We would like to thank both reviewers for their comments. Detailed answers to the comments are provided in blue below, with changes applied to revised manuscript given in bold.

RC1 Anonymous

This manuscript presents a well-documented test of precision and reproducibility of ID-TIMS U-Pb dates using pre-spiked solution of natural zircon, undertaken by the EARTHTIME community. The paper is well written, and in principle can be published in the present form. Still, it can be improved by greater attention of details in presenting analytical data and interpretations.

Thank you for the supportive review. Below we address the main comments from the text file of the review as well as several longer comments copied here from the pdf.

First of all, I would like to express my opinion about the design of this study. In my opinion, some design decisions made at the start of the study were not optimal, and significantly reduce the value and utility of this test.

- First and foremost, the choice to use a 205Pb-233U-235U spike without 202Pb. It is mentioned in the text several times that this decision was driven by low availability of 202Pb and ET2535 mixed spike. However, without the numbers showing the size of the remaining stock of ET2535, and the amount of spike used in this test, it is impossible to see whether this decision was justified. The data presented in this manuscript clearly show that instrumental isotope fractionation of Pb causes the greatest component in the age uncertainty (and this is emphasized in the text many times), but without Pb double spike, it cannot be properly quantified and accounted for. In my opinion, the authors must present complete quantitative information that lead them to the decision to use single Pb spike. If there was any chance to use ET2535 without critically depleting the remaining stock, I would consider the choice of spike made in this study an unforgivable mistake. Using the spike containing 202Pb and 205Pb would provide a lot of additional valuable information, without losing any currently available information, because the Pb isotope analyses with 202Pb-205Pb can be reduced with both internal and external fractionation correction.

We sincerely hope the reviewer will forgive us for the choice we've made here! There is no doubt that using a ^{202}Pb–^{205}Pb spike would bring more options to our analysis. However, as a counter point, the systematic inter-lab differences we currently see are clear even with a single Pb spike (where each lab propagated a large uncertainty on the Pb fractionation factor), so we would argue that we as a community are not yet at a point of reproducibility where we would need a double spike to resolve our issues. As such, after some discussions we concluded there was a need to deplete the supply of ^{202}Pb for this experiment given that there is about a decade's worth left in stock.

We hope that improvements brought about by experiments such as this one will get us there in the next years. We cannot change the spike used in this study, but we agree we should consider using a double spike in future experiments.

- The decision to distribute only the pre-spiked solution. In my opinion, a better way would be to split the solution into two portions, pre-spike and equilibrate one of them, and then provide each participating lab with two aliquots: spiked and unspiked. The participating labs should have been allowed to use the spike of their choice for analysis of solution provided without spiking. This approach would have at least two advantages compared to the one used in this study. First, it would allow to determine the magnitude of errors related to sample-spike homogenisation (and to spike calibration, in the cases where spikes other than ET525 and ET2535 are used). Second, it would allow participation of the labs that are involved in the measurements of natural 238U/235U, and hence avoid handling enriched 235U. The only downside is the increasing the number of analyses, but additional 10 (or even 20) analyses is a fairly modest burden for a lab that specialises in U-Pb dating and performs hundreds of analyses (in some cases many hundreds) each year.

Thank you for this, this is an interesting idea that we should definitely explore in future exercises.

- The third is the decision not to accompany analyses of zircon solution with analyses of synthetic age solutions, e.g. ET100 or ET500. I consider this a missed opportunity to check whether any systematic differences between the labs vary in the same way for two or more age solutions.

This is another good idea to consider next time. We opted for a natural zircon material and decided to use the spike supply that was easily available on PLES535 because it most closely resembles routine zircon unknowns. But this could easily be expanded to ET solutions; probably best pre-spiked and then distributed as done here.

