Thanks for the reviews and helpful comments, we appreciate them and addressed them as specified in the table below.

We would like to mention that some of the information is challenging to answer, since this is the first experiment of this kind; which is why we are missing any theory or other previous work in which we could build. This has also been acknowledged by Reviewer 1: “The authors have identified a significant gap in the literature.”

We hope that you perceive our refinements of the Registered Report as feasible.

In the pdf, we have highlighted changes in blue.

1. Justification of N=30 for the survey study Given the importance of the survey study for choosing the input parameters for the MAIT and MPIT interventions, the precision of these estimates in the survey study seems crucial. At the moment, the sample size justification for this part of the design is defined too imprecisely and arbitrarily. Instead, please provide a formal sampling plan based on the required level of precision (for guidance see the section “Planning for Precision” in https://psyarxiv.com/9d3yf/). This could be achieved analytically or through simulations.	Thank you for this valuable comment and the useful reference. Since we can not perform any inferential statistical analysis with the survey data (the MAIT and MPIT are selected based on their modal values only), we can not plan our sample size on the basis of effect sizes and the corresponding confidence intervals. We have now clarified in the survey how we obtain the values for MAIT and MPIT. Furthermore, we have explained in the report that our sample size is restricted due to the fact that we are recruiting a very specialized sample with a small population. Furthermore, as our study is one of the first in this respect, we have no literature from which we can derive effect sizes. Or in other words, our focus is much more on a detailed description and definition of the characteristics of our special samples and about the likelihood for our categories.
2. Sub-samples within the survey study On p3 you note: “In addition, we will distribute a second version (to distinguish both populations) of our survey through our social media networks.” How will this be taken into account in generating the parameter estimates? Will the different samples be distinguished or collapsed to produce the payoff functions?	clarified in the manuscript (we will analyze both groups for differences in the variables of interest. If there are no differences, we will collapse both samples. In case there are differences between the samples, we will only use the data of our personal contacts.
3. Clarification of the statistical sampling plans for the experiment. The are two issues to address in relation to the sampling plans.	Thank you for your valuable input. We have added a power analysis for a range of 20 - 35 participants in the report.

• You plan on recruiting 20 participants per group but also reserve the option to collect additional participants. In order to control the Type I error rate, standard power analysis requires a fixed stopping rule, which in turn requires committing to a specific sample size. If you want to employ a flexible stopping rule then you will need to implement a sequential design that involves regular inspection of the data between the minimum and maximum N, with the error rate corrected en route (see https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2023). • At present the only reference to statistical power is in the design table: "Furthermore, we will conduct an a posteriori power analysis to reason on the power of our tests." Power is a pre-experimental concept, and post hoc power analysis (or "observed power") is inferentially meaningless because it simply reflects the outcomes. A formal prospective power analysis is required, to either define the sample size required to detect a smallest effect size of interest (*a priori* power analysis), or to define the smallest effect that can be detected given a maximum resource limit (so-called *sensitivity* power analysis). At present, given N=20 per group, and a strictest Holm-Bonferroni correct alpha of .0083 for the lowest ranked p-value (assuming you apply the H-B correction for 6 tests across both hypotheses), your design has 90% power to detect d = 1.3. Any d > 1 (i.e 1 standard deviation) is in the conventionally-defined "large" range. Unless you would be happy to miss an effect smaller than d=1.3, the sample size needs to be substantially increased. To progress I would suggest the following: (1) try to establish what the smallest effect size of interest is for H1 and H2, either based on theory, or the smallest practical benefit of your intervention in an applied setting, or based on prior software-engineering experiments; then justify the rationale for this smallest effect of interest in the manuscript. (2) if you have no upper resource limit on sample size then perform an *a priori* power analysis to determine the sample size necessary to correctly reject H0 for this effect size with no less than 90% power. If you *do* have an upper resource limit on sample size (which is very reasonable) then instead perform	

