Dear editor, Dear Dr. Feeney,

We would like to thank you for the opportunity to respond to the valuable comments raised by The BMJ’s manuscript committee and the reviewers and to revise our manuscript entitled ‘Accuracy of COVID-19 self-tests with unsupervised nasal or nasal plus oropharyngeal self-sampling in symptomatic individuals in the Omicron period’ accordingly.

Please find below a point-by-point response to the comments, and a description of how and where the responses have been incorporated in the manuscript. The comments have tremendously helped to further improve the manuscript. For your convenience we submitted a clean and track changes version of the manuscript. Page and line numbers listed refer to the version with tracked changes.

We hope these revisions meet your expectations and very much look forward to your positive judgment.

Yours sincerely,

Karel Moons, corresponding author

Karel GM Moons,
Professor of Clinical Epidemiology
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Utrecht University
Universiteitsweg 100
3584 CG Utrecht
The Netherlands
E-mail: k.g.m.moons@umcutrecht.nl
Comments put forward by the manuscript committee:

- The text says, “the analysis only included those with symptoms at time of sampling” but only 60% said they were testing due to symptoms. This seems inconsistent. Can the authors clarify this?

Response: We understand this may seem contradictory. All participants who had symptoms at the actual time of visiting the COVID-19 test site and undergoing the RT-PCR were eligible for participation. This, however, does not necessarily mean that they also had symptoms at the time of scheduling the test or that they scheduled the test because of symptoms. Their initial reason for testing may have been having had close contact to an infected individual, or because they had a positive self-test at home and needed confirmation. Prompted by the comment of the reviewer we have clarified this now further on page 7, line 173 (newly added text in italics): “Current analyses only include individuals who reported any SARS-CoV-2 infection-related symptom at the actual time of sampling, regardless of the reason for visiting the test site.”

- Why did the authors have people go home and take the test instead of in the office?

Response: The explicit aim was to assess the diagnostic accuracy of the self-tests as used by individuals in their own environment, in the absence of any professional test site personnel, i.e., unsupervised testing. We explicitly aimed to optimally mimic the situation in which the self-tests are being used. Test site personnel may inadvertently influence the way participants perform and interpret the test.

- Can the authors comment on why sensitivity and specificity was different in different groups even though in theory they should be invariant to disease prevalence?

Response: It is not entirely clear to us what the committee refers to with “different groups”, but we assume this relates to the different strata that we used in our analyses. Indeed, the sensitivity and specificity of the self-tests differed across these subgroups, as did the prevalence. However, that does not mean that the difference in prevalence is the cause of the difference in sensitivities and specificities. We did not want to leave this impression, and have therefore rephrased the text where necessary. Disease prevalence typically differs across subgroups for every disease (not only COVID-19), which is often due to underlying characteristics of these subgroups, and these underlying characteristic may affect test performance. For example, confirmatory testing and a previous SARS-CoV-2 infection clearly affected the sensitivity and specificity of the Ag-RDTs. Since individuals in these two subgroups are also not equally distributed across age and gender groups, the diagnostic performance in age and gender subgroups may be affected as well. We further hypothesize that diagnostic test performance in the epidemic setting mostly depends on SARS-CoV-2 viral load in the body area that is being sampled and on the quality of the sample. Our study provided direct empirical evidence for the former as using a viral load cut-off greatly improved sensitivity. Some of the subgroups that we evaluated may have had lower viral loads on average. For example, the immune responses mounted by individuals who have had COVID-19 in the past or have been vaccinated may inhibit the virus from replicating. And indeed, our data showed lower sensitivities in these groups. Parts of this response have now been added to the discussion section on page 14, lines 410-419: “Differences in diagnostic performances across subgroups may be explained by differences in the underlying characteristics of these subgroups. For example, confirmatory testing and a previous SARS-CoV-2 infection clearly affected the diagnostic performance. Since individuals in these two subgroups are not equally distributed across age and gender groups, the diagnostic performance in age and gender subgroups may be affected as well. We further hypothesize that diagnostic test performance in the epidemic setting mostly depends on SARS-CoV-2 viral load in the body area that is being sampled and on the quality of the sample. Our study provided direct evidence for the former as using a viral load cut-off greatly improved sensitivity. Some of the subgroups that we evaluated may have had lower viral loads on average. For example, the immune responses mounted by individuals who have had COVID-19 in the past or have been vaccinated may inhibit the virus from replicating. And indeed, our empirical data showed lower sensitivities in these groups.”