- The fourth problem is insufficient supporting technical information from the labs that does not contain many potentially important pieces: useful ion yields for Pb and U analyses (should be included in the supplementary excel table), gain and baseline history, any determinations of cup efficiency performed on the same instrument, type of the ion counting multiplier (including manufacturer and the model), details of the deadtime calibration and linearity assessment, any analyses of the interference patterns (if

> performed), any in-run correction of oxygen isotope fractionation in UO2+ analyses. If the same instrument is used for high-precision (ppm level) isotope analyses of other elements, these data could also be useful. With these data, the evaluation of uncertainties related to baseline, gain, linearity, cup efficiencies and the like would be much better constrained, and not as speculative as in the current version. Fortunately, it is not too late to request these data from participating labs, and include them in the paper.

This is similar to a comment by Reviewer 2, please see our answer there. Briefly, we are of the opinion that these are topics worthy of entire papers (e.g. how to calibrate ion counter deadtime) and including detailed descriptions of how each lab calibrated each part of the detection system would only serve to distract the reader from the main point here – which is to create a snapshot of the current level of lab-to-lab reproducibility and provide insight into where the community needs to improve. With this paper we hope to initiate discussions about to how to move forward and optimize all these practices, but we are unlikely to solve all these issues right now and with this manuscript. We would therefore prefer not to add these details.

More specific questions are marked in the attached annotated manuscript. They are the inherent part of the review, and I encourage the authors to consider them as seriously as the text above.

Below we copied major comments from the annotated manuscript that go beyond small wording corrections (and do not duplicate comments above):

L. 116-118: I find it hard to accept this statement without firm support with solid figures on availability of 202Pb and the amount of the remaining stock of the ET2535 spike.
You must present the numbers. What amount of spike was used in this experiment? How large is the total remaining quantity of the ET2535 spike? Without numbers, these are empty words.

This is addressed above. We do not think it is particularly relevant how much spike is left exactly, the important point is that we did not want to waste it. We could have done so but the added benefit was not considered worth it at the current level of inter-lab reproducibility. Perhaps the reviewer disagrees with this assessment, but there is not much that can be done now in the context of this manuscript.

L. 131:
20 mg of zircon split between 30 capsules makes about 0.7 mg per capsule. This is a HUGE amount of zircon per capsule - about 2-3 orders of magnitude greater than in regular practice of zircon dating. Formation of precipitates, in particular REE

fluorides, is very likely in these conditions. Have you checked your solutions for possible presence of colloidal material? E.g., by running REE elements on ICPMS?

This is a very good point. We were completely aware that zircon load for dissolution is greater than the usual amount of zircon and we have taken care to add excess HF for dissolution to avoid that the solution becomes supersaturated (which may lead to the formation of small clusters or aggregates). The presence of colloidal particles was not checked after the dissolution, but we aimed at complete fluoride conversion to chloride form via re-dissolution under high pressure and temperature. Importantly, we did not attempt here to date Plesovice but to measure U and Pb in the obtained solution – whether it is fully dissolved Plesovice or not. The results, which agree well with literature data for Plesovice, suggest that we managed to get U and Pb in solution at the right proportion.

L. 210:
It would be very useful to include useful ion yield values for each measurement in the Table 2. And this is easy to do. If you had these numbers, there would be no need to speculate.
L. 381: Ionisation efficiencies for analyses in this study should be reported in the supplementary table, along with other analysis details.

This would indeed be very helpful, but it is not as easy as the reviewer suggests. We could compile measurement durations and intensities, but our Pb measurements are essentially never run to exhaustion because U is loaded on the same filament and measured after Pb. So, we are probably quite far from detecting all Pb ions (likely less so for U); consequently, the calculated ion yields would be off.

L. 258:
Proof?

The consistent results in the experiment (more specifically, good repeatability within individual labs) proves that the solution was close to homogeneous. There are some examples (line 248+) where this could be questioned; we do not have a good explanation for these cases. In any case, the sentence in line 258 refers to the intention of producing a homogeneous solution and contrasts it with natural zircon crystals; it is not focused on proving the case.

L. 300:
How do you quantify interferences, considering that they are isotope-specific and vary significantly throughout the run?

