**QUESTION**

How does the random sample response agreement between expert answers and click worker answers differ for low and high lameness differences?

What does this tell us about the number of click workers we should hire?

**METHODS**

Divide the original data frame into 2 dataframes of high and low lameness differences.

Filter questions with expert rating of more than 1 or less than -1 (easy questions).

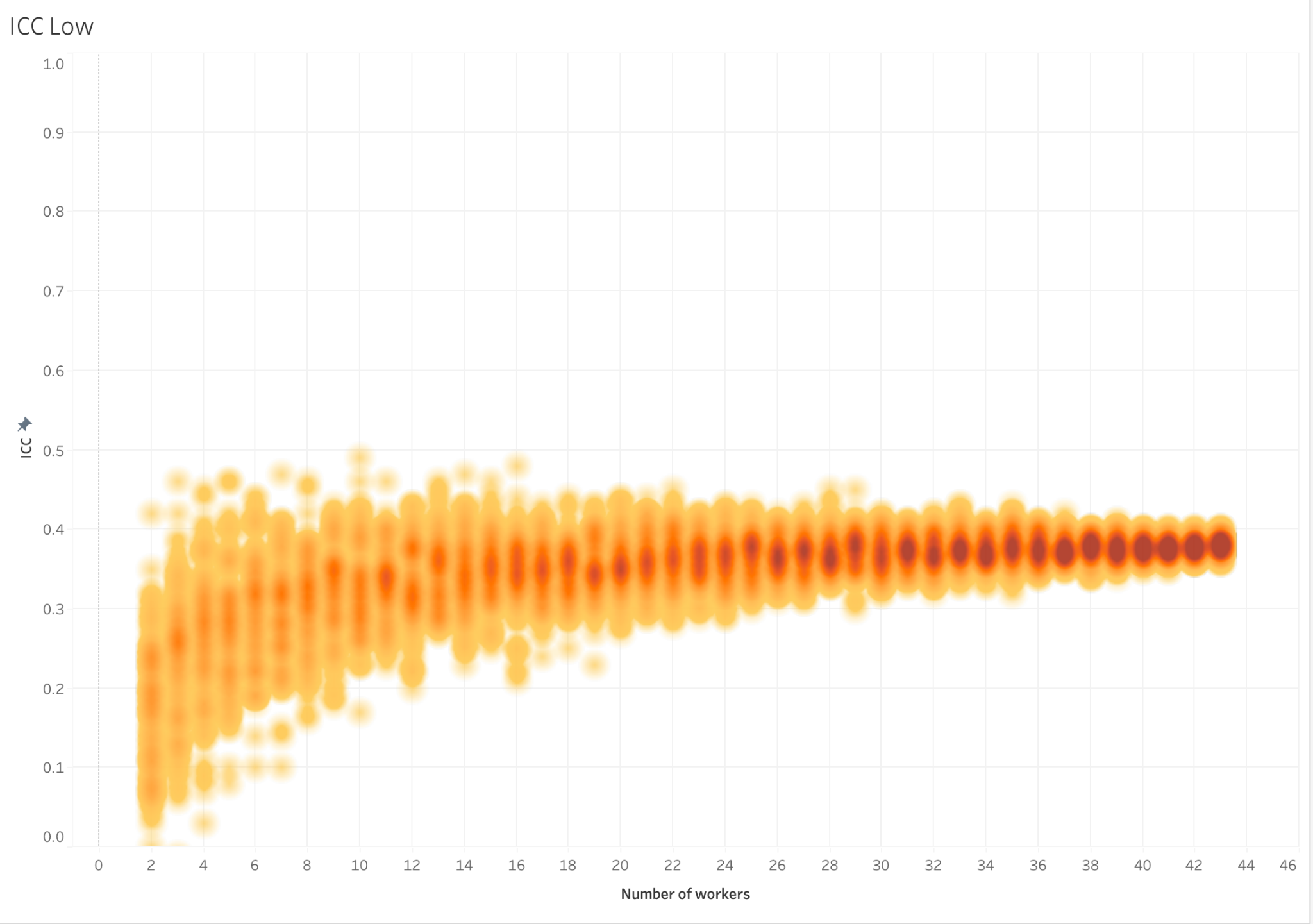
Filter questions with expert rating of less= to 1 and more= to -1 (hard question).

For both, randomly sample 43-2 workers 100 times from all HIT-s and assess agreement with experts.

Generate heat maps for both.

**RESULTS**

Figure 1: low lameness differences (hard questions)

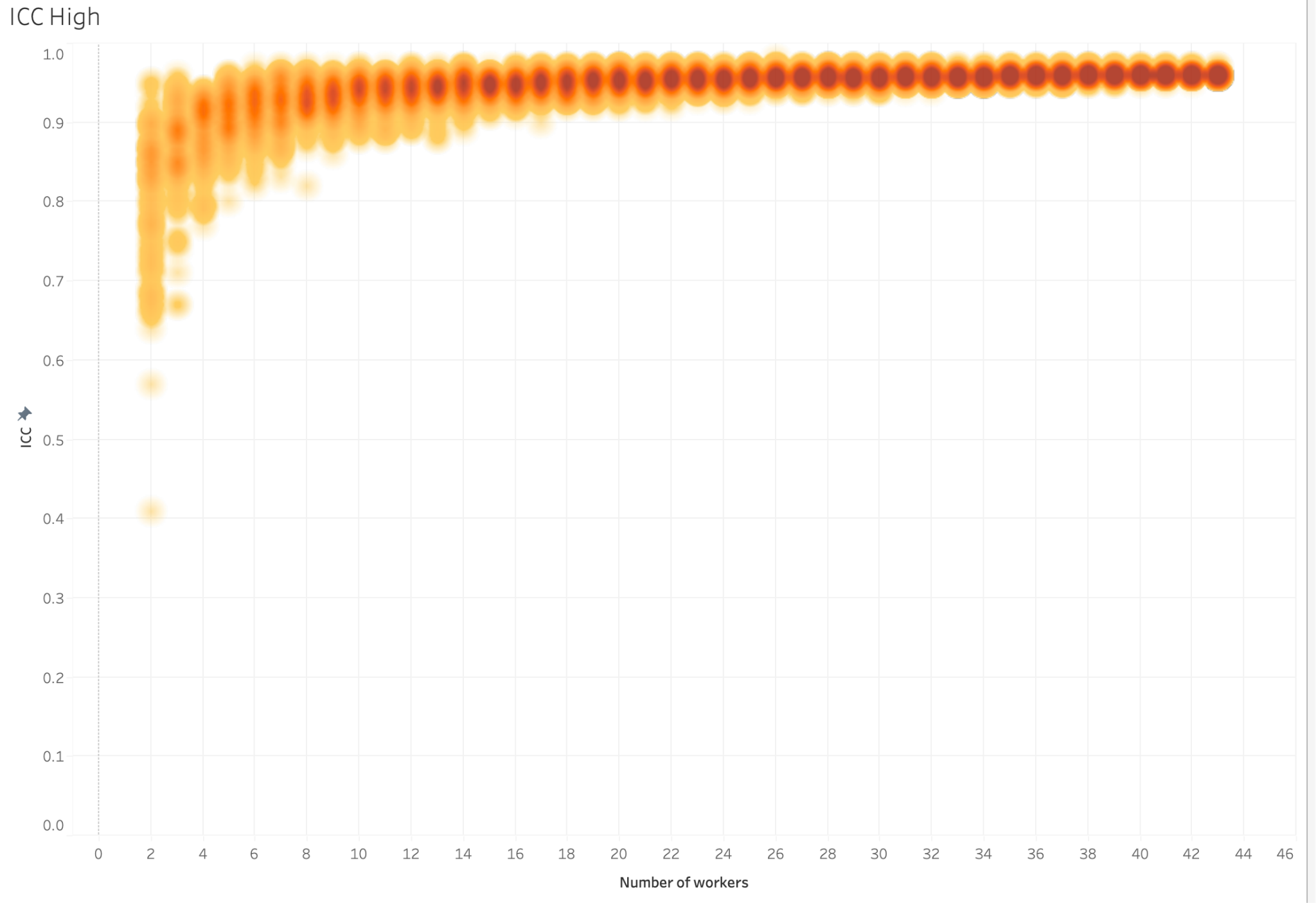


Highest ICC between average click worker and average experts = 0.39, increasing the number of click workers does not improve the results

ICC among experts = 0.2.