- We are interested to know more about the much higher sensitivity (~15%) in the confirmatory testers. Is it mainly attributed to the generally higher viral load, or the proficiency of the testers? If it’s the latter, are these assays ready for self application in general population?

Response: In the Netherlands, all available SARS-CoV-2 self-tests are lateral flow antigen tests. Numerous previous studies, including our own studies, have shown that antigen tests require a higher viral load to...
become positive than molecular tests such as RT-PCR. This was confirmed in the current study: the mean viral load in confirmatory testers was indeed higher than in the non-confirmatory testers. Confirmatory testers also had more self-testing experience than the non-confirmatory testers (having performed a self-test more than 10 times was reported by 37.2% vs. 30.0% in the Flowflex group, 42.0% vs. 25.7% in the MPBio group, and 22.5% vs. 19.6% in the Clinitest group, respectively). Furthermore, if self-testing experience were to impact sensitivity, one would expect a higher sensitivity in those who had performed more than 10 self-tests in the past vs. those who only performed 1-3 self-tests. However, in non-confirmatory testers, we found the opposite (51.3% vs. 73.1% for Flowflex, 47.9% vs. 54.0% for MPBio, and 41.4% vs. 55.9% for Clinitest, respectively). These data suggest that inexperienced individuals are similarly capable of performing these tests unsupervised at home than experienced individuals. A summary of this response had been added to the discussion on pages 13-14, lines 380-392: “In the Netherlands, all available SARS-CoV-2 self-tests are lateral flow antigen tests. Previous studies, including our own studies, have shown that antigen tests require a higher viral load to become positive than molecular tests such as RT-PCR. This was confirmed in the current study: the mean viral load in confirmatory testers was indeed higher than in the non-confirmatory testers. In a post-hoc analysis, we assessed the impact of self-testing frequency. Confirmatory testers did have more self-testing experience than non-confirmatory testers (having performed a self-test more than ten times was reported by 37.2% vs. 30.0% of the participants in the Flowflex group, 42.0% vs. 25.7% in the MPBio group, and 22.5% vs. 19.6% in the Clinitest group, respectively). However, if testing experience were to impact sensitivity, one would expect a higher sensitivity in those who had performed more than ten self-tests in the past vs. those who only performed 1-3 self-tests. However, in non-confirmatory testers, we found the opposite (51.3% vs. 73.1% for Flowflex, 47.9% vs. 54.0% for MPBio, and 41.4% vs. 55.9% for Clinitest, respectively). These data suggest that inexperienced individuals are similarly capable of performing these tests unsupervised at home than experienced individuals.”

-Is this a select group of patients? It seems those included here might have done more testing in the past and be more likely to obtain an adequate sample for testing. Is this still making these tests look better than they should?

Response: Please see our response to the previous comment.

-PLEASE make sure the manuscript reports the exact time window of the study--to us it is unclear.

Response: The explicit study dates per self-test evaluation are presented in Table 1, and are now also referred to in the text explicitly at the end of the “Study design and population” paragraph (page 8, lines 178-179): “Inclusion dates per test location and phase are provided in Table 1.”

-Can the authors include a box with information about where these tests are used and how commonly used they are?

