This study was aimed at investigating sex differences in the frequency of pain-related behaviors after tail docking and emotional reactivity as a measure of emotional state. The study is interesting and welcomed as there is limited information on this topic in veterinary medicine (most available data related to lab animals). Overall, higher frequency of pain-related behaviors was observed in female than male lambs after tail docking, and in lambs after tail docking when compared with sham surgery. Additionally, lambs after tail docking showed greater duration of behaviors related to fear than those undergoing sham surgery. The manuscript concludes with the potential welfare concern related to the large flock of female ewes which could be affected by chronic pain. The research is relevant and adds to the literature. However, there are important issues that must be addressed. Although some of them cannot be resolved, they must be clearly identified and discussed. These are listed below in order of importance:

1) It is unclear why an ethogram was used instead of a validated pain scale (Lamb Grimace Scale; doi:10.1016/j.beproc.2016.09.010) that is already published and involved the same painful procedure in lambs. Using such a scale would have strengthened the validity of the results. As a consequence, the authors cannot assume with a strong level of confidence that they are indeed measuring pain (e.g. construct validity) or whether the assessment if reliable (e.g. inter- and intra-rater reliability) (doi: 10.1097/j.pain.0000000000002474). For these reasons, this reviewer asks the authors to discuss these issues and reword references to 'pain sensitivity' making it clear that the outcome measures were frequency of the occurrence of pain-related behaviors (assuming that is the case – please see below).
2) Baseline assessments were not performed. This would have strengthened our confidence that the behaviors observed after surgery were indeed related to pain. A sham group was included and behaviors were different between sham and tail docked indicating that those were likely related to pain. However, it is unclear how the observer was blinded if in theory they could see whether the tail was docked. Another alternative would have been to have a positive control group when an optimal analgesic protocol would be administered after tail docking to male and female individuals.
3) The actual data that were generated are unclear. It seems that pain-related behaviors were related to the number of times a behavior was observed regardless of its duration whereas emotional reactivity data were related to the duration of different behaviors. Additionally, the actual data should be reported (perhaps as supplementary material or in a table). This information is extremely important as it will help the reader to understand the methodology and interpret the findings.
4) The authors are invited to consult and use the ARRIVE guidelines to improve reporting of the study. The methodology is unclear in several instances.
5) Pilot study: The methodology is unclear and, as is, the study could not be replicated. Considering that the methodology from the pilot study was not used at all in the main study, all information related to the pilot study could be deleted as it would make the manuscript more concise and objective. However, if the pilot study is kept in the manuscript, then a lot more information needs to be added.

Line 79: Please clarify: Was the decrease in sensitivity at 12 days when compared with 1 day? Or thermal nociception was decreased in male when compared with females at 12 days?

Lines 81-83: Please specify for which species this evidence is available. Also, what does 'being more sensitive to analgesic treatment' mean? Please clarify/reword.

Lines 83-85: This sentence is difficult to understand. Please reword.

Line 90: Why would pain be more difficult to asses after the first hour? Please explain your rational.

Line 93: What can be indicative of negative affect? It seems that this sentence is referring to a specific finding in emotional reactivity tests, no? Please clarify/reword.

Lines 87-101: Why is there no mention of the Lamb Grimace Scale? This tool has been validated in lambs following tail docking (doi:10.1016/j.beproc.2016.09.010). Thus, it is highly relevant for the study in question as it involves lambs undergoing the same surgical procedure. Also, what is the level of evidence for the validation of emotional reactivity tests in lambs? A reader unfamiliar with pain assessment might interpret that behavioral assessment of pain using pain scales is not reliable and that emotional reactivity tests provide a better assessment of the emotional state of the animal which might be related to pain or not. If the validity of pain assessment tools is being put in question, then a throughout literature on the topic is warranted. Finally, what are emotional reactivity tests actually testing in this study? It seems that they are testing the emotional reaction to neutral, positive or negative events after restraint (lateral recumbency and tail manipulation) +/- surgical pain. It is clear from the objective of the study that the authors clearly understand these differences. However, the message on these lines do not necessary reflect this belief.

Line 124: The abstract says '5 cohorts of 16' and here it says '6 cohorts of 16'. Please correct.

Lines 153-159: It seems that habituation was done in groups of 4 for 7 days and then in pairs for 2 days. Please clarify. Also, how were the pairs chosen?

Lines 160: Please provide more detail so that the methodology could be replicated. When lambs were in pairs, how long were they kept in the pens?

