32x. That could be a super easy way to find problematic network behaviors.

Like · Reply · 1y · Edited



... It does matter quite a bit about how we define problematic as you say; I suspect a lot of the 17% may
just be from EB so it would be worth doing more granular strength increases to understand the % of VPVs eligible for
specific demotions.

Like · Reply · 1y



... amazing work ... thank you. had a few clarifying questions regarding the last part of your note:
what do you mean specifically by "collateral damage"?

1. "We focus on producers, i.e. voice", what does this mean? and why should we do that?
2. "We limit the downsides of Integrity" what do you mean by downsides of integrity?
3. "We develop other measures / proxies for what users want to see" By developing other measures/proxies do you mean
measures other than MSI and WYT? if yes, why?

thank you.

Like · Reply · 1y

... So a lot of this is discussed in previous notes. Re "voice", that's this idea that we should make sure
that producers are able to express themselves and that our demotions don't unintentionally silence people. By the
downsides of Integrity, this is mostly referring to our impact on engagement and 'what users want to see.' For other
measures, this is mostly getting at other measures for WYT, specifically more proxies for 'what users want to see.'
This could mean understanding this kind of question from the perspective of what is entertaining, informative,
meaningful, etc.

Like · Reply · 1y

Chats

REDACTED FOR CONGRESS

1. we focus on producers, i.e. voice , what does this mean? and why should we do that?

2. "We limit the downsides of Integrity" what do you mean by downsides of integrity?

3. "We develop other measures / proxies for what users want to see" By developing other measures/proxies do you mean measures other than MSI and WYT? if yes, why?

thank you.

Like · Reply · 1y

So a lot of this is discussed in previous notes. Re "voice", that's this idea that we should make sure that producers are able to express themselves and that our demotions don't unintentionally silence people. By the downsides of Integrity, this is mostly referring to our impact on engagement and 'what users want to see.' For other measures, this is mostly getting at other measures for WYT, specifically more proxies for 'what users want to see.' This could mean understanding this kind of question from the perspective of what is entertaining, informative, meaningful, etc.

Like · Reply · 1y

I'm just here for the Let's Go cover photo.



Like · Reply · 1y

Excellent note !! I want to make sure I am understanding this properly... Would you agree that a take home message is that increasing the strength of current integrity demotions comes with important tradeoffs to engagement or other topline metrics. Therefore a potentially better path forward is to improve the relevance of the demotions, thus identifying content that users would agree more we should remove?

Like · Reply · 1y

Hmm I generally do agree that many demotions should just be constructed such that they are in line with what users want. That's a big opportunity and at that point collateral damage becomes a very different thing because our goal is effectively just the same as the goal of Feed ranking generally, i.e. make sure people like what they see. That said, I think we would probably need to strengthen current demotions quite a bit (especially for non-EB) to see much meaningful top line effect.

The other part of this work is probably better described in <https://fburl.com/gpz7dyek>; basically I think that if a demotion does not align with user value, then we should tune it such that we optimize for ecosystem health subject to constraints on situations where users don't get to see what they want to see.

Like · Reply · 1y



FYI



Like · Reply · 1y

This is super interesting and a key thing we've been thinking about on recommendation integrity as well. I'm curious, did you do any analysis on the impact for entities versus content? cc

Like · Reply · 1y

Chats

REDACTED FOR CONGRESS