a sensitivity power analysis (see section 3.1.2 here if using G*Power) to determine what effect size you have 90% power to reject at your maximum feasible sample size, and then justify why this effect size is sufficiently small for your experiment to provide a sufficiently sensitive test of your hypotheses (H1 and H2).	
4. Clarification of which specific outcomes will confirm or disconfirm the hypotheses. **For H1:** In the design table you state: "We find support for H1, if our participants' performance in NPIT is worse AND if the tests between any of our experimental treatments are significant with $p < 0.05$ (after correcting with the Holm-Bonferroni method)." Does "worse" here refer to each of the pairwise comparisons, or does it mean that NPIT must be numerically worse than *all* of (or the average of) the other conditions (OSIT, MAIT and MPIT)? If I understand correctly, the second part of your specification means that H1 is supported if *any* of the following contrasts is statistically significant: (NPIT < MPIT) OR (NPIT < MAIT) OR <NPIT < OSIT). If so, I suggest making this crystal clear by adding italics and including these in the "interpretation" cell of the table: "We find support for H1 if our participants ' performance in NPIT is significantly lower than in *any* one of our experimental treatments at $p<.05$ (after correcting with the Holm-Bonferroni method): (NPIT < MPIT) OR (NPIT < MAIT) OR <NPIT < OSIT)". **For H2:** If I understand correctly, any significant difference in any direction between OSIT, MAIT and MPIT would be considered support for H2. So H2 is supported if: (MPIT < > MAIT) OR (MAIT < > OSIT) OR (OSIT < > MPIT). If so, please make this clear in the interpretation cell of the design table.	Thank you for this precise hint, we refined our descriptions accordingly!
5. Definition of the F1-score. Please provide a precise explanation and definition of the F1-score (including a worked example of how it is calculated), and make clear that it is the *only* outcome measure that will be used to evaluate H1 and H2.	We clarified the F!-score measure as the only measure in our experiment.
6. Clarification of exclusion criteria. On p7: "We do not plan to remove any outliers or data unless we identify a specific reason for which we believe the data would be invalid." For a Registered Report, the precise rules for excluding data must be exhaustively specified, both within participants and also at the level	Thanks for pointing out this issue, we aimed to clarify in the report.

of participants themselves. Where participants are excluded, make clear that they will be replaced to ensure that the target sample size is reached.	
7. Eye-tracking acquisition and analyses Please provide additional details on preprocessing (e.g. filtering, smoothing) of eye-tracking data to ensure that the procedures are fully reproducible. Presumably eye-tracking analyses are reserved for exploratory analyses (with no prospective hypotheses) and will therefore be reported in the "Exploratory outcomes" section of the Results at Stage 2. If so, please note this explicitly in the revised manuscript. Alternatively, if you have specific hypotheses for the effect of incentivization on the eye-tracking measures, ensure that they are fully elaborated in the main text and study design table.	We specified more clearly what software/hardware and version we are using, also that this is part of the exploratory analysis only.
8. Robustness analyses On p7 you state: "Though the share of participants who will use eye trackers will be constant among all treatments, and thus should not affect treatment effects, we will further check whether the presence of eye trackers affected performance. To increase the statistical robustness, we will also conduct a regression analysis using the treatments as categorical variables and NPIT as base. As exogenous variables, we include: age, gender, experience, and arousal of the participants." Make clear that these are exploratory analyses.	As for point 7.
9. Other points p7: "We will first check whether the assumptions required for parametric tests are fulfilled, and if not proceed with non-parametric tests." Make clear which assumptions (e.g. normality) you are going to test for, and how, and then specify the alternative tests that will be used (e.g. presumably Mann Whitney U test?) p7: "For the significance analyses, we will apply a confidence interval of $p < 0.05$ and correct for multiple hypotheses testing using the Holm-Bonferroni method." Do you mean "alpha level of .05" instead of "confidence interval of $p < 0.05$"?	Thanks for highlighting these unclear statements, we adapted them accordingly.
Reviewer 1 (anonymous reviewer)	
The authors have identified a significant gap in the literature. Current studies, in general, do not consider the impact of financial incentives in affecting behaviour and performance developing software.	