Also, "Once lambs were observed .." Was one lamb removed and the other left alone? Please provide more detail.

Lines 171-173: Please provide more detail. For example, freezing response was a categorical (yes/no) or numerical variable (duration of freezing). Were duration and frequency recorded for all behaviors? Please describe how 'avoidance of the dog location' was defined. How was behavioral change from baseline defined?

Lines 202-203. It seems there is a verb missing.

Lines 206-207: It does not seem that the methodology is aimed at 'measuring the emotional state' (i.e. the emotion is not being quantified). From reading the methodology it seems that duration of different behaviors was quantified. Please clarify.

Line 208: Please confirm if sham lambs were also placed on their back in a marking cradle.

Line 222: How was the observer blinded? It seems one would be able to see whether the tail was docked or not. Was this observer involved in the experimental phase? How long after video recording were assessments done? Was a single individual doing the evaluations? What was the sex of the observer? Was there training on the assessment of pain-related behaviors performed?

Lines 221-228: It is nice that pain-related behaviors previously described in the literature were included in an ethogram. However, and ethogram is only a list of behaviors which is quite different from a validated pain assessment tool. It is unknown whether these behaviors are really measuring pain unless scientific validation is undertaken (e.g. construct validity). Using a validated pain scoring system such as the Lamb Grimace Scale would have provided more reliable results. These limitations need to be discussed.

Also, how were the behaviors in Table 1 recorded? Yes/no, duration and frequency, using scores? Please clarify.

Please confirm whether recording occurred in the pen where the other 3 lambs and their ewes were also housed. In that case, how were animals identified?

Why were baseline behaviors not assessed? Baseline pain assessment is extremely important and allows comparisons with pain behaviors after the painful stimulus. It helps to confirm that behaviors are related to pain and not something else. This limitation should be discussed as well.

Lines 237-238: After highly stressful +/- painful events (removal from pen + restraint + placement in lateral recumbency in a device +/- tail docking) lambs were transferred to the test arena in isolation to then be exposed to novel and startle stimuli. Thus, there were numerous novel and stressful events occurring in close temporal association. These should be taken in consideration for the interpretation of the results.

Tables 1 and 2: Please add a column describing the possible response options for each behavior. This will help us understand the data that were generated. Also, what is 'Y' in Table 2?

Line 280: 'The acute pain behaviors were all summed'. It is not clear what was summed as it didn't seem that behaviors were scored. Please see the previous comment.

Line 286: All animals should be n=80 (5 X 16), no?

Statistical analyses: Were adjustments for multiple comparisons performed? For example, for each behavior, it seems that contrasts were done between each cohort. This results in numerous comparisons and one would think that significant findings would occur by chance.

Line 353: What is KMO? Please define and explain.

Line 354: Replace 'the' for 'that'

Line 360: Define AIC as it is the first time it appears. If this information (i.e. AIC and the choice of model) is being added to the manuscript, then a small explanation should follow. Otherwise, you might omit this.

Line 373: Replace 'the' for 'that'

Lines 357-358 and 376-377: What were the criteria to name the components as 'fearful', exploration' and 'active response'? Similar question for lines 400-402.

Line 392: What is a 'high reaction'?

Line 414: "Pain research in livestock often focuses on male animals" – Please provide references.

Lines 416-417: What would be the practical consequences from learning that a sex is more sensitive to pain? Please elucidate on this somewhere in the discussion.

Lines 432-437: This discussion is interesting. But please, add to it by presenting evidence related to what is expected to be different between male and female lambs after 8 weeks of age that would influence pain sensitivity.

Lines 443-445: The information in this sentence is contradicting. First is says that lying behavior is reduced after analgesia, then it says that standing behavior is displayed after tail docking. Please rewrite.

Line 450: Missing 'in' before lambs

Lines 451-455: Please clarify from which species this evidence pertains to

Lines 468-473: Another possible explanation is that the sex effect on pain behaviors is a type I error. This should be discussed.

Lines 527-528: "Lambs that underwent tail docking displayed more behavioral indicators of pain.. sensitivity". The wording herein seems more appropriate than in previous references to the findings and should be used throughout. Assuming that the data refer to frequency of behaviors, and accepting that the ethogram is not a validated pain scale and that we cannot ascertain that the study is measuring 'quantity of pain', then using the terms 'number of pain-related behaviors' or 'number of behaviors believed to be associated with pain' or anything similar, is more adequate.

Line 535: Delete ';'