Response: The Flowflex, MPBio and Clinitest tests have all been freely distributed by the Dutch Ministry of Health, Welfare, and Sport across various target audiences. Primary schools, secondary schools, universities, institutions caring for vulnerable people, and organizations who aid civilians who cannot afford to buy tests were amongst those receiving tests from the Ministry for distribution to their constituents. From April 2021 onwards, the Ministry distributed almost 120 million Ag-RDTs for self-use, of which 10.6 million were Flowflex, 28.7 million MPBio, and 12.4 million Clinitest. In addition, the Ministry still has over 11.7 million Flowflex, 14 million MPBio, and 7.1 million Clinitest tests in stock. A box with the above information has now been added to the manuscript after the introduction on top of page 7.

-PLEASE share how you will disseminate or share the evidence in addition to publication, blogs, specific groups, social media, opinions written with an end-user etc. This is mandatory for The BMJ, and should be placed in the endmatter.

Response: Thanks for this request. The following statement has been added to end of the manuscript (page 25, lines 621-627): “Dissemination to participants and related patient and public communities: The Dutch Outbreak Management Team that provides guidance to the Ministry of Health, Welfare, and Sport on COVID-19 policy have advised, based on the results of this study, that rapid antigen tests can be used in the home setting for detection of a SARS-CoV-2 infection in symptomatic individuals, and that confirmation by a RT-PCR test at a test site is no longer necessary. As such, the results of our study have directly been
disseminated and are currently incorporated in the nationwide testing policy. At the time, this change in testing policy was covered by multiple national news outlets."

-Do the authors think performing a self-test after having just been tested by healthcare workers would have increased the quality of self-testing since the user would have an example to go off of?

**Response:** The interval between the RT-PCR and self-test had to be short to avoid viral load increases in the interim. The self-test result is available to the participant immediately whereas the RT-PCR test result is only available after several hours (more than 3 hours) or the next day. The only viable solution therefore was to do the RT-PCR at the test-site first, followed by self-sampling in the home environment within 3 hours. We believe that the alternative option – self-test followed by RT-PCR – likely would have resulted in individuals changing their behavior based on the self-test result, including differential loss to follow-up. At the time of the study, the Dutch test sites only performed RT-PCR testing. These samples are taken deeper in the nose and throat. Participants were made aware of the fact that the sampling method for the self-test differs from the RT-PCR test, and that they should explicitly follow the instructions that they received with the self-test.

**Reviewer: 1**

**Comments:**
The study performed by E Schuit et al. evaluate whether accuracies of Ag-RDTs with nasal self-sampling changed over time and to quantify whether addition of OP to nasal self-test sampling provide better sensitivities Authors used 3 different rapid tests, Flowflex, MPbio and Siemens Clinitest and were compared to the RT-PCR as reference standard test using the Cobas SARS-CoV-2 test on either Cobas 6800 or 8800 platform. They used a viral load cut-off as a proxy of infectiousness of >= 5.2 log10 SARS-CoV-2 E gene-copies/mL, based on a previous study performed by the same team and published in the BMJ, which indicate that under this level viral culture did not grow and they conclude that under this level the patient is not contagious.

**Main remarks:**
- Flowlex is a rapid test commercialized by Acon, a company based in San Diego, California, but linked with Chinese manufacturers. Siemens Clinitest is the not alone distributor of the chinese Orient Gene test, which is available on other brands all around Europe. Acon and Siemens are only distributors, as SD Biosensor is a South Korean test and Roche did not manufacture the test they sell.

**Response:** We thank the reviewer for providing this information. We now added more detailed descriptions of the three tests accordingly and based on the information provided in their instructions for use in our supplementary material. We changed the description of the Flowflex (page 10) from “The Acon Labs Flowflex COVID-19 Antigen Home Test is a CE-marked Ag-RDT” to “The Flowflex COVID-19 Antigen Home Test is a CE-marked Ag-RDT manufactured and distributed by Acon Laboratories, Inc. San Diego, CA, USA.” The description of the MP Bio tests (page 10) was changed from “The MP Biomedicals Rapid SARS-CoV-2 Antigen Test Card is a CE-marked Ag-RDT, originally manufactured by Xiamen Boson Biotech Co., Ltd., Xiamen, China and distributed by MP Biomedicals Germany GmbH, Eschwege, Germany.” The description of the Clinitest (page 10) was changed from “The Siemens-Healthineers CLINITEST Rapid COVID-19 Antigen Test is a CE-marked Ag-RDT” to “The CLINITEST Rapid COVID-19 Antigen Test is a CE-marked Ag-RDT, manufactured by Healgen Scientific Limited Liability Company, Houston, TX, USA, and distributed by Siemens Healthineers, Erlangen, Germany.”