This is again left to the experience of each lab. We are mostly concerned about BaPO2 and Tl (interfering at 204 and 205) which are variably monitored by looking at mass 201 and 203, respectively. Additionally, some laboratories may have lowcount rate interferences, present at all mass stations but particularly problematic on Pb-204, particularly at the beginning of runs, that are attributed to volatile organic compounds. Whether any intensity on these masses is due to these particular ions can be occasionally checked by verifying isotope ratios in blank measurements – but other ions interfering with other Pb masses are also possible. This is particularly the case if the instrument is used to analyze other elements. For the purpose of modelling here, we simply assume additional counts (up to 50 cps) on all masses (line 328) and explore what that would do to measured ratios. Details of the modelling can be explored and modified in the supplementary file.

L. 309:
Weird and misleading wording. Ion counters (both Daly and SEM) have their own mass bias, whereas Faraday cups don't, at least at the level of precision of this study. So for Faraday measurements we observe just evaporation-induced fractionation, whereas for ion counters, we see a combination of the latter and the detector-induced biases.

That's fair, we have corrected this (now line 321).

L. 503:
The question is how exactly to do this. Should the uncertainty be propagated to the individual analyses or to the final ages? These two approaches will yield very different results.

The final ages. The inter-lab reproducibility that we calculate refers to those (it compares weighted mean ages) so that would probably be the most appropriate way to do this.
This is already suggested by the current phrasing "propagate the uncertainty (...) onto final boundary ages" (l. 524-525).

RC2 Brian Jicha

Szymanowski et al present the results of an interlaboratory experiment on a pre-spiked zircon solution to evaluate the thermal ionization mass spectrometry for U-Pb analyses and the inherent corrections associated with them. The manuscript is impressive, clear, and well-written. All steps of the process, including preparation of the solution, are clearly explained. The manuscript can be published after some minor modification.

Thank you for this assessment and the helpful comments. We address them one by one below.

The novelty of the community experiment is a little oversold as they have eliminated many variables from the typical U-Pb process. It is often very smart to eliminate variables in an analytical protocol to understand the limitations of some of the steps in the process, however, the manuscript needs to do a better job of discussing the limitations of this experiment.

We stand by the general strategy we chose, and the significance of the results. Before this study, we had very limited understanding of how results would compare for the larger community. That understanding, as described in the introduction, was based on comparisons involving zircons dated by 2 labs (e.g. when a study was moved between labs e.g. Schoene et al. Geology 38: 387-390, 2010), 3 labs (e.g. Schaltegger et al. 2021) or maximum 5 in the case of a few zircon reference materials (e.g. Nasdala et al. 2018). Other clues have come from synthetic U-Pb solutions which do not perfectly mimic the analytical workflow of a normal zircon analysis. Such comparisons have serious limitations, which are currently described.

In contrast, this is the first time we dated the same homogenized zircon material at every lab. The experiment was designed as a first-order test of how well we can date the same material – the main limitation being that we cannot yet deconvolve the exact sources of disagreement. We can discuss the possibilities fairly well (see Fig. 6), but we cannot pinpoint them exactly. This would be possible if we eliminated even more variables (e.g. using a double Pb spike to eliminate a blanket mass fractionation correction, as suggested by reviewer 1), not fewer. Future experiments could focus separately on Pb and U isotope analyses, for example.

A significant limitation of this experiment is that, depending on the nature of the samples, the uncertainty budget for U-Pb analyses can be more or less sensitive to various quantities. In this experiment, because it focused on a “best case” scenario with relatively large amounts of sample, and only used a single ^{205}Pb tracer, the ^{206}Pb/^{238}U dates are primarily sensitive to the mass bias correction. This experiment therefore may not reflect the quality of inter-lab agreement for other types of samples or conditions. For example, when samples have small amounts of Pb*, they will be more sensitive to blank corrections, and therefore be most sensitive to the blank isotopic composition. Unfortunately, given the wide range of possible conditions, it is impossible to test each one in a single experiment.

We made sure that these limitations are spelled out more directly in the text. In section 6, which already covered the low-Pb* case, we have now added a sentence about Pb/Pb ages. Thus, this limitation of the experiment is made very explicit in the most accessible, "general audience" section.

