Dear authors,

Please find attached some comments regarding your revised manuscript.

I have three main points of critique:

- The presentation of results should be improved, to make it easier for the reader to grasp the findings of the study. Sometimes results are exclusively shown in the appendix with little or no information about these findings in the main part of the manuscript (e.g., L294-295), sometimes findings are presented for a first time in the Discussion (e.g., L355-359).
- It should be ascertained that readers who are not fully familiar with all the previous work (e.g., L419) or how the statistical analysis was done or can be interpreted (e.g., Appendix-4 to 7) can follow the line of argumentation. Please accommodate more of an outside view when revising by providing more of an explanation.
- It is clearly important to discuss the implication of the variability in tapping force regarding the stability interpretation of tests relying on hand tapping. However, this section needs much more of a balanced discussion combining findings (L364-365) with previous research on this topic.

In the following, please find some more detailed comments including on the three points mentioned before.

I hope this feedback helps to improve the manuscript.

Comments:

L12-13: Consider removing “Peak forces and loading rates are the metrics chosen to quantitatively compare the data.” This statement seems redundant as the two metrics are used in the following sentence, and as no additional explanation is given in this sentence.

L13: Consider mentioning that peak force approximately doubles from one loading step to the next. This seems to be a key finding.

L14: “Significant” overlap – Does this refer to statistical significance in the overlapping proportions? From L357 I understood that the distributions are significantly different, but there is no mention about the overlapping proportions being statistically different? Consider reserving “significant” only when referring to statistical significance, and reword using something like “considerable”, or similar. It may also pay to mention the approximate proportions of overlap in brackets.

L16: introduce ECT abbreviation, when first used.

L22: Correct, but a bit strange that the concept of snowpack stability is introduced in the sense that it has been modelled in this way. What about: “Snowpack instability describes the propensity for a slope to avalanche (CITE). Failure initiation and crack propagation are key components of the avalanche release process (CITE).”

L24-26: It is certainly correct that you state this but it interrupts reading the introduction. Consider rephrasing or moving to a different section.

L30: Is this really the reason why stability tests were invented. Please cite the respective work where this is written. Otherwise, please rephrase more along the line that snow stability tests can support decision making in case of conditional stability (e.g., Birkeland et al, 2023).

L34: Consider to start the sentence with “In contrast, in situations…”

L35: Consider explaining that this division is based on informational entropy, by emphasizing that the mentioned signs (L34) allow direct interpretation of instability (class 1), while stability tests provide more indirect information (class 2).

L69: “simulate portions of” – maybe rephrase to “reflect the”

L72: “component of the ECT” – consider adding the CT: “component of CT and ECT”

L127: Please rephrase to accommodate that Griesser et al. (2023) not only performed tests in an indoord setting but also in the field.

L134: “improved measurement device” – please specify relative to what this improvement is intended to be.

L199: Introduce EAWS when first using this abbreviation.

L212-213: Consider removing, as it is stated in the sentence before.

L258-259: Consider removing “After this second… “.

Sect. 2.4: Consider reordering this paragraph in the way you present the results: (1) methods used to compare tapping force, (2) methods used to analyze explanatory factors. Please explain in a little more detail what ANOVA actually compares and provide a reference. (for instance, I had to look it up as I rarely use this technique). Consider bringing the second sentence first, then the first sentence. Provide references to all methods and tests.

L271: Is “trend” the right word for your analysis? Consider using an alternative word.

L272-277: this is essentially a repetition of Table 2. Consider shortening, highlighting one or key facts, and referring the reader to the table for more information.

L276: Consider removing “~0.0%” as you state that it happened once (in about 8000 tests).

L281: remove “(outliers removed using 1.5 times IQR)” as this is already introduced on L272-274

L283: “(wrist, elbow, shoulder)” – consider deleting, as the loading steps are explained several times before

L285: “each load step (i.e. loading step)” – consider removing one of these

Table 3: Explain abbreviation “Std.” when first being used in the manuscript.

L289-292: This is essentially the same as the caption of Figure 4. Consider removing from text or from caption.

L294-295: Why make this statement or add this table shown in Appendix-3 if you don’t explain what is shown? Moreover, I would move this way of looking at the data in a more prominent way to the Results section, as it provides another way at interpreting the data, with results which allow a simple way of summarizing the proportions of agreements and disagreements. - I suggest incorporating this into the main part of the manuscript, but at least to rephrase the current statement giving the reader at least the essence of what is shown in Appendix-3. For instance, findings like “defining non-overlapping band-widths of tapping force based on the distribution of tapping forces shown in Table Table 3 and Figure 4 showed that between 49% and 54% of the taps of a loading step were within this range. However, a share of about 1% of the taps applied a force corresponding to the range of two loading steps higher or lower”. or similar, could be mentioned.

L305: “cannot explain the bulk of the variance” – Please be more specific with regard to how much these factors could explain or consider rephrasing.

L303: Consider renaming the section title to something like “Explanatory factors impacting peak force” or similar.

L303-309: For the reader it is hard to follow the line of argumentation as the reader is being referred four times to different tables in the Appendix. I suggest summarizing these four tables in a compact table shown in the main part of the manuscript. The full details can still be presented in the Appendix. Moreover, there is no explanation regarding the interpretation of the four tables in the Appendix. Please provide more explanation and/or a reading example for one of the tables.

L326: repeat again where the median and IQR values are shown “Using the median metrics along with their 25^{th} and 75^{th} percentiles (Table 3), the force curves idealized as Gaussians are shown in Figure 5.”, or similar.

Section 3.3

In this section, shown in the Results, you almost exclusively describe your approach to derive force curves. Results are mentioned on L332-333 and in Figure 5. Should this be better split up in a part, which belongs to the Methods section, and a part, which is results? Consider moving L349-350 to this section, as it introduces results for a first time.