It should be explicitly stated how they plan to mitigate the thread to validity of having colleagues perform code reviews.	We have provided some more details on our sample in the report to clarify this point. We are not planning to have colleagues perform the code reviews.
In general, the report is very well written. One thing I would change is “Seemingly, this resulted” to “This has resulted” in the abstract though.	Fixed
Given that the experiment will be conducted in a controlled laboratory setting, the authors should state what threats this could present in terms of the results being transferred to industry and how such threats could be mitigated. Cost functions are discussed solely in terms of motivating participants. The authors could add a discussion on the different types of organisational objective functions that may be at play in an industrial setting. Such as, the organisational culture and the degree to which code quality is important to the software being the developed and the extent the organisation would be willing to compensate employees in this manner.	Thank you for these interesting ideas. We have now clarified in the report how we are planning to mitigate threats to validity posed by variables like organizational culture. We will recruit people from different companies and control for various external factors (e.g. industry, management practices).
The authors should state the sample size or the number of people that will partake in the study to justify the potential statistical results.	Clarified
Reviewer 2 (*Edson OliveiraJr)*	
This RR presents a two-package study on how financial-incentivization might impact in code review. The first study is a survey with practitioners in which researchers will observe the most applied and the preferred payoff methods. From this survey, they will define a set of such methods (4 a priori) to conduct a controlled experiment with students and, potentially, practitioners. In such an experiment, the researchers will analyze how different payoff schemes impact the performance of software developers during code review. This is a relevant research topic. The protocols of the studies are generally well-designed and explained. However, I have some points to discuss towards improving such studies:	
* Survey: - a major concern is on the open vs. non-open source projects. Literature clearly emphasizes they have different motivations and activities from general software engineering projects. I didn´t see any	We aim to mimic the motivations of OSS developers using incentives, building on research in experimental economics. We clarified this in the report.

discussion on these potential threats in the survey protocol. How do you can extrapolate such threats as you will provide a set of payoff methods to the controlled experiment, which will be performed with students and, potentially, practitioners (non-open projects)?	
- Why do exactly you expect at least 30 participants? Is this because of the probability's Central Limit Theorem? If so, please make this explicit.	The primary reason for aiming for 30 participants in our survey is the availability/accessibility of corresponding experts. We explain the sampling in more detail.
- What is the minimum experience time expected for the participant's profiles?	In the survey, we are excluding participants who state that they do not have any experience at all. We have now added this to the report.
- I'd suggest to run the instrument evaluation tests with practitioners rather than students, as students are not the target audience of the survey.	We clarified that the survey is not intended for students, but practitioners (i.e., personal contacts in companies).
- As you will use the mean value for the weights, how will outliers or extreme values be treated?	Thank you for this remark. We have now included in the report how we are planning to analyze the outliers.
- I'd suggest providing a complete feedback on results for participants at the end of the study, as a way to motivate them to take other surveys.	We are not certain that this would add much value, since the participants will mostly enter preferred ratios of incentives. A summary could only show the averaged opinions of all participants, but without a detailed analysis this would not add much value.
- It is not clear to me, if the participants may choose more than one payoff method in the survey questionnaire. If so, 30 participants seem ok, otherwise, the sample size should be considerably larger.	The participants can select multiple payment components (i.e., incentives) that they apply or prefer. We clarified this in the report.
* Lab Experiment - I'd like to see clearly the declaration of independent and dependent variables in the "Metrics" section. This is essential for readers to understand the chosen Experimental Design.	We clarified, thanks for pointing out that we did not have this yet.and
- In the "Experimental Design" section, please provide the design chosen in terms of factors and treatments, for example, 2x2, 1xn, etc..	Added, thanks for pointing out that this was missing.

- in the "Inferential Statistics" I'd suggest running an effect size test to provide the strength of the p-value over the null hypothesis results.	Yes, we added this.
All in all, the RR is well-written and easy to follow. Congrats on it and success on the studies.	Thank you.