For example, on line 87 and again on line 90 and later in the discussion, the text states that they have chosen to avoid local tracer addition and avoid sample-spike equilibration. Why? Are there recent data available to suggest that this step is an issue? If so, please cite it or explain further why they have opted to eliminate this fundamental step from the experiment.

The main reason is that we tried to eliminate the lab-to-lab variability of spike compositions from consideration, as not every ID-TIMS measurement uses the same spike. Even if every lab used the EARTHTIME tracer (assuming they have access to it, which should not be a requirement of taking part in this open experiment), we have seen small differences in the minor Pb isotopic composition of different bottles of spike. This may have something to do with the bottles themselves and labs are careful about it, but we believe eliminating this variable is key to this and any further interlab experiments. For a proper comparison, we simply have to make sure we are analysing the same sample-spike mix.

Regarding sample-spike equilibration, we do not think there is an issue, but doing this once for the common "mother" bottle of solution rather than for every aliquot at every lab seems like an obvious choice. The main risk here is that the solution in the mother bottle is not well equilibrated, but this seems to have worked pretty well given the coherent results.

Moreover, the reasons for the differences between the labs could be better conveyed or displayed so that people can see what analytical protocols produced the best results. Right now, the reader can only deduce what instrument/collector configuration produced the best results. The manuscript could serve as a very useful guideline for instrument operators if more of the instrument parameters were given.

We do not mean to give the impression that there is a meaningful range in overall data quality or that there are best results. The participants in this experiment were making analyses representative of typical conditions, not maximizing precision. For example, since the Pb measurements were singly spiked, the Pb fractionation correction was the largest component of the uncertainty budget, and therefore some analysts made more efficient, albeit lower precision, isotope ratio measurements, knowing that collecting additional data would not affect the precision of the derived date. There are obvious advantages of certain hardware aspects, such as collecting using Faraday cups with low-noise amplifiers vs. ion counters (as mentioned in the text) but being able to leverage improved hardware is a function of the ion yield, which can also vary as a function laboratory specific procedures and materials. The key differences in analytical precision or internal repeatability of a lab are likely due to how well we can ionize and how well we calibrate the detectors and not really because there are differences between the mass spectrometers used.

To address this comment, we have stressed the notion that the available TIMS hardware does not appear to be the limiting factor in obtaining data comparable to the rest of the community (lines 272 and 465).

On line 163 it says that the dead time calibration and baseline settings have been left to the decision of the participants. These parameters for each lab/instrument should be added to Table 1 so that other users of these instruments can see how they are being utilized. Thus, Table 1 should probably then be subdivided into two tables or a Table 1a and 1b with one showing additional instrument parameters (dead time, baseline settings, acquisition times, # of cycles, # and frequency of standards (ET 2535 or SRM 981) used for Pb mass fractionation correction, etc) and another table with blank information.

This comment is similar to what's suggested by reviewer 1 so we answer both of these here. Briefly, we believe adding all these parameters will be distracting while not bringing significant new understanding to the manuscript.

For instance, there are no standard techniques to characterize deadtime on ion counting devices used in TIMS measurements, and the participants used a variety of methods to characterize their detection systems. At the < +/-1 ns level, it becomes increasingly difficult to account for different effects on absolute and relative ion beam intensities, such as fractionation, blanks, and detector non-linearity and mass dependent response.

Baseline settings are calibrated by each lab and there are also different approaches, e.g. measuring in-method or offline ("electronic baseline"). Labs should make sure that baseline variability is properly accounted for, but describing how this was done will not help the reader decide which way is better – the users should investigate what works best for their particular setup.

Acquisition times are a function of intensity, beam stability, and the expected analytical precision. It is also not standardized, as the required precision of measurement varies with the characteristics of the sample; analyses are generally stopped manually when satisfactory precision is obtained. As such, displaying this would not be particularly helpful for the reader, would not necessarily show a useful correlation with other variables of the experiment, and may be misleading to a novice user.

