

## Reviewer RC1

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  $^{205}\text{Pb}$ - $^{233}\text{U}$ - $^{235}\text{U}$  spike without  $^{202}\text{Pb}$ . It is mentioned in the text several times that this decision was driven by low availability of  $^{202}\text{Pb}$  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  $^{202}\text{Pb}$  and  $^{205}\text{Pb}$  would provide a lot of additional valuable information, without losing any currently available information, because the Pb isotope analyses with  $^{202}\text{Pb}$ - $^{205}\text{Pb}$  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  $^{202}\text{Pb}$ - $^{205}\text{Pb}$  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, we did not believe there was a need to deplete the supply of  $^{202}\text{Pb}$  for this experiment.

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  $^{238}\text{U}/^{235}\text{U}$ , and hence avoid handling enriched  $^{235}\text{U}$ . 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  $\text{UO}_2^+$  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 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 BaPO<sub>2</sub> and Ti (interfering at 204 and 205) which are variably monitored by looking at mass 201 and 203, respectively. Additionally, some laboratories may have low-count 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 will get this corrected.

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.