- The viral cut-off chosen do not meet the MDCG (Medical Device Coordination Group) which have defined the quality criteria of a rapid test based on Ct level. The viral load was determined according to standard curves to transform the Ct in quantified viral loads. In that case, before each PCR run a standardized quantified sample of 5.2 log10 copies/mL has to be done at each PCR series to be sure about the concordance of the CT level given by the RT-PCR corresponding to the quantified Viral load. It is not mentioned that it has been done in each center in the study.
**Response:** This study was conducted at Dutch Public Health test sites during the Dutch epidemic and we therefore had to follow routine test site procedures. These procedures are described in detail in the supplement, part 3. Determining viral loads was at the time of the study unfortunately not part of the routine testing procedures and we could therefore only make estimates based on a previous study that used standard curves (this study was cited in the supplement, part 3). Prompted by this comment of the reviewer, and to further clarify this, we added additional information to the statistical analysis section on page 9, lines 230-231 of the manuscript as follows: “The viral load of each sample was estimated from the Ct value of that sample using formulas based on the results of a previous study (supplement 3).” We have further clarified the limitations of this method in the discussion of the manuscript on page 15, lines 442-444 as follows: “The viral load calculations were based on standard curves in a previous study (reference). These standard curves were not repeated with each RT-PCR run in this study. The viral loads should therefore be considered estimates and not as gold standard viral loads.”

- When a patient initiates a COVID infection, Ct level is high during the first or second day then decrease and then increase again over a week (less or more). When the Ct is over 30 for example, on one day, it could decrease on the day after and the patient considered by this hypothesis not contagious become contagious. The principle of a positive rapid test is to isolate the patient form other people to prevent the spread of the infection. Using this cut-off has no public health implication. The patient has really a SARS-CoV-2 and need to be isolated as it is not possible to forecast what will be the evolution of the viral load in the next days

**Response:** We fully agree with the reviewer that Ct values change over time depending on the stage of the infection, and that the Ct value in itself should not impact the decision to isolate. However, many have argued that antigen tests are better predictors of infectiousness than RT-PCR tests because they turn positive at higher viral loads and do not remain positive for as long as RT-PCR tests do when a patient is recovering from COVID-19. Policymakers need information about this because RT-PCR tests are more difficult to make widely available than antigen tests, and RT-PCR tests are also not suitable for self-testing.

We therefore believe that our sensitivity analyses using the viral load cut-off do provide very useful information to policymakers, to further tailor a country’s testing policy. We have provided our own opinions on how the results of our study may be used in testing policy communication in the discussion on page 16. As was correctly pointed out by the reviewer, we also warn the reader in the discussion that a negative Ag-RDT test has a high probability of being false-negative and transmission can therefore not be ruled out.

- Bekliz et al. published a study on retrospective samples to evaluate the sensitivity of 7 rapid tests to detect delta and omicron. It was based on a limited number of samples but Abbott Panbio for example showed a decrease from 67.7% to 36.1%. In that study the Flowflex was quite good with a sensitivity of 91.2% for delta and 88.9% for omicron. In a study we have performed (not published) on SD biosensor, the sensitivity was as low as 15%

**Response:** We thank the reviewer for pointing out this study. Similar to the studies of Deerain et al. (reference 9) and Osterman et al. (reference 10) that we refer to in the introduction, Bekliz et al. assessed the analytical performance of several rapid antigen tests. Since we specifically focused on real-world performance of self-tests in our study, we did not add this reference to our manuscript.