When using the traditional manual locomotion scoring system, the absolute score difference between 2 cows median = 0.3, ranging from 0 to 1.

Figure 2: high lameness differences (easy questions)



Highest ICC between average click worker and average experts = 0.95. Only need about >= 9 click workers to reach high ICC

ICC among experts = 0.88

When using the traditional manual locomotion scoring system, the absolute score difference between 2 cows ranging from 0. to 3, median = 1.2.

**CONCLUSION**

for questions between -1 to 1 degree of lameness difference, there is really not much you can do about it, just treat it as a tie

“how much more lame” does not matter that much, it’s only useful for us to identify hard and easy questions

**QUESTION**

Should we still do attention checks when launching new video pairs in the phase 2?

**METHODS**

Literature search and google search

**RESULTS**

https://docs.google.com/document/d/16GCiG4OljHPO0QLLrAyesHNFOfMbDnbHBLmH\_Sv6dB0/edit?usp=sharing

**QUESTION**

“How much more lame” from pair-wise assessment reflective of the absolute lameness score difference between the 2 cows?

**METHODS**

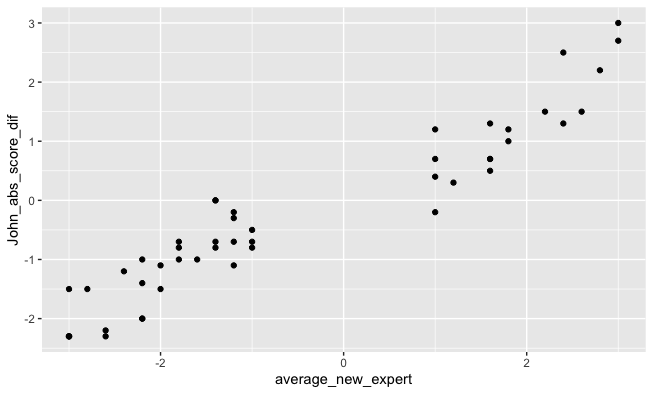
Assess interobserver reliability among experts for both questions with low and high lameness differences.

Assess interobserver reliability among click workers for both questions with low and high lameness differences.

Create a scatterplot with the average expert rating as x-axis and John’s absolute scoring difference as the y-axis.

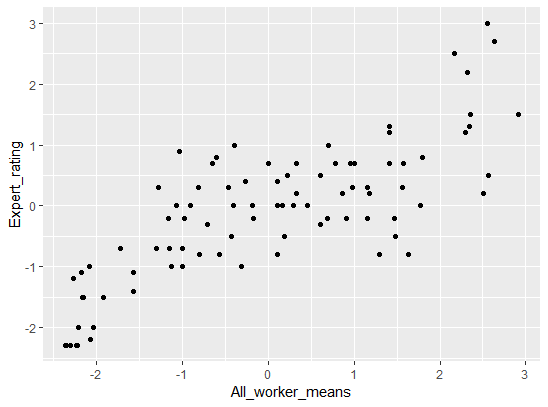
**RESULTS**

Figure 1



On the x axis is the average “how much more lame” assessed by the experts based on pairwise assessment, y axis is the absolute lameness score difference assessed using traditional manual locomotion scoring system. They do agree with each other.

Figure 2



On the x axis is the average “how much more lame” assessed by the click workers based on pairwise assessment, y axis is the absolute lameness score difference assessed using traditional manual locomotion scoring system. The “how much more lame” is indicative of the absolute score differences, but not always the case.

**CONCLUSION**

“how much more lame” does not matter that much, it’s only useful for us to identify hard and easy questions

**QUESTION**

How to identify easy (average expert response for absolute lameness difference > 1, ICC > 0.95) and hard (average expert response for absolute lameness difference <= 1, ICC < 0.3) questions based on click worker response (click worker recruited n = 10, 15, 20)?

Based on our previous result, we tentatively propose: if the average response has an absolute lameness difference > 1, we are certain that this is an accurate response. If the average response has an absolute lameness difference <=1, we just treat them as tie.

**METHODS & RESULTS**

1. We evaluated:
2. For hard questions (identified based on average expert response, n = 47), does the average response from all 50 click workers also reside between -1 and 1, so that the absolute lameness difference also <=1?
   1. Answer: NO. We got 3 questions with average click worker response < -1 (range -1.04 to -1.28). We also got 14 questions with average click worker response > 1 (range 1 to 2.51) 17 questions in total.
   2. Using the tentatively proposed method, we calculated sensitivity (true positive) and specificity (true negative), classify hard questions as positive cases, and easy questions as negative cases.
      1. sensitivity = 0.78
      2. specificity = 0.66
   3. Insight: why do we got way more on the positive side than the negative side? Positive means cow on the right much more lame. Are we seeing the effect of people are right-handed, so they are more likely to pick those on the right hand side?
   4. Insight: If we use the average response only from workers that pass both attention checks (negative and positive ones), and filtered based on cluster (pass\_cluster\_worker\_means), we still have 9 questions with average click worker response < -1 (range -1 to -1.89), and 13 questions with average click worker response > 1 (range 1.11 to 2.78). However, for all these questions, almost all of them, the average click worker agree with average expert on “which cow is more lame”, also the lameness difference is tiny based on expert response.
3. For easy questions (identified based on average expert response, n = 41), does the average response from all 50 click workers also reside > 1 or < -1, so that the absolute lameness difference also > 1?
   1. Answer: NO. We got 4 questions with average click worker response reside between 0 and -1 (range -0.91 to -0.57). We got 3 questions with average click worker response reside between 0 and 1 (range 0.7 to 0.98). 7 questions in total.
   2. Insight: so our proposed approach (if the average response has an absolute lameness difference > 1, we are certain that this is an accurate response. If the average response has an absolute lameness difference <=1, we just treat them as tie. ) can’t be used. We risk treating the cows that have a significant difference as a tie, and treating cows that are truly tie to have significant difference.
   3. Insight: If we use the average response only from workers that pass both attention checks (negative and positive ones), and filtered based on cluster (pass\_cluster\_worker\_means), all questions have average click worker response of absolute lameness difference > 1 (range 1.11 to 3).

**CONCLUSION & SUGGESTIONS**

It’s very possible that we need to **re-define what is hard and easy questions based on click worker r**ameness difference between 2 cows?

**esponse**. It seems like among the hard questions (defined by average expert response), the click workers that passed the strictest attention checks and filters on average could agree with experts on “which cow is more lame” on many hard questions. Click worker responses amplified the degree of difference observed by the experts, which potentially suggest again that the “how much more difference”, is not that useful for generating lameness ranking, it’s more useful in determining the “level of confidence” for click workers to identify which cow is more lame, and more useful in being used as a threshold for separating hard and easy questions.

**Borbala comments:**

Before spending much time and effort on determining which questions are hard and which are easy it would be good to zoom out and see how we will use this information when creating a lameness hierarchy or sorting the cows. Knowing this is important because it will inform us how accurate we need to be and to what extent we can get away with a less accurate easy and hard distinction.

Another way to think about it: What would be the consequence of using the proposed method of -1/+1 threshold? What is the consequence of inaccuracy down the line? What do we gain by trying to fit a model to our average expert responses as ground truth? Is this gain worth the extra time and money (more click workers for each question, more analysis etc)

**QUESTION**

How to use a machine learning algorithm to generate a validated cutoff point in order to identify video pairs that should be treated as ties and those that have a clear lameness differences

**METHODS**

use machine learning models to learn the features of easy and hard questions.

Question labeling

* Hard questions = indistinguishable pairs that needs to be treated as ties. They are identified as if < 3 out of 5 experts agree on the directionality of the pairwise lameness comparison (which cow is more lame), or if click workers and experts do not agree on average which cow is more lame.
* Easy questions = distinguishable pairs that there is a clear lameness difference between the 2 cows. They are identified as if >= 3 out of 5 experts agree on the directionality of the pairwise lameness comparison (which cow is more lame), and click workers and experts agree on average which cow is more lame.

