Title: Using MTurk Survey Data to Model $GME Ownership Among U.S. Households
Author: edpeepers
Created 2021-09-07 05:02:53 UTC
Permalink: /r/DDintoGME/comments/pjgcn7/using_mturk_survey_data_to_model_gme_ownership/
Url: https://www.reddit.com/r/DDintoGME/comments/pjgcn7/using_mturk_survey_data_to_model_gme_ownership/

Tl;dr I replicated the survey research of u/Get-It-Got analysis using Amazon Mechanical Turk to a) extend the longitudinal aspect of the research and support meta analyses, b) replicate and validate using a different survey sampling platform (MTurk), and c) experiment with the survey methodology to improve accuracy. My conservative estimates align with the previous studies. My less conservative estimates should be taken with a huge grain of salt but are thoroughly tit-jacking. Read on!

As always, I'm standing on the shoulders of giants and offer my kudos to everyone investigating this from different directions.

Hypothesis + Why did I do this?

I'm a nerd and inherently somewhat distrusting. I don't know jack about Cayman Island tax avoidance schemes, but I know survey research! So what better way for me to validate part of the hypothesis myself, and transparently share the results with others? The underlying hypothesis remains the same – retail ownership of GME exceeds the float – but I wanted to see if I could build on prior work by asking about household data + estimate a less conservative upper bound to ownership. In addition to collecting more data to improve confidence in the results, it's also important to sample from different populations, so I surveyed using MTurk to gain a slightly different perspective. If the results align with survey research using Google's consumer population, that further increases confidence that the data is representative of the larger population.

Last but not least, I have long been fascinated by organic, online communities and spontaneous leadership emergence, from online gaming to political movements. The unfolding $GME events over the past year have not escaped my attention!

None of this is financial advice. I am not a financial advisor. I have personally invested in GameStop using a cash account with a reputable broker, only investing what I am comfortable losing, with a Buy & Hold plan. I enjoy the taste of crayons and you should read the DD and come to your own conclusions when making any investment decision.

Methodology

I set up a branching survey through Amazon Mechanical Turk, administered via Qualtrics, to collect data over the 2021 Labor Day weekend. The survey and MTurk task launched after market close on September 3rd and the survey ran through the evening of September 6th. The timing was intended to collect a snapshot of data in the absence of market activity (U.S. markets were closed Monday, September 6th).

Participants

Participation was restricted to U.S. adults (18 years or older) using MTurk's built-in filtering criteria. To be eligible to complete the survey, participants also needed to have completed at least 1,000 HITs (MTurk tasks) with a HIT approval rate greater than 98%, in line with standard MTurk survey guidance (if anything, I was less strict than the guidance for academic research). Participants were awarded $0.10 for completing the survey. 532 participants completed the survey.

Survey Branching

The survey included 3 primary branches. Branch A (GME Control, n = 202) asked participants to respond to the following question regarding individual $GME stock ownership, and then terminated.

Branch B (AAPL Control, n = 172) was identical to Branch A but asked participants about their $AAPL

ownership as a control.

Branch C (GME Households, n = 164) asked participants to respond with regards to their household, a key differentiator from previous studies. If the participant indicated their household owned one or more $GME shares, the survey then randomly branched to Path C.a (share ownership in multiple choice format) or Path C.b (share ownership as a raw numeric input). If the participant's household owned no shares, the survey terminated.

Analysis & Results

I am not extrapolating estimates based on Branches A & B – I actually recommend that data be aggregated with the prior research. I do think it's important to note that
* I found individual $GME ownership to be roughly in line with the prior studies. Even though I added additional share buckets (while retaining backwards compatibility), they did not change the median reported share ownership.
* I found individual $AAPL ownership to be quite a bit higher than the control study previously reported. I don't know why this is, but it's possible that Google's consumer survey population is particularly light on Apple device owners (and Apple device owners may own more AAPL shares)
* If there were a blanket bias or difference in individual ownership for both stocks I would suspect sample bias. Since it was just AAPL, I believe it has something to do with that stock. Get-It-Got noted in the previous study that the reported AAPL shares seemed REALLY low, so this may be a indicate a problem with the control rather than an issue with the GME estimates.

[Branch A ($GME Control)](https://i.imgur.com/GpDDs6N.png)

[Branch B ($AAPL Control)](https://i.imgur.com/oDyxnvC.png)

Given that the individual GME ownership was in the ballpark with the previous, most conservative estimates, I'll spend more time on the household data!

Remember that the household data was presented first as an ownership yes/no, followed by random assignment to either the open ended numeric field or a multiple choice question.

[Reported $GME Household Ownership %](https://i.imgur.com/icPDIjm.png)

29% of Branch C participants reported their household owned $GME. They were randomly assigned to one of two follow-up questions.