- Using this cut-off has another limit: the specificity of the test will decrease at 90%, but recommendations for rapid test specificity need to be more than 99%

**Response:** The current requirements from the WHO are that “COVID-19 self-test kits should meet the existing World Health Organization (WHO) standards for Ag-RDTs (≥ 80% sensitivity and ≥ 97% specificity among symptomatic individuals)” (source: https://apps.who.int/iris/rest/bitstreams/1413257/retrieve). Indeed, the reviewer is correct that most of the self-test-sampling method combinations that we evaluated did not fulfill these explicit criteria: before applying the viral load cut-off the sensitivity is too low and after applying the cut-off the specificity is too low. We therefore added the following statement to the end of the first paragraph of the discussion (page 12, lines 321-323): “Only the MPBio self-test combined with OP-N sampling met the World Health Organization’s standards for Ag-RDTs (≥ 80% sensitivity and ≥ 97% specificity among symptomatic individuals).”
Minor remarks

- Participants who did not complete the RDT within 3 hours of their test-site visit were contacted with the request to perform the self-test. How many of participants delayed to perform the rapid test. What were the range of the delay? If the rapid test was performed even 24 hours later, it is difficult to compare with a virus load measured 24h before as the VL will change form one day to another

**Response:** Although the deadline was 3 hours, all participants were asked to perform the test “as soon as you get home”. We did not have sufficient data to determine the exact interval between the RT-PCR test and the antigen test. However, we were able to approximate it by assessing the time difference between data entry at the test site (always within 30 minutes of the initial arrival) and subsequent online questionnaire initiation by the participant at home (we assume that this was usually done directly after performing the Ag-RDT). The median time interval was 1.0 (IQR 0.61 to 1.8), 0.87 (IQR 0.58 to 1.7), and 0.52 (0.20 to 1.5) hours for Flowflex, MPBio, and Clinitest, respectively. Overall, 84.5% of participants completed their questionnaire within 3 hours. Positive Ag-RDT test results were reported by 42.6% of the participants who completed the questionnaire within 3 hours and 41.9% of the participants who took more than 3 hours (73.4% and 72.8%, respectively, of those whose RT-PCR came back positive). Therefore, we do not expect delays to have impacted the results of our study. Prompted by the reviewer’s comment, we added the following text to our results section (page 10, lines 261-263): “Most participants (84.5%) performed the Ag-RDT within 3 hours of visiting the test site. We found no differences in Ag-RDT results overall, nor in the RT-PCR test positive group, between participants who completed the questionnaire within 3 hours or after 3 hours.” In addition, we added the following text to the limitations section of the discussion on page 15, lines 448-451: “We did not collect detailed information on the exact timing of the RT-PCR sampling and the Ag-RDT testing. Therefore, the time interval was approximated by assessing the difference between the time a participant was registered at the test site (generally minutes after the RT-PCR sampling) and the time the online questionnaire was opened by the participant.”

- How the variant have been determined. Did mutation panel kit used for every patient and which mutation were included in the panel mutation kit. What it the Roche mutation panel kit?

**Response:** Variants were not determined for individual participants in our study but were based on nationwide COVID-19 pathogen surveillance. This source is referred to in the paper. More information on the pathogen surveillance can be found via a reference in the manuscript: https://coronadashboard.government.nl/verantwoording#varianten-van-het-coronavirus.

- Page 10: sensitivities of all three Ag-RDT was 28.6%. I think that, according to the next number it should be 82.6%

**Response:** The reviewer accidentally mistook the Omicron share for the sensitivity of the tests. The Omicron percentage in the nationwide surveillance was 28.6%: “Sensitivities of all three Ag-RDTs were highest during the first week (Figure 3) when the Omicron share was 28.6%: ‘’Sensitivities of all three Ag-RDTs were highest during the first week (Figure 3) when the Omicron share was 28.6%: ’’”.