Data preparation and Model selection

* I created 18 different dataframes to run on 2 simple machine learning models (decision tree and SVM) to classify easy questions and hard questions based on click worker response
* I identified 4 factors that could impact the performance of the models, and shall test these out
  + 1. Filtering effect (how do we filter workers whose responses we would use)
    - All\_worker: all click workers were used without filtering
    - Easypass: only click workers who passed the positive attention check (2 cows have significant lameness differences)
    - Pass2: only click workers who passed both positive and negative attention checks (same cows mirrored)
    - Cl\_all: only click workers that formed big cluster based on Eucledian distance clustering
    - Cl\_easypass: only click workers who passed the positive attention check and formed big cluster based on Eucledian distance clustering
    - Cl\_pass2: only click workers who passed both the positive and negative attention check and formed big cluster based on Eucledian distance clustering
  + 2. Number of workers effect (how many workers do we use after filtering)
  + 3. Feature selection effect (which features do we use to predict hard and easy questions)
  + 4. Model selection effect (which machine learning method would perform better)

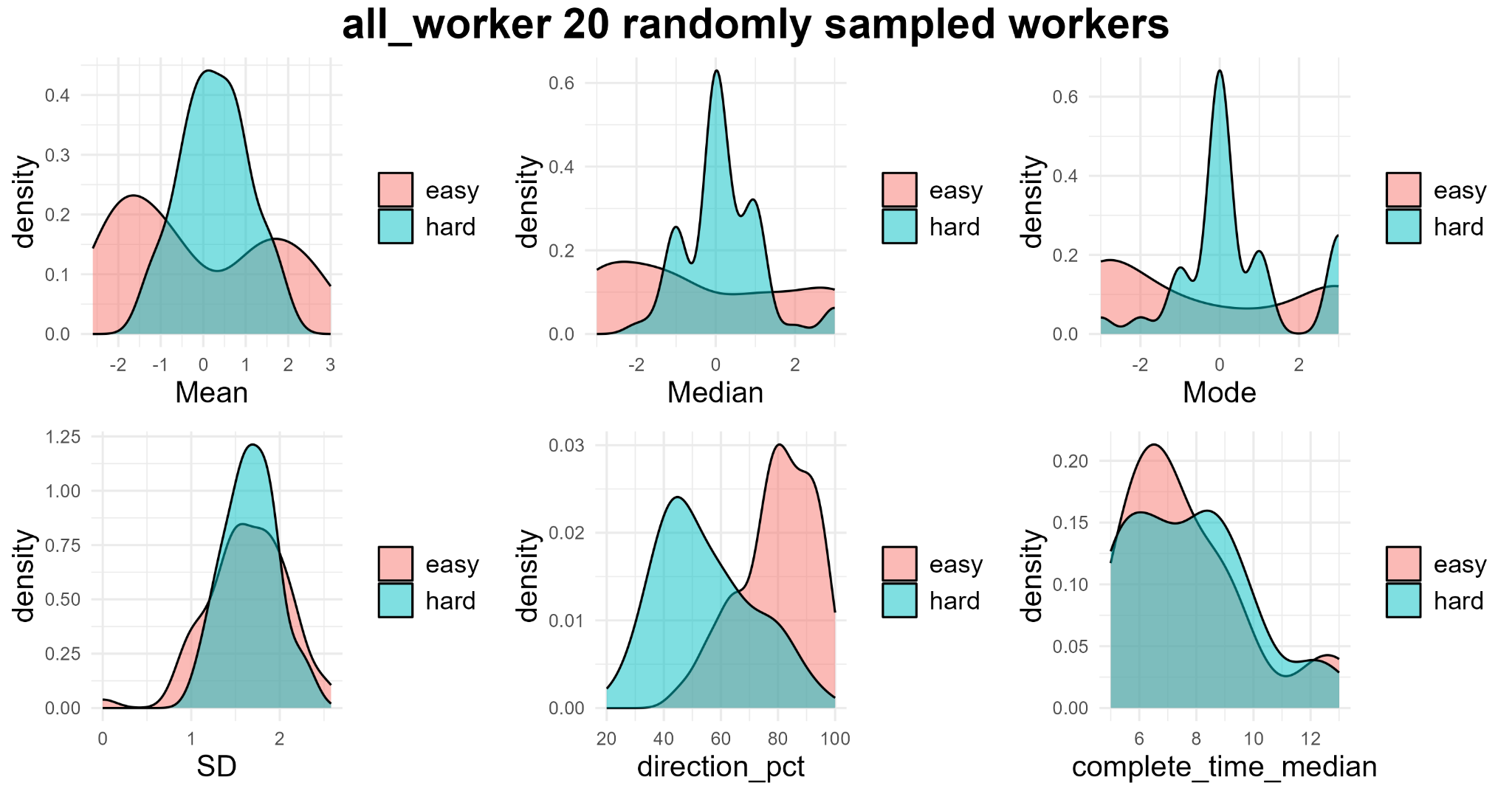




**RESULTS & DISCUSSION**

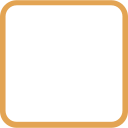
**[Exploratory data analysis]**

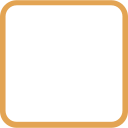
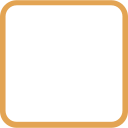
1. **Distribution plot**: Using each filtering method, after randomly sampling n = 10, 15, 20 click workers, plotted distribution of each feature for hard and easy questions



**Insight**: Mean and direction\_pct are the top 2 features that are able to separate easy from hard questions, Median might also work as well. The other features are bad features for this classification task.

**Question**: Is this the same regardless of sampling method and number of workers used? Does filtering of workers make it easier for certain features to separate easy from hard questions?

