

## Reviewer 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.

Another 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  $^{205}\text{Pb}$  tracer, the  $^{206}\text{Pb}/^{238}\text{U}$  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 will make sure that these limitations are spelled out more directly in the text.

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 will stress 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, but how we calibrate and monitor it is critical. The variables of note are already listed.

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  $< \pm 1$  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. They are 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 will mention the run temperature as an additional factor in the revised manuscript.

The text states that the large differences in the precision of  $^{208}\text{Pb}$  determinations reflect reduced  $^{208}\text{Pb}$  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  $^{208}\text{Pb}$  to calculate Th/U in the zircon and perform a correction for initial  $^{230}\text{Th}/^{238}\text{U}$  disequilibrium. The precision of the  $^{208}\text{Pb}$  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  $^{208}\text{Pb}$  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.

Thank you again!