- Finally, in the context of omicron first generation RDT sensitivities decrease and second generation ultrasensitive tests are arriving on the market (Hotgene, Fosun, Intec, Autobio, Bioperfectus...). Using this cut off and to the traditional CT levels is confusing, and the study appears to make more a promotion of some distributors to help these companies to stay on the market. The cut-off increase artificially the sensitivity of each test and do not help to select the best rapid test in the present context

**Response:** The tests included in our study were selected by the Dutch Ministry of Health, Welfare, and Sport. The Ministry requested these tests to be evaluated because they were widely available on the Dutch market and were among the most widely used self-test in the Dutch public and private sectors (e.g., schools, supermarkets, drug stores). We have now added a box to clarify this (see our response above). All organisations involved in the current evaluation of these tests are public not-for-profit organisations with no financial or any other ties to the diagnostic test companies. We have explained our reasons for including the overall results as well as results after applying a viral load cut-off above.
Reviewer: 2

Comments:
Dear Editor,

First of all, thank you for inviting me to review this interesting paper, by Schuit et al., entitled "Accuracy of COVID-19 self-tests with unsupervised nasal or nasal plus oropharyngeal self-sampling in symptomatic individuals in the Omicron period".

The authors aimed to assess the performances of rapid antigen diagnostic tests (Ag-RDTs) with nasal self-sampling, and oropharyngeal plus nasal (OP-N) self-sampling, which show useful results for researchers working in a similar field of work.

There are a few issues that need to be addressed including some further clarifications to enhance the overall clarity of the manuscript and potential impact on routine care.

TITLE:
- maybe better to mention that this is a diagnostic accuracy study (instead of prognostic accuracy for example)

Response: We have changed the title accordingly, and have included the study design according to BMJ’s format: “Diagnostic accuracy of COVID-19 self-tests with unsupervised nasal or nasal plus oropharyngeal self-sampling in symptomatic individuals in the Omicron period: cross sectional study”.

WHAT THIS STUDY ADDS:
- Some of the percentages are shown with 1 decimal figure, others are not. Best to be consistent here.

Response: We removed all decimals throughout this section.

- the statements are still very verbatim what the results show. It would be better to mention here what these results mean for clinical practice and further research in a few statements.

Response: We agree with the reviewer and rephrased the section to:

- “The sensitivities of three commercially available Ag-RDTs performed with nasal self-sampling decreased during the emergence of Omicron from 87% to 81% for Flowflex, from 83% to 76% for MPBio, and from 80% to 67% for Clinitest, with only the latter reaching statistical significance.
- Addition of oropharyngeal to nasal self-sampling improved the sensitivity of the MPBio test from 70% to 83% and the Clinitest test from 70% to 77% (not done for Flowflex), most notably in individuals who visited the test site for other reasons than to confirm a positive self-test.
- Based on these findings, the manufacturers of MPBio and Clinitest may consider extending their instructions for use to include combined oropharyngeal and nasal sampling, and other manufacturers may consider evaluating this as well.”

ABSTRACT:
- interventions: There is a different timeframe defined for the different test methods, how is this taken into account in the analyses and how comparable are these periods?

Response: The reviewer is correct that nasal and OP-N sampling were conducted in different time periods of the study. However, in both time periods, the Omicron variant was present in >90% of samples in the national surveillance. Potentially, the share of the Omicron variant may have been higher during OP-N sampling period. Since we observed a decline in diagnostic accuracy with an increasing share of the Omicron variant in the nasal sampling period, the higher share of the Omicron variant in the OP-N sampling period may have led to an underestimation of the true difference in diagnostic accuracy between both sampling methods. So we are confident that OP-N sampling is superior to nasal sampling in the Omicron era. Prompted by this comment, we have added the above to the limitations section of the discussion on page 15, lines 451-457: “Nasal and OP-N sampling were conducted in different time periods, but the Omicron variant was present in >90% of samples in the national surveillance in both periods. Potentially, the share of the Omicron variant may have been higher during OP-N sampling period. Since we observed a decline in
diagnostic accuracy with an increasing share of the Omicron variant in the nasal sampling period, the higher share of the Omicron variant in the OP-N sampling period may have led to an underestimation of the true difference in diagnostic accuracy between both sampling methods. Therefore, we are confident that OP-N sampling is superior to nasal sampling in the Omicron era.”