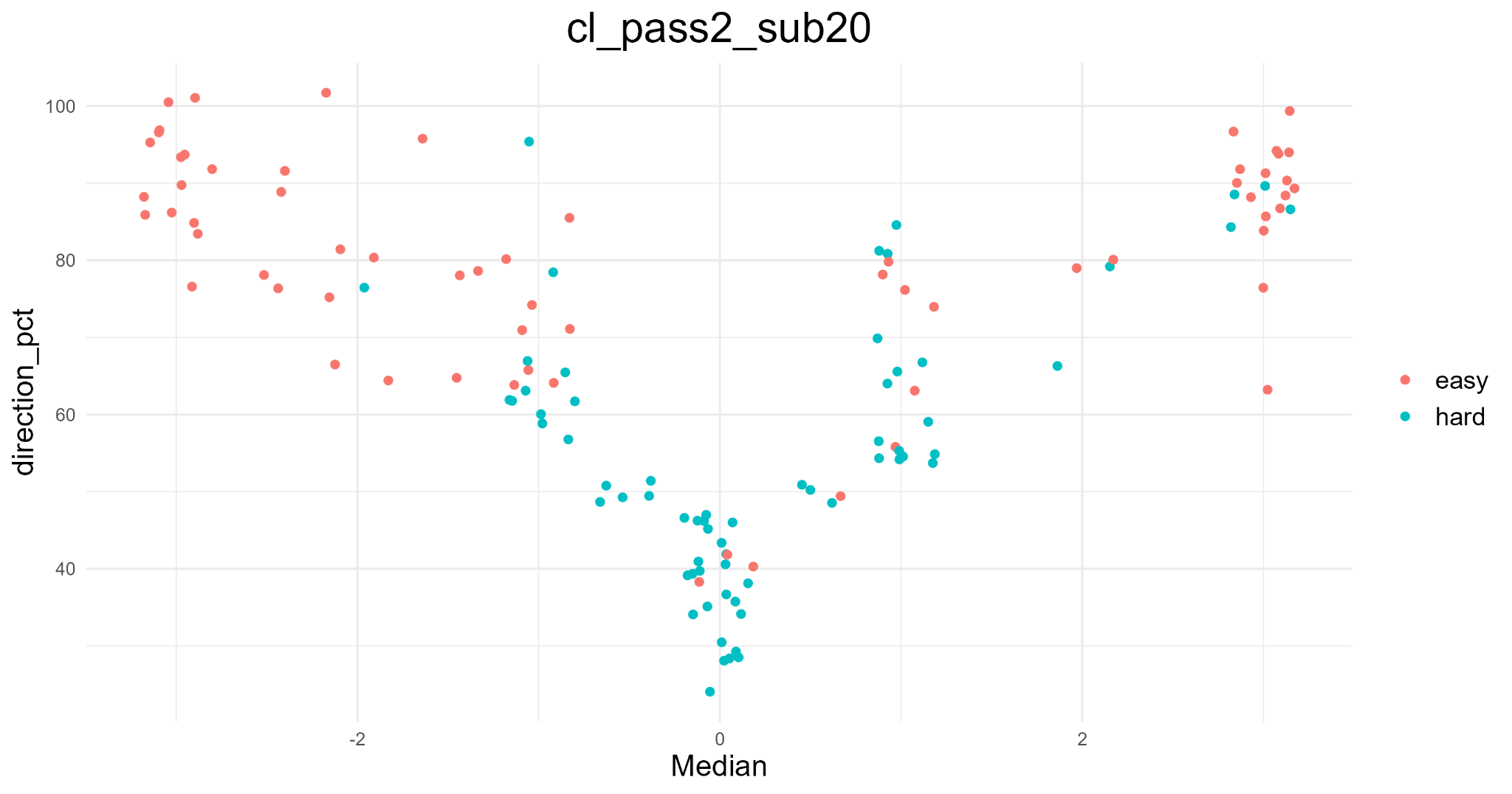
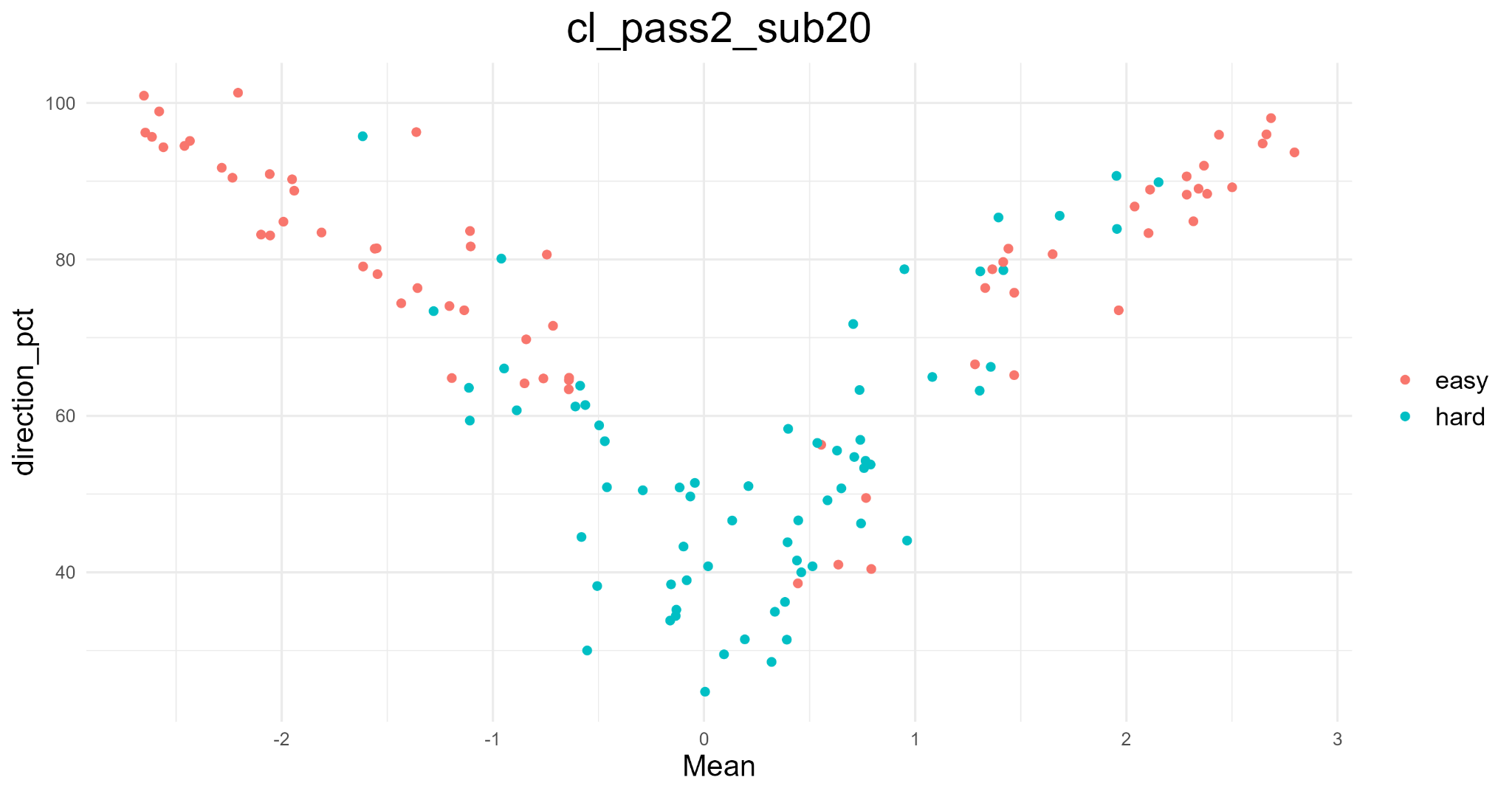




**Insight**: It seems like filtering workers based on Eucludian distance clustering + those who passed both positive and negative attention chek cquestions aid in making the 2 classes more seperable based on direction\_pct feature, also a little bit on Mean. The number of workers randomly sampled did not appear to have a significant effect based on eyeballing the plots (did not attach the plots here).

We could use direction\_pct to look for a natural discontinuity between easy and hard questions