[Path C.a - $GME household ownership, multiple choice](https://i.imgur.com/xrcWMjf.png)

[Path C.b - $GME household ownership, raw entry](https://i.imgur.com/tK6uS7Y.png)

Note that I extended the upper ranges of the responses. Across the two paths, 7 participants indicated they owned more than 101 shares. One participant indicated they owner 5,000 shares, making them a clear outlier. Still, I suspect we can use this data to draw a more accurate estimate than in prior studies (which to their credit were aiming for an incredibly conservative estimate that STILL showed retail ownership greater than the float). I attempt here to extrapolate U.S. household ownership with a conservative estimate (similar to Get-It-Got), a moderately conservative estimate, as well as a widely speculative estimate.

Conservative Estimate (n = 48)
Extrapolating to [120,756,048 U.S. households](https://www.census.gov/quickfacts/fact/table/US/HSD410219) at 29.27% reported $GME ownership, with a median value of 6 shares, we get 212 million shares. Higher than last month's survey study update (could be due to recent run-up?) but not too far away. I consider this confirmatory of the underlying, conservative hypothesis that retail ownership greatly exceeds the float. Note the median share ownership is actually lower than the median I found in the individual data – and the basic extrapolation is still well over 200 million.

Moderately Conservative Estimate (n = 48)
Now things get very interesting. If we calculate the mean value based on the conservative buckets (e.g. the 5,000 share participant gets trimmed down to 1,001 shares in the aggregated bucket), we end up with more than 1.7 billion shares owned by U.S. households. That seems high but... not completely out of the picture? That represents a mean of just over 50 shares. Gallup recently reported over 50% of Americans own stock, so the number is within the realm of possibility. Even trimmed though, the outlier is really pulling up the mean – excluding them would drop the mean result to about a billion shares. That is still a lot of shares.

Wildly Speculative Estimate
What if we ignore the conservative, multiple choice buckets and focus on the raw reported data (n = 20) from Path C.b? I'm glad you asked! Throwing caution and outliers to the wind, the mean of the raw response data gives us a modest 10.5 billion shares owned by U.S. households. I really don't believe that one – the participant reporting 5,000 shares has tremendous influence on the mean! But I had to share it for fun and full transparency. One last interesting note – if we drop the 5,000 share outlier again, that brings the raw data average back down to about 49 shares per household, which extrapolates out once again to 1.7 billion shares. I like it when numbers converge. If the survey samples do represent the broader U.S. population (admittedly a big if), then the true U.S. household ownership could now be somewhere in that 1-2 billion range.

[Link to estimates](https://i.imgur.com/He8F2CZ.png)

Summary
* I found conservative estimates of individual $GME ownership in line with u/Get-It-Got previous research (GME Control)
* I found conservative estimates of individual $AAPL ownership higher than u/Get-It-Got previous research (AAPL Control), with a hypothesis as to why
* I found conservative estimates of median household $GME ownership in line with u/Get-It-Got previous research
* I found estimates of mean household $GME ownership much higher than u/Get-It-Got previous research

Limitations & Caveats (aka Cool Your Tits)

Does a filtered MTurk sample represent the entire country? That's probably not the case. So how is it different?

[Prior research](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198835) found MTurk participants trended younger and were more likely to be college educated. We do not know whether that was the case in my MTurk sample (I didn't pony up for all the extra demographic data) but we can look to our Branch A control to piggy-back, albeit crudely, on the demographic data collected from prior GME studies conducted by u/Get-It-Got. That said, I would hypothesize the results from the MTurk study are more likely to overinflate the number of shares reported (for GME or AAPL).

What if multiple participants from the same household responded?

These results may need weighting to reflect an uneven distribution of ownership across households. For example, if we had data from every adult in the U.S., many of them would of course be referring to the same households and we would ideally weight responses based on size of household and potentially other factors. As expected, the reported household ownership was higher (29%) than individual ownership (22%) indicating participants were responding with their larger household in mind when prompted. Collecting additional data (household size, household income, etc) could help us better weight and extrapolate the results, but this study does help address a concern (extrapolating from individual responses to household ownership) with a conservative estimate of ownership that supports the hypothesis that retail GME ownership is much higher than the float.

Targeted weighting in Qualtrics

I thought I was setting up a very deliberate weighted branch in Qualtrics. 40% of participants would route

to Branch A, 10% to Branch B, and 50% to Branch C. This would give me plenty of sample for the new household versions of questions, and reasonable samples of the individual ownership questions to serve as controls. Well, Qualtrics screwed up despite my best QA and multiple manual resets, instead routing participants to all three primary branches evenly. That is why you see roughly similar number of responses in each branch. I recommend future researchers avoid Qualtrics for surveys with sophisticated, uneven branching flows, since they are apparently unreliable at best. That, and eat fewer crayons while conducting research in case user error was involved.

Did a participant really report owning 5,000 $GME shares?

We have no reason not to believe it. MTurk users are motivated primarily by speed. The faster they can complete tasks, the more money they make per hour. Taking the time to enter a number with more characters (vs 1-9) does not make logical sense. Could someone have been messing around? Possibly. More importantly – do they represent U.S. households? That seems less likely and is why I report the value, but discount it as an outlier in the least conservative estimate and offer estimates both with the outlier included as well as excluded.