Response: According to the reviewer’s suggestion, “molecular testing” was changed to “RT-PCR testing”.

- Some of the percentages are shown with 1 decimal figure, others are not. Best to be consistent here.
Response: Decimals were added throughout the abstract.

- conclusions: please mention the clinical implications of your findings for current practice.
Response: Per suggestion of the reviewer, we added the following implications to the conclusion on page 4-5, lines 121-130: “A positive self-test justifies prompt self-isolation without need for confirmatory testing. Individuals with a negative self-test should adhere to general preventive measures because a false-negative result cannot be ruled out. Ag-RDT manufacturers may consider extending their instructions for use to include combined oropharyngeal and nasal self-sampling, while other manufacturers should consider evaluating this as well.”

INTRODUCTION:
- line 121: suggest to add "diagnostic" to "accuracy"
Response: added accordingly.

- line 124: " yields higher diagnostic accuracies": please specify which diagnostic accuracies you are focussing on here (both sensitivity and specificity, or just sensitivity?). This is important to assess the clinical relevance of your findings.
Response: changed to “yields higher diagnostic performance”, without specifying the actual performance measures as these are elaborated on in more details in the statistical analysis section.

METHODS:
- specimen collection and testing: the different tests were not performed simultaneously, thus a head-to-head comparison is possible. Furthermore, the order of tests was not randomised, so the presence of viral material over time could have influenced the results.
Response: While the RT-PCR sample collection and the Ag-RDT were not performed simultaneously, they were performed within a 3-hour window within the same individual. We therefore believe that it is justified to speak of a direct comparison. The three different Ag-RDTs, however, were not compared head-to-head because they were not performed in the same individuals. We did consider asking participants to perform all three Ag-RDTs (plus the RT-PCR at the test site) within a few hours but explicitly decided against this because of the high burden on participants, because the quality of the first sample may be better than the quality of subsequent samples, and because the first test result may influence the interpretation of subsequent test results.

- the description of the methods for sampling demonstrates large heterogeneity in recruitment and sampling, which further complicates the analyses and the generalisability of these findings.
Response: Indeed, the slight differences in sampling methods for the reference test, i.e., the RT-PCR test, might have slightly influenced the results of the study. However, we think that RT-PCR test performance is very high regardless of whether the slightly less invasive OP-N or more invasive OP-NP sampling is used. Indeed, the test site that evaluated the Clinittest used the less invasive OP-N sampling method (compared to OP-NP method at the other two test sites) but the Clinittest performance was in fact worse (rather than better) than the performances of the other two tests. We therefore do not expect that the sampling method of the reference test substantially impacted our results or generalizability. The following text was added to the limitations section of the manuscript on page 15-16, lines 457-463: “Slight differences in sampling methods (OP-N vs. more invasive OP-NP) for the reference test, i.e., the RT-PCR test, might have influenced the results of the study. However, we think that RT-PCR test performance is very high regardless of the
sampling method. Indeed, the test site that evaluated the Clinitest used the less invasive OP-N sampling method but the Clinitest performance was in fact worse (rather than better) than the performances of the Flowflex that used the OP-NP sampling method. We therefore do not expect that the sampling method of the reference test substantially impacted our results or their generalizability."

- apart from the comparison between sensitivities ("Sensitivities in the first and last week were compared by Chi-square tests"), I cannot find the methodology of the time-trend analyses mentioned in the study design-section.