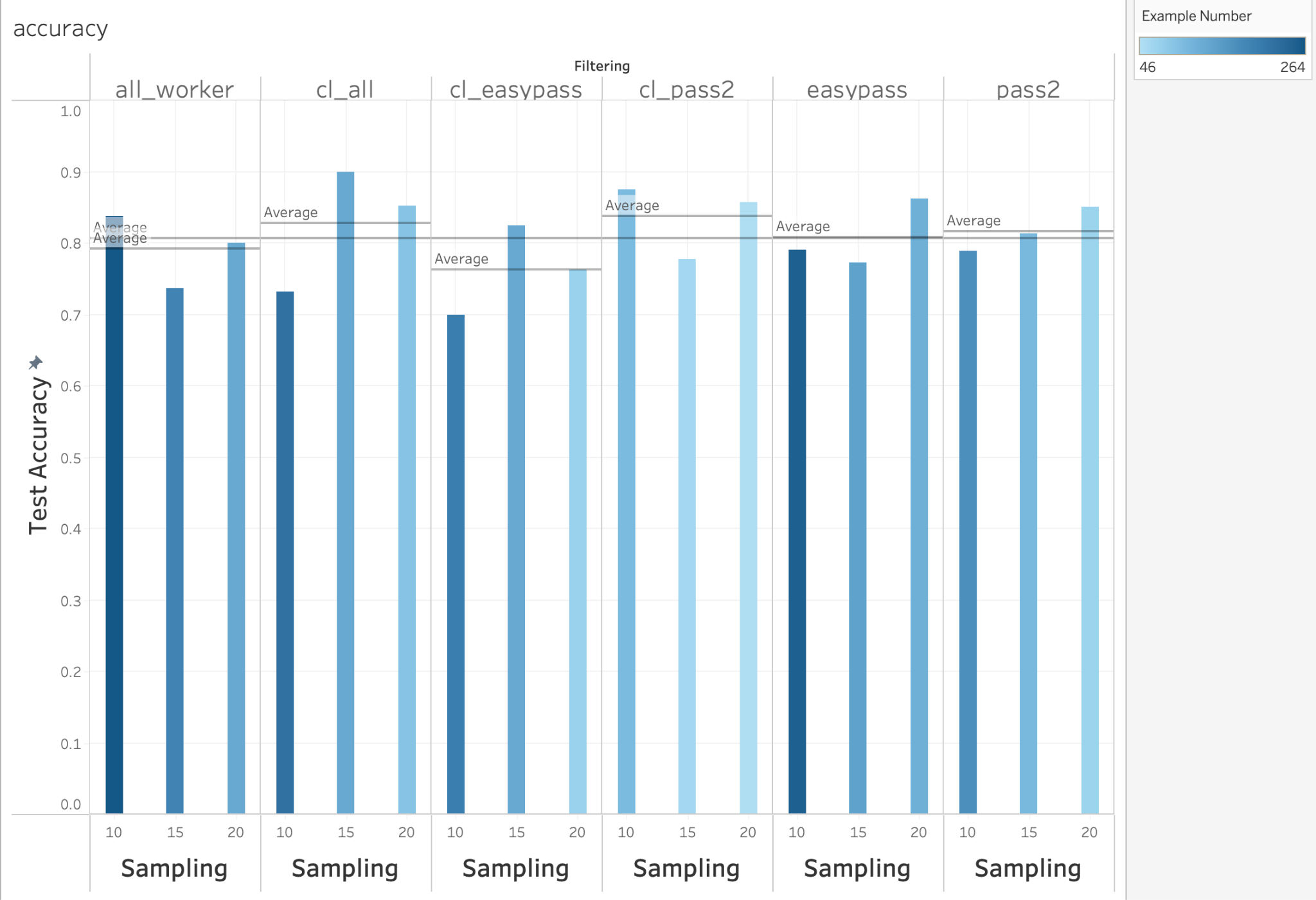
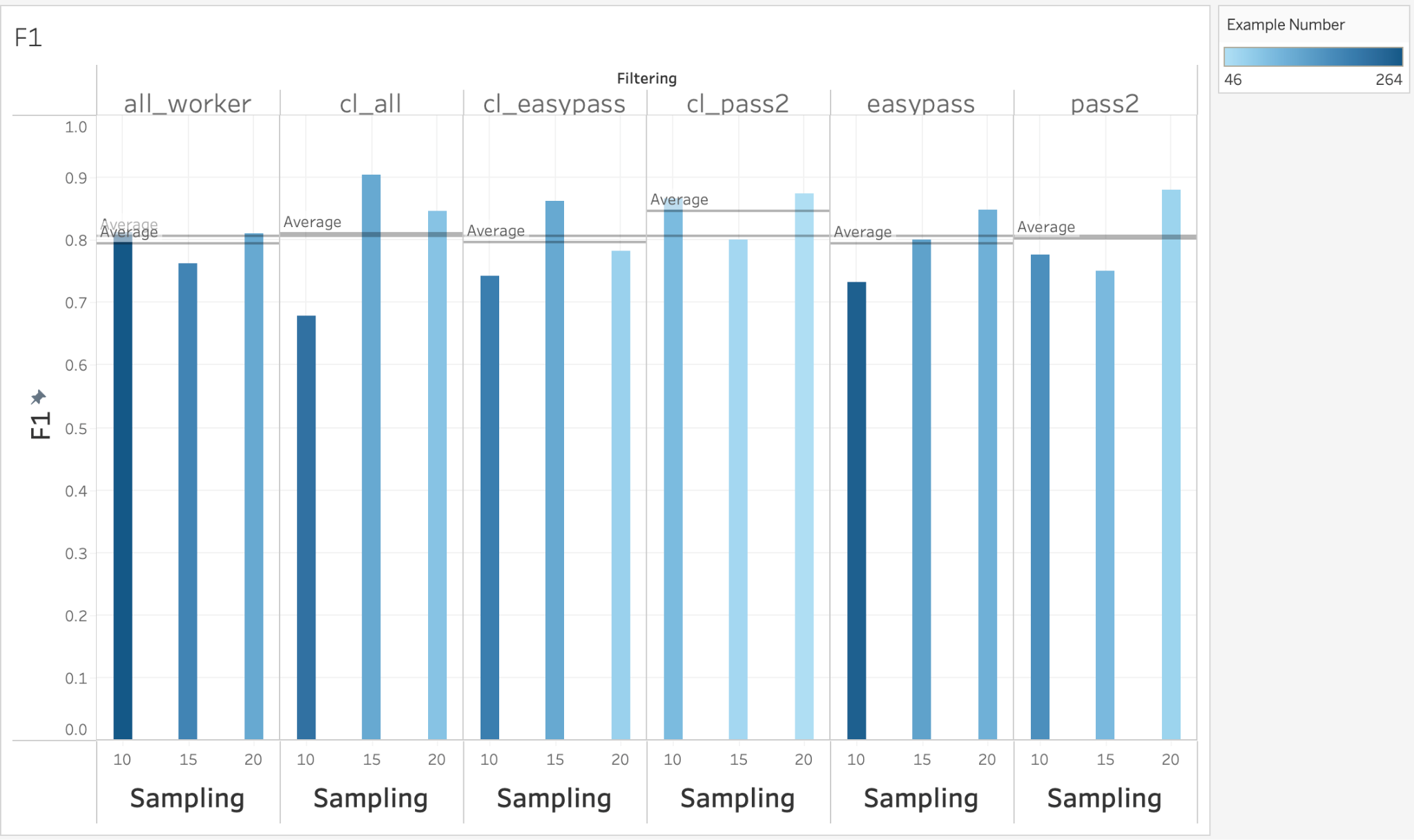
1. **Scatter plot**: Plot scatter plot for Median X direction\_pct; and Mean X direction\_pct, to see if together they can separate the 2 classes better.



**Insight**: It seems like both Median X direction\_pct and Mean X direction\_pct are doing a somewhat good job in separating the 2 classes, although not perfect. Mean X direction\_pct seems to be better at separating the 2 classes. Suggest testing this 2 feature combinations, or using Mean, Median, direction\_pct alone in fitting the models.

