Review of: Raitzsch et al. - Technical Note: Single-shell $\delta^{11}B$ analysis of *Cibicidoides wuellerstorfi* using femtosecond laser ablation MC-ICPMS and secondary ion mass spectrometry

Dear Editor and Authors,

Raitzsch et al. present an interesting and timely manuscript about a comparative study of B isotopes in the benthic foraminifera *Cibicidoides wuellerstorfi* analysed using LA-MC-ICP-MS and SIMS. Despite the relevance of $\delta^{11}B$ as a paleo-pH-proxy, very few studies have been published showing intra and inter foraminfera test (shell) $\delta^{11}B$ variability as these analysis have proven to be challenging due to low B concentration and fragility of foraminifera tests. This study provides an interesting comparison between different heterogeneity levels within and between individual foraminfera which will be of widespread interest and should be published after revision of the issues listed below:

Main comments:

1. Data processing:

The manuscript lacks a general explanation of how the data were treated after collection.

Fig. 2 shows a typical time-resolved laser ablation profile for a clean and a contaminated (clay filled) foraminifer. In the caption, it is mentioned that some points have been removed from the ablation trend by a 2-sigma outlier test, however in the methods there is no explanation of the data processing involved.

It would be important to mention the general data reduction routine that was employed.

Furthermore, the ablation intensity profiles appear very bulgy and do not present apparent plateaus. Please report how the shell signal was extracted from the rest?

2. Sample size estimation:

The estimation of the required sample size to resolve 0.1 pH unit is a very important part of the manuscript but the R function "combn()" used for that purpose lacks a detailed explanation in the manuscript – in addition it is unclear if the presumptions made in the manuscript are correct or lead to an underestimation of the required sample size.

In detail:

On line 237 it is reported that the sample size simulation is based on the assumption that the entire population (P) consists of the 18 shells analysed. Although this holds true for this particular study it is not a representation of the actual (true) population size which is what future studies would be interested in to estimate required sample sizes. In other words, the presumption of P = 18 holds only true within this study but has no real world application. Instead it should be discussed what population sizes are realistic within similar pH-environments and simulations should be based on these.

Furthermore, it is not clear how the simulation are carried out using the 18 shells as they have not been measured in the same way according to the Supplemental Material. The "large crater" was

analysed on 16 shells and the "umbilical knob" was also analysed on 16 shells suggesting that for the simulation using 18 shells two different measurement "types" were merged which further complicates its validity. It would be more informative to separate the two and report required sample sizes based on measurement type i.e., for measurements on the "umbilical knob" and for "large crater" measurements.

Considering the sample size of 16 or 18 it appears to be useful to consider the use of a conventional sample size estimation approach in comparison to a resampling approach as drawing from the same small population may result in errors. In the figure below, the estimated sample sizes required for e = 0.5 (2SD) and 1-α = 0.05 in relation to the population size is given as estimated by the R function "sample.size.mean()" (https://CRAN.R-project.org/package=samplingbook) for both measurement "types". Given an overall population size of e.g., 500 specimens in the same pH-environment, it would require n = 87 specimens if the "umbilical knob" was measured and n = 40 if the "large crater" was measured (based on the variability observed in this study) to achieve the desired significance level. Even if the population size consisted of only 16 individuals, the estimated sample size would be n = 16 and n = 14, respectively and thus twice as large as reported in the manuscript.
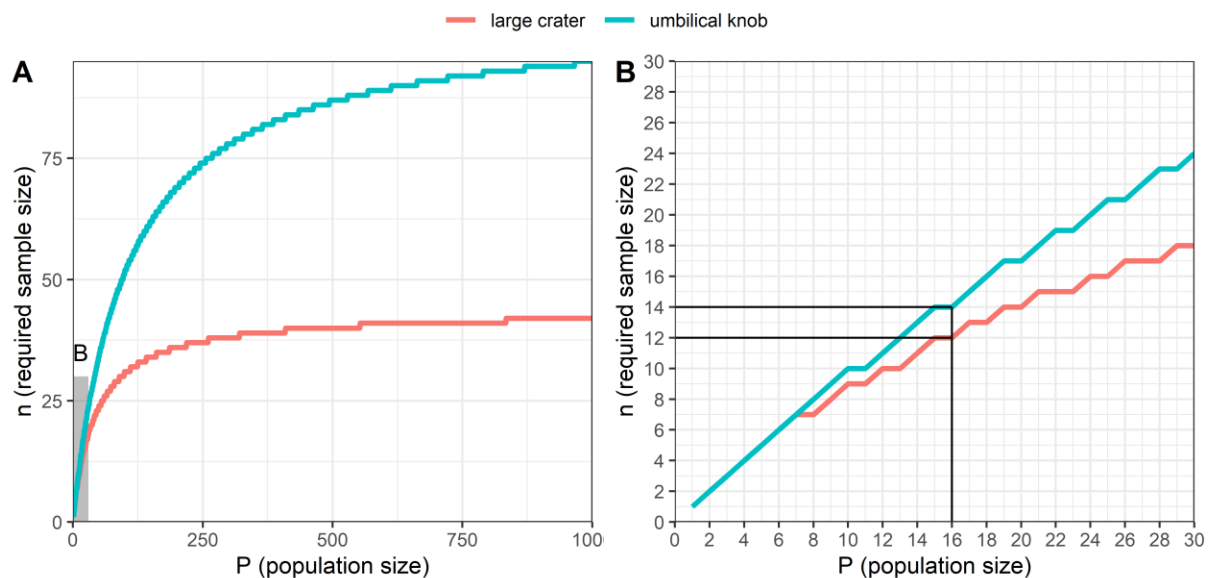
Minor comments:

This is a non-comprehensive list of minor issues

- Line 30: Consider removing the last clause of the abstract. "Vital effect" is a loaded term and since it is not further discussed (Line 185) of little value for this manuscript.
- Line 35: Space missing between 27.2 and ±0.6 ‰
- Line 57: Comma missing after "Also"
- Line 69: Considering that this study looked at a total of 23 specimens the term "tens of specimens" seems excessive, better report the actual number of individuals.
- Line 179: Why was a non-parametric test used? Please specify what data the test was used on? Please report the Wilcoxon-Mann-Whitney test summary i.e., (W = XXX, p <0.001)
- Line 184: Space missing between "large-scale" and "suggesting"
- Line 186: "Somewhat" not useful, report how much $\delta^{11}B$ was elevated in the umbilical knob
- Line 197: a total of 18 shells "were" used

Sincerely,

Dennis Mayk

Department of Earth Sciences
University of Cambridge
Downing Street
Cambridge
CB2 3EQ
United Kingdom