**Response:** We realize that the term “time trend analysis” may suggest statistical assessment of changes over time. Instead, we assessed sensitivities in different inclusion weeks and statistically compared the sensitivities in the first inclusion week with the sensitivities in the last inclusion week. We revised the phrase “time trend analysis” to “assessment of sensitivity over time” throughout the manuscript and have added the previous sentence to the methods. In addition, we rephased the following sentence of the methods section on page 9, lines 225-227 from “by analysing the Ag-RDTs’ sensitivities and specificities per week. [...] Sensitivities in the first and last week were compared by Chi-square tests.” to “by assessing the sensitivities and specificities in different inclusion weeks and statistically comparing the sensitivities in the first inclusion week with the sensitivities in the last inclusion week by Chi-square tests.”

- Please also specify the software used for your analyses.

**Response:** The following statement was added at the end of the statistical analysis section (page 9, line 238): “All analyses were performed in R version 4.1.2 (2021-11-01) "Bird Hippie."

**RESULTS:**
- the time-trend analysis description is a bit vague and it is difficult to assess its relevance.

**Response:** We changed this section on page 11-12, lines 303-305 to avoid ambiguity: “With emergence of Omicron, sensitivities decreased to 80.9% for Flowflex (Chi-square test statistic 2.0; p-value=0.16), 73.0% for MPBio (Chi-square test statistic 0.28; p-value=0.60), and 70.3% for Clinitest (Chi-square test statistic 5.0; p-value=0.025).”

- the heterogeneity inherent to the data is addressed by performing subgroup analyses, however further diluting the overall message.

**Response:** This is something we also struggled with when writing up our study. The subgroup analyses were prespecified in our study protocol and therefore in our view had to be reported. For the reason mentioned by the reviewer, we already reported some of the discussion of the results of the gender and age subgroups in the supplementary material to maintain focus on the main findings in the main text.

**DISCUSSION:**
- The first section contains a lot of results, which are already shown earlier. Consider focussing on the main findings only to further set the stage of your discussion section.

**Response:** We agree with the reviewer. Per BMJ’s recommendations for the discussion section, however, we started the section with summarizing our main findings. Per suggestion of the reviewer we revised the first paragraph on page 12, lines 316-323 to: “This large diagnostic accuracy evaluation of three commercially available SARS-CoV-2 Ag-RDTs used unsupervised by symptomatic individuals with nasal self-sampling showed a decline in overall Ag-RDT’s sensitivities with the emergence of Omicron. The observed decline was, however, only statistically significant for Clinitest, not for Flowflex or MPBio. Sensitivities increased when the Ag-RDTs (assessed for MPBio and Clinitest only) were combined with OP-N self-sampling instead of nasal self-sampling only. Sensitivities were substantially higher in confirmatory testers than in those who visited test-sites for other reasons. Only the MPBio self-test combined with OP-N sampling met the World Health Organization’s standards for Ag-RDTs (≥ 80% sensitivity and ≥ 97% specificity among symptomatic individuals).”

- The heterogeneity is, in my view, not sufficiently highlighted in the discussion section. It would be useful for the readers to learn more about the potential impact of your choice of design/methods on the study results and how this could feed in to future research.
Response: Based on the reviewer’s previous comments, the time interval between RT-PCR testing and performing the Ag-RDT, the different time frames for the different Ag-RDT sampling methods, as well as the heterogeneity in sampling were added as limitations to the discussion section on pages 15 & 16.

- "We found trends towards lower sensitivities": not sure what is meant here, but unless you have shown a statistical significant trend, it might be better to avoid such statements.
  **Response:** We agree with the reviewer and rephrased this part on page 14, lines 394-396 to “Sensitivities were significantly lower in participants that had a previous SARS-CoV-2 infection (67%) compared to those who had not (83%) for Flowflex with nasal sampling. Non statistically significant differences over 10% were found for the MPBio with nasal sampling, and for Clinitest with both sampling methods.”