**[Model fitting–Decision tree]**

1. **Decision tree:** performance shown below



**Insight**:

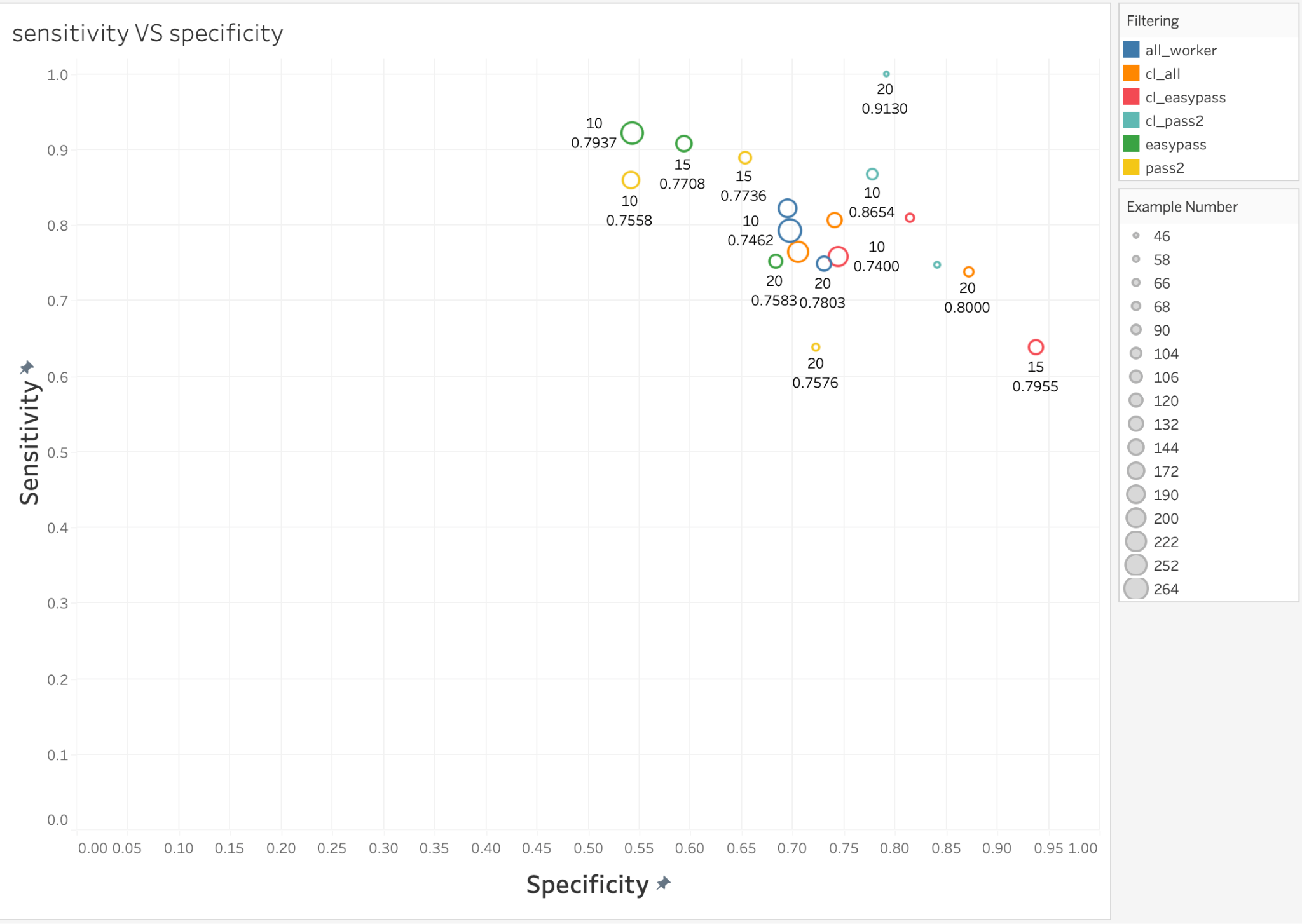
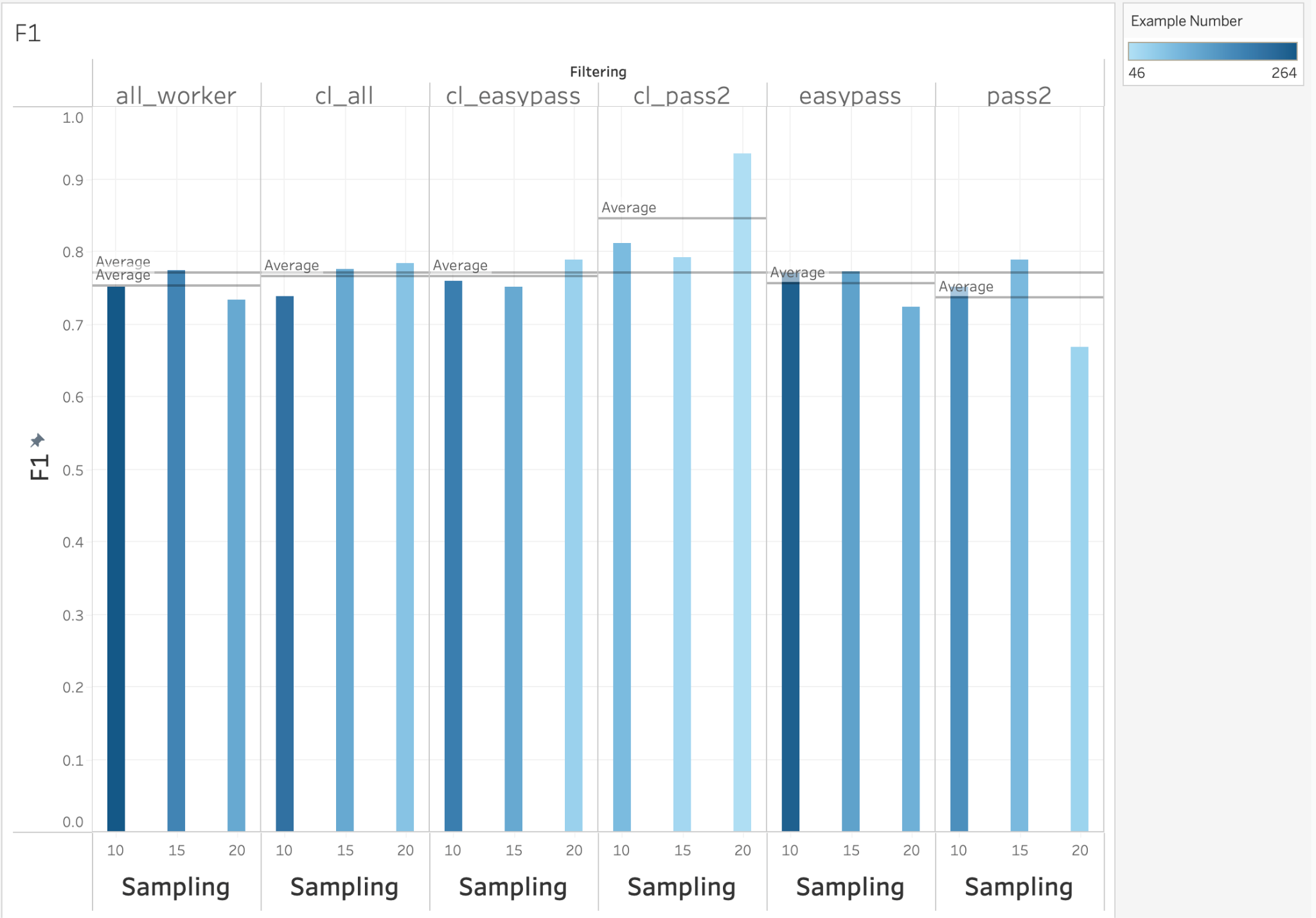
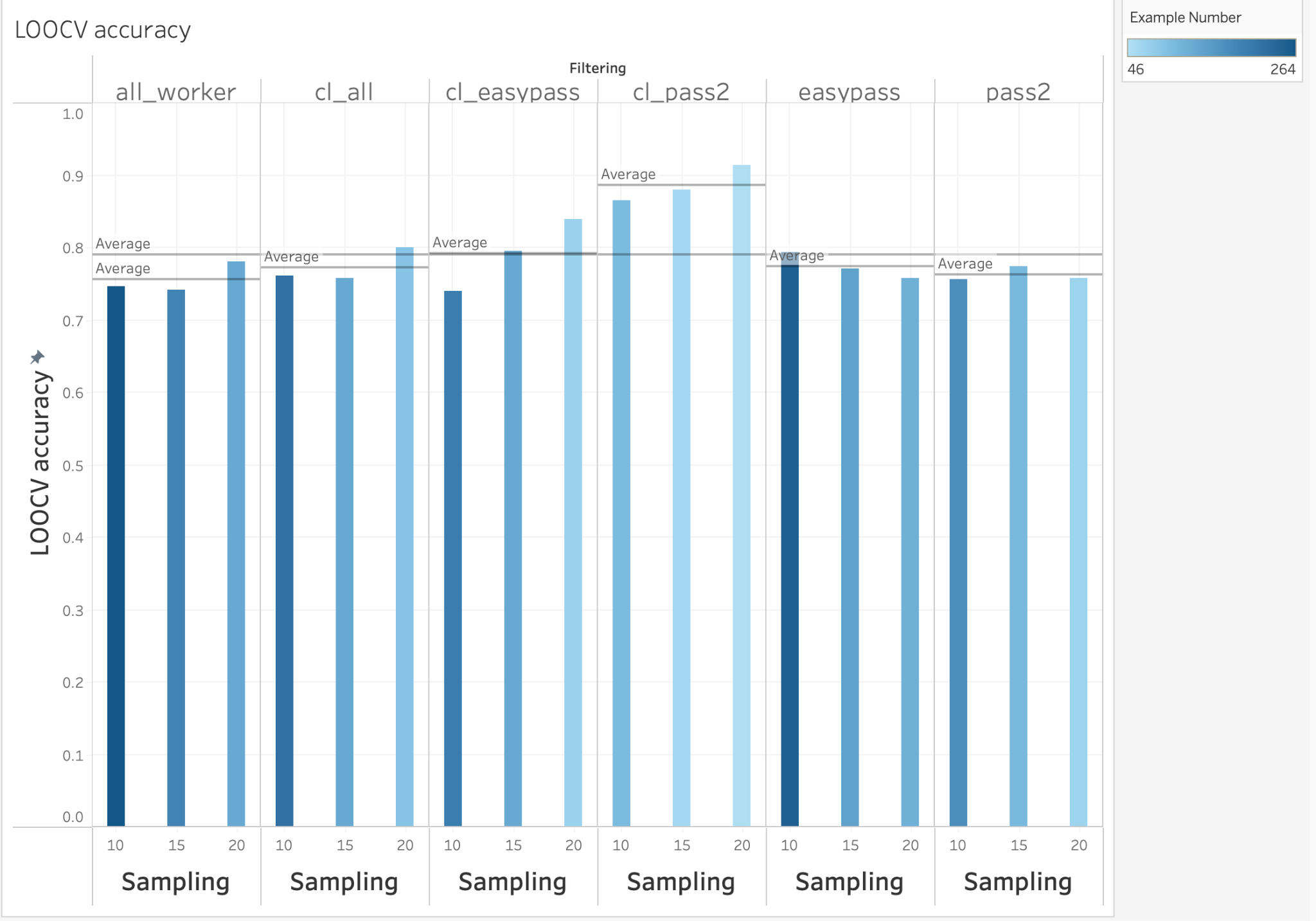
* Models built on workers that were filtered based on “clustering (cl)” and “passing both negative and positive attention checks (pass2)” seems to have better performance in distinguishing hard and easy questions.
* Number of workers does not have a consistent trend, but it seems like sampling 20 workers gives the most accurate results.
* The most useful features are Mean, Median, and directionality percentage (especially for models built after filtering workers based on clustering)