Given these complications and the lack of standardization, and given the large number of participating labs, such a compilation effort would be extremely time consuming, substantially expand the manuscript, and detract from the main goal, which was to create a snapshot of the current level of lab-to-lab reproducibility and provide insight into where the community needs to improve.

Line 260: The text states that by using a pre-spiked solution, it offers the opportunity to exclude geological bias. While this statement is true, the limitations of using a pre-spiked solution are not stated and should be. This would be another place to add a brief discussion of the limitations of the experiment.

See above.

As noted by the authors, in Figure 3, there are large and systematic differences in raw ratio precision among labs. They suggest that there are systematic differences in terms of Pb ionization efficiency or acquisition time. To further illustrate this point, a plot of raw ratio precision vs acquisition time would be beneficial within Figure 3. This plot could potentially address if some of the differences are a function of acquisition time vs. ionization efficiency. The same thing can be done for Figure 2.

This is similar to a comment by reviewer 1 who suggested calculating useful ion yields. While we do not think that is possible with our data because samples are not run to exhaustion (so many ions are never counted, see answer to rev. 1), a plot like the one suggested here could go some way towards deciding whether the differences in analytical precision are more likely due to different acquisition times or differences in ionization efficiency. However, it would neglect the third major factor which is the sample evaporation rate. Simply put, an analysis run at higher filament temperature~evaporation rate (which could be shorter due to quicker exhaustion of material) could result in greater precision even if ionization efficiencies were equal. However, there are many differences between Re filaments (between batches, manufacturers, impurities, thickness characteristics, degassing histories) so different labs' filaments can have a different evaporation rate at the same temperature.

Thus, even compiling run intensities, acquisition times and temperatures from the participants we wouldn't be able to distinguish the relative importance of evaporation rate and ionization efficiency. As a result, the proposed representation or its modification will not give us the answer we're looking for. The only way to understand this would be to perform a separate experiment where we can control key variables, for example by using a single batch of filaments run at a predefined temperature.

We have mentioned evaporation rate (~temperature) as an additional factor in the revised manuscript (lines 215-219).

The text states that the large differences in the precision of 208Pb determinations reflect reduced 208Pb counting times preferred by some labs. Again, the counting times of each lab could be shown with an additional panel in Figure 2 (instrument vs counting time OR 208/205 precision vs counting time) or as a column in Table 1, preferably both. The "Next Steps" #4 says that common protocols for ion detector performance are needed. If all these parameters are given for each lab in this manuscript, it will become clearer to the community what protocols may be needed.

This is an observation that was made in the study, but its relevance to the U-Pb dates of zircon is limited; we only use ^{208}Pb to calculate Th/U in the zircon and perform a correction for initial ^{230}Th/^{238}U disequilibrium. The precision of the ^{208}Pb analysis is not very important given the dominant uncertainty in the assumed Th/U of the magma. As a result, we prefer to not expand the discussion of ^{208}Pb determinations in the manuscript, as this aspect of the measurement is not significant to the outcomes of this study.

Lastly, I will offer a comment. The group has chosen a 337 Ma zircon solution. This age represents a sweet spot for U-Pb geochronology as there is adequate parent and daughter available for analysis. I would encourage any future intra- or inter-laboratory experiments to be conducted on zircons that are closer to the extremes of Earth's history to fully assess some of the limiting factors on laboratory protocols. For example, instead of a 337 Ma zircon solution, the group should consider ~3 Ma zircon or ~3000 Ma zircon solutions.

Thank you, this is a good idea for future experiments. We chose Plesovice precisely because it is easily available, has enough Pb and U to perform high precision analyses of both elements, and is the age range where the quality of both U and Pb measurements matter. It is indeed where U-Pb geochronology excels. The old and young "end members" are in our future plans, though it is important to note that such experiments will be mostly testing other inputs/corrections, specifically the accuracy of the Pb blank correction for 3 Ma and Pb ratio measurements for 3 Ga.

This is a solid manuscript. Nice job to all involved.

Brian Jicha

University of Wisconsin-Madison

Thank you again!