Figure 5: Consider showing the 6σ intervals in the figure, including the impact duration. You could also show the respective values for the peak force, making it much easier to grasp the key findings you derived from these curves.

Figure 5: You explain that you used the median and IQR values shown in Table 3 to derive the curves. I don’t really understand why the idealized shoulder tap peaks at about 370 N, which seems very similar to the 373 N shown for the median in Table 3, rather than 343 N shown for the median. For wrist and elbow taps it is hard to judge from the figure whether this would also be the case. - Please explain why there is this difference.

4. Discussion: Obviously it is a matter of preference but consider re-ordering the Discussion section in four sections: (a) Explaining the variability in tapping force (currently in 4.3 and 4.5.2), Comparison with previous studies (currently in 4.1), Idealization of taps as Gaussian functions (currently in 4.3), and Implications for practitioners (currently in 4.2, 4.5).

Table 4: nice summary of the findings in the three studies

L340: Important to address these differences between studies. Consider mentioning that the differences between your study and Griesser et al. (2023) are even more astounding as about 62 participants participated in both studies at the EAWS General Assembly (Appendix-1).

L349-351: To me this is a finding. Consider moving to the respective Section 3.3, possibly after the sentence on L333.

L349: “We estimate the” – I really had to search for where you defined how you estimated this (a half-sentence on L316). Please make it easier for the reader (throughout the manuscript) to follow how you reached your results or conclusions, either by pointing the reader to respective parts of the manuscript, or by briefly repeating important facts, or by moving results closer to their definition.

Sect. 4.2 and 4.3: Consider changing the order of these sections and/or merging to one section where you first discuss potential reasons for variations, and then the implications of such variations. This would include discussing of body characteristics and gender, and possibly, the qualitative

observations (L461-466), which may all have partly contributed to the observed variations. It may also be possible to combine the implications of this variability with Sect. 4.5.2, where you again take up this topic.

L354-367: Consider rephrasing this section to something like “Variability in tapping force – implications for stability interpretations”.

L355-359: From my perspective, this belongs to Section 3.1 as you present the results from a statistical test including p-values, which were not presented in the respective Results section.

L360-367: From my perspective, it is absolutely warranted to discuss the impact of the variability in tapping force with regard to interpreting stability test results (ECT or CT, in this case). But please provide a more in-depth discussion, combining your findings, how they translate to practice, and how your interpretation differs or aligns with previous work. Currently, it is hard for the reader to follow your statement: “especially in cases with potentially fatal outcomes”. Explain how conducting a stability test is linked to potentially fatal outcomes. In that context I suggest to emphasize what (hopefully) is obvious “… no tests provide a definitive “go/no go” result. With accuracies of around 80%, tests are obviously not reliable enough to bet your life on them.” (Birkeland et al., 2023) It is also not explained what your statement “our interpretation aligns with the principle of ‘err on the side of caution’” specifically means (ignore the tapping number in test interpretation, I guess?). When discussing the test interpretation, it may also be worthwhile to state that in 21% of the cases two side-by-side ECT in the same snowpit didn’t show the same propagation result regardless of tapping number (Techel et al. 2018). Similarly, Marienthal et al. (2023), who compared many previous studies and analyzed large data sets reported false-stable rates of 0-40% (previous studies) and 16-31% (new data), despite considering primarily the interpretation scheme based on fracture propagation. In other words, not only tapping force contributes to variability but there is a lot of (spatial) variability inherent in localized tests, with a high number of false-stable results, which again supports Birkeland’s statement.

L379-380: This may be one explanation, but it is also possible that such a difference does not exist. – Please rephrase.

Sect. 4.5: important section to translate the findings to practice. Consider naming the section to “Implication for avalanche practitioners”.

L419: You introduce the stuffblock test. Please add reference to the original paper and explain to the reader how this test works compared to CT or ECT.

L425: Consider changing from “limit” to “Revisiting the stability interpretation of CT and ECT”

L435: this simple, binary interpretation is widely accepted and not debated even by proponents of finer-scaled interpretation schemes (e.g., Techel et al., 2020 (p. 1949): “Quite clearly, whether a crack propagates across the entire column or not is the key discriminator between unstable and stable slope.”). Consider/mention also that in your example, ECTP20 would be considered as “unstable” and ECTP24 as “intermediate” stability (but not as “stable”) by Winkler et al. 2009. In other words, it is just a more gradual interpretation, with shades of grey in between rather than black and white what you seem to propose. Clearly, simplifying has advantages, but also comes at the cost of losing information. Therefore, providing a more in-depth discussion on the benefits/drawbacks of either approach is important when questioning whether three different loading steps are needed, and when proposing to revisit the ECT interpretation.

L461-466: I am not sure whether these qualitative observations are a limitation rather than an actual finding, which possibly may have impacted the force measurements? Consider moving this to the section where you discuss the variability between participants.

L495: Rephrase the “we” in the two questions to “avalanche practitioners” as done on L17-18 as the reader is not necessarily part of the “we”.

References

Birkeland et al. (2023): https://arc.lib.montana.edu/snow-science/objects/ISSW2023_O9.04.pdf

Griesser et al. (2023): https://www.cambridge.org/core/journals/annals-of-glaciology/article/stress-measurements-in-the-weak-layer-during-snow-stability-tests/26DBD00E7309A4EF1C50F9FED88186F7

Techel et al. (2020): https://nhess.copernicus.org/articles/20/1941/2020/

Marienthal et al. (2023): https://arc.lib.montana.edu/snow-science/item/3009

Winkler and Schweizer (2009): https://www.wsl.ch/fileadmin/user_upload/WSL/Mitarbeitende/schweizj/Winkler_Schweizer_Stability_tests_CRST_2009.pdf