**Pitfall**:

* Small data set: we don’t have the same number of examples (data points) in each model, and in general 46-264 examples are a very small dataset, the high performance in some models could happen by chance (too little test data points, 30% of the total examples) → next step TODO: k-fold cross validation and leave-one-out cross validation
* In reality it is hard if we decide to use the “clustering” filter method.This is because the clustering filter method is a Euclidean distance based on clustering, if we want to have 20 click workers to be selected after clustering, we need to recruit more than 20 click workers, and not sure exactly how many, as the clustering method does not guarantee this input output ratio.

**[Model fitting–Decision tree (k-fold cross validation and leave-one-out cross validation)]**

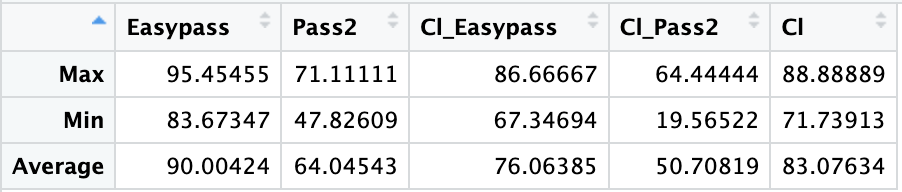
1. I applied both k-fold cross validation(k = 3), to calculate average accuracy, average F1 score, average sensitivity and specificity. I also applied leave-one-out cross validation (LOOCV) to calculate average accuracy to be more precise. Positive class is “hard” questions.



**Insight**:

* In consideration of achieving high accuracy and F1 score while taking into consideration the number of data points in training the model (models with low number of examples need to be taken with a grain of salt), I would pick “cl\_pass2 + 10 click workers used”, or “cl\_easypass + 15 click workers used”, or “easypass + 10 click workers used”
* However, I know that I am okay with having some hard questions (2 cows are the same) labeled as easy (there is a difference between the 2 cows), and would rather have most easy questions labeled out as long as there is any lameness differences between the 2 cows. Therefore, as hard questions are positive class, I want high specificity, and can have low sensitivity. Then I would select “cl\_easypass + 15 click workers used”, second choice “cl\_pass2 + 10 click workers used”
* TODO: create a flow chart to say why I am doing each step
* TODO: redefine what is the ground truth

**[Calculating the retention rates for each filtering methods]**

1. I found out the number of click workers by making a list of the number of columns in every data frame for each filtering method (since each column represents a click worker).
2. Then I calculated the retention rate of each task by dividing the number of click workers before any filtering with the remaining number of click workers after filtering. I did this for every task for each of the 5 filtering methods and compiled them into a single data frame.
3. Afterwards I found the retention rates across all the tasks by making a dataframe that shows the minimum, maximum, and average retention rate for each of the 5 filtering methods. 

**TASK**

Sort out the videos to come up with an artificial group of 30 cows, with good videos.

| **GS** | 4 | 3 | 2 |
| --- | --- | --- | --- |
| **Cow number** | 4 | 8 | 18 |

Upload the videos to Amazon Storage 3. Video only plays on Safari and cell phone, not on Chrome or Edge. It’s because the videos are too big.

Compressed all videos and reuploaded to Amazon Storage 3, got a list of URL for each video that is publicly accessible

Generate a dataframe to compare every cow with every cow

Generate the HTML. 10 questions each, 1 positive attention check, 1 negative attention check.

**Insight**: Some cow videos that were scored as 4, I do not think they look like a 4.

**TODO**: Wali need to watch those 30 videos and redo the gait scoring by watching those 30 videos.

**QUESTION**

How do we defeat bots? Are Amazon click workers still reliable? Should we switch to Prolific with the combination of using Qualtrics?

**METHODS**

Talked with Lara, Erin, Katie, they are very disappointed with Amazon MTurk, reporting tons of bots. Erin is moving on to Prolific, citing papers proving that Prolific workers are more naive, and provide better quality responses. Erin reported 80% of the responses are bots

They shared multiple papers with me, read them. Previous lit reported 70-86% of bots.

I found the API for both Prolific and Qualtrics, if needed, I can move on to use their API, but it may be time consuming for me to get familiar with the new platform.

**My proposed decision**: Let’s test and see if Amazon click workers are truly horrible or still amazing! I prefer to stay in Amazon if possible, it would save me more time that would be otherwise used in getting familiar with a new platform. I can launch a test HIT on Amazon, details about this HIT below:

# click workers: 100

Payment per HIT: $1.2

Time limit: 30 min

Click worker qualification: >= 90% approval rate (have done 10 other tasks) with master qualification

Estimated total cost: $174

# video pair questions: 10

Collect worker IP: Yes

Bot detection:include a text entry box only visible to bots but not humans, if detect anything, stop them from proceeding

HIT website design:

[Page 1] Purpose statement, approval/rejection criteria statement, HIT description & instruction

“Thank you so much for working on this task! We are a team of scientists from an animal welfare research lab, dedicated to improving the welfare of dairy cows. Your work is highly valuable in helping us identify a condition called lameness in dairy cows. The cows and us would appreciate your attention while completing this task!

**Work approval and rejection creteria**: This task must be completed on a laptop, PC, desktop computer, or a tablet, you can not proceed to complete this task if you are not working on one of these devices. This task includes attention checks to ensure quality. Failure to pass these checks will result in rejection of your work. If you complete the task with high quality and accuracy, you will be granted qualification to future tasks of a similar nature.

**Instructions**: Please use the "Next Question" and "Previous Question" buttons to navigate between questions. You would only be able to submit this HIT after all 13 questions are answered. In each question you will be asked to watch a pair of videos, showing two different cows walking at the same time. Your task is simply to indicate which of the two cows is more lame. Click below to indicate which cow is more lame, and by how much. Example: if the Left Cow is much more lame then click the option to the far left of the colored buttons. If the two cows are about equally lame (or not lame) then click the option in the middle.”

[Page 2-11] video pair questions

this HIT will have 12 test questions (video pairs) **Randomize the sequence**

| Video pair type | Compare GS 2 with GS 4 (difference = 2) | Compare GS 2 with GS 3 (difference = 1) | Compare GS 3 with GS 4 (difference = 1) | Compare GS 2 with GS 2+ (2.49)  (difference = 0) | Compare GS 3- with GS 3 (2.51) (difference = 0) | Negative attention check (same cow video) |
| --- | --- | --- | --- | --- | --- | --- |
| #questions | 3 | 3 | 2 | 2 | 1 | 1 |

[page 12] contact info

1. “You are welcome to contact us if you have any feedback and suggestions about this HIT. You are also encouraged to email us if your work is rejected but you paid close attention when completing this HIT and want to be paid. Email: skysheng0306@gmail.com”

After all responses were submitted, I need to approve or reject their work:

Rejection criteria: they will be rejected if meeting 1 of the conditions below:

If they failed the 1 negative attention check

If they are bots (entered text into the hidden text intry box)

If they clicked the same answer for all the 10 video pairs

Check for duplicated IP with duplicated answer, reject duplicated work

TODO: review the completion time, see if any speeders exist

Selection criteria:

Accuracy >= 70%

If they passed the 1 positive attention (GS differences between the 2 cows are 2)

**TASK**

AWP internal test

**METHODS**

Send out HIT0 (12 questions), to all AWP members, ask them to send me the coded response

**RESULTS & DISCUSSION**

11 people finished the HIT

Feedbacks and advice:

**[1] Katie:** Check grammar: The cows and us would appreciate your attention while completing this task!

-The question page (the first one) has some overlapping/odd formatting with text (the scales)?

-Do you need a submit and next question button on each page? I clicked next but wasn't sure? Maybe just on the final page?

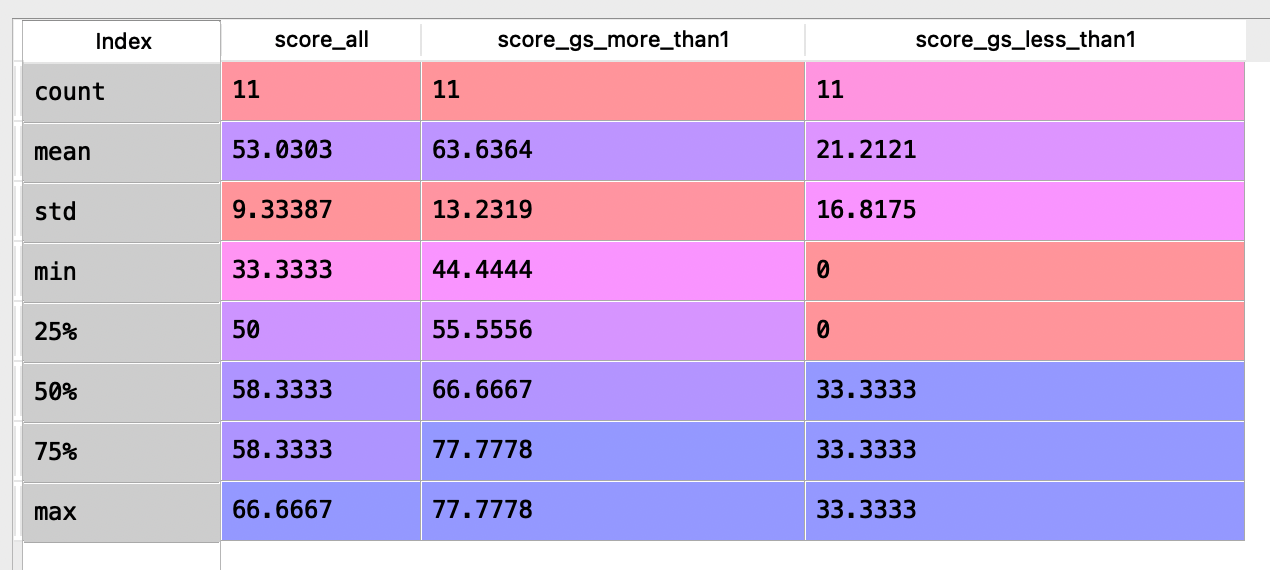
-Also, I was unable to select the entire text at the end. It was selecting some text, but then it was disappear? I think it did this when I had to scroll up. I am not sure if this is just me, but I tried a few different times and it wasn't working super well. I had to do it in chunks!

**[2] Emeline:** A couple questions I found a little more difficult (sorry I forgot to take notes which ones tough…), but I think it would have helped if both video were closer together. Maybe that depends on the size of your screen too?

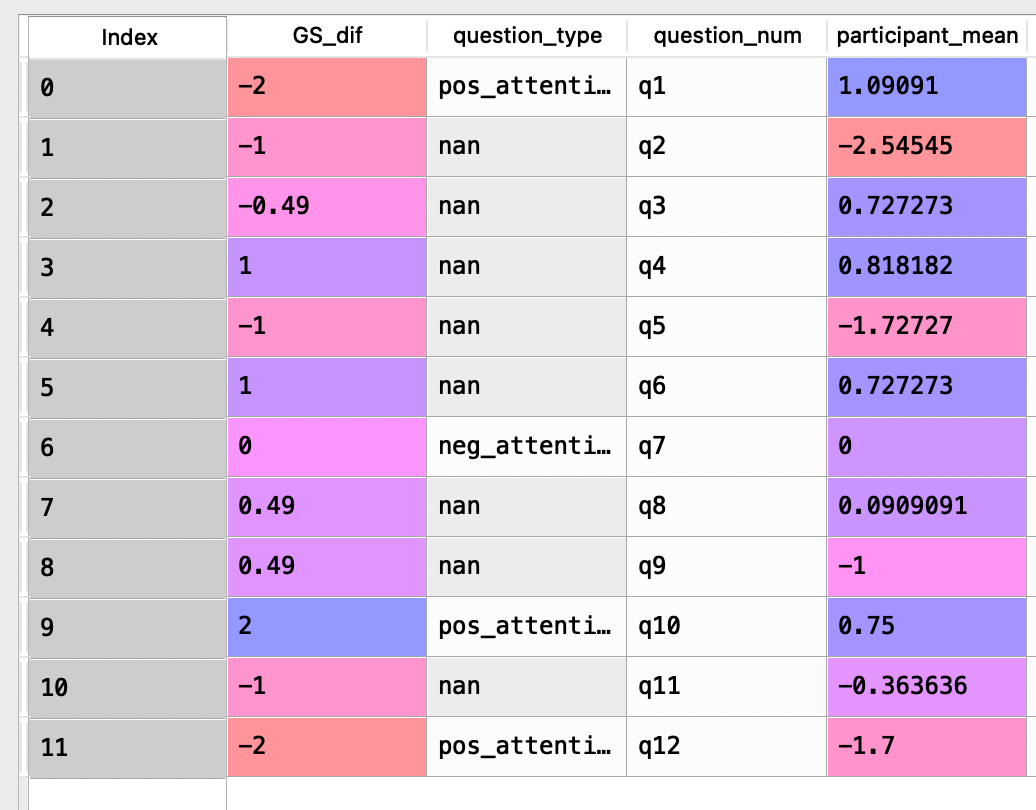
**[3] Lara**: I actually had a little bit of a hard time viewing the 2 videos side by side…maybe I need to do it on a smaller screen or I am just bad at watching 2 things at once, but I felt like it was hard to watch both at once and I found myself tending to watch first one video and then the other and then being confused if I was over-analyzing differences. Did you do any pilot tests (or know of any literature) on effects of having the videos horizontally side by side versus vertically top and bottom, or any side biases, ie individuals more likely to see cow on left as more lame or vice versa?

Observation from 11 people’s responses:

* IP collection failed for 2 people (maybe their browser set up blocked IP collection?)
* Emeline was detected as a bot
* 11 out of 11 passed negative attention check
* 8 out of 10 passed positive attention check
* For all the other potential positive attention checks (GS difference = 2)
  + 1 out of 11 passed q1
* 6 out of 8 passed q10
* The average, min and max of each individual’s respons score:



* The comparison between average response from participants and the GS\_difference golden answer:



* From my observation, we get to the same place as our conclusion in phase 1 amazon project: again they are not excellent individually, but they are correct together (averaging all 11 people’s response)

**TASK**

Lameness experts do all 54 HITs on Amazon MTurk’s sandbox. I share my password and account with them

**METHODS**

Launch all 54 HITs on Amazon MTurk, record the list of HIT address in a csv

**RESULTS & DISCUSSION**

**QUESTION**

How to validate or assess the accuracy of the lameness rank

**METHODS**

* Rank all the cows for lameness from the worst to the best, using gradient color and create color zones, top \*\* color zones are cows that are lame enough needs to be checked
* Generate the rank multiple times, if 18/20 ranks agree with each other, then it’s reliable