Review of *How hard do we tap during snow stability tests?* by H. Toft et al.

1 General remarks

The study analyzes the (un)reliability related to the force applied during hand tap tests by almost 300 avalanche practitioners, using a device which simulates performing a snow stability test like a Compression Test (CT) or Extended Column Test (ECT). Beside the analysis of force measurements, the study also briefly addresses potential explanatory factors for observed variations in the applied force (like body height). Based on the results, the authors propose that (a) the instructions for performing these tests should be revised and that (b) the interpretation of test results should be revisited.

The subject matter of the paper primarily appeals to the snow avalanche community, particularly to avalanche practitioners. While the paper is overall generally easy to read, there is a need for reorganizing its structure. Certain sections, such as the Introduction or Methods, lack essential context or references. Additionally, parts of the discussion tend to repeat information already presented, and in some instances, they offer background that would have been beneficial if introduced earlier, i.e. in the Methods section. The technical setup, the extraction and analysis of the force measurements are clearly described, and the applied methodology seems appropriate.

The subject matter is not novel as, recently, two studies addressed practically the same topic using a (rather) comparable approach (Sedon, 2021; Griesser et al., 2023, summarized below). However, the presented study has the advantage of relying on a likewise large sample ($N = 286$ vs. $N = 69$ and $N = 62$, respectively). While the similarities in study design and results clearly decreases the novelty of the research, from my perspective, the study still warrants publication, if it were to focus more strongly on the (un)reliabilty of applying force when performing these tests, and on how to mitigate these effects. While the authors do propose potential ways forward, their propositions unfortunately remain vague rather than providing more specific, applicable recommendations. However, I feel that the authors are actually in a good position to provide data-driven actionable recommendations based on the results obtained from their own comparably large data-set, which is fully supported by the findings from the two other studies. This would considerably strengthen the submission, by adding novelty and by linking science to practice. To reach this objective, a number of changes would be necessary in the manuscript, which I explain in more detail below.

1.1 General comments

1.1.1 Previous research and specification of research gap.

My most important remark relates to the similarity of the research with the two previous studies:

- Sedon (2021) used a setup very similar to the one described here, where force was measured directly on the force measuring device. Sedon (2021) also explored the correlation between the applied force with explanatory factors like the persons' height, anger, or hand. Sedon (2021) made some propositions on how to increase consistency, as for instance by applying force dropping a ski pole from a specified height rather than using the hand to tap.
- Griesser et al. (2023), which was also co-authored by the main author, focused on measuring the stress observed in the snow, but also presented some measurements where force was applied directly to the device. The main differences between these parts of the studies seems to be the device used for measurements and the number of participants. Moreover, the data presented here was also used by Griesser et al. (2023) to compare the two devices. Moreover, Griesser et al. also explored variations in tapping force as a function of the persons' body features.

From my perspective, the methodologies (though using different devices) and findings are remarkably similar between these studies, even though absolute force measurement values differ (due to the device, I guess). For instance, the findings presented in Figure 4 are similar to the findings presented in Figure 10 in Griesser et al. (2023), and in Figure 2 in Sedon (2021). The same seems true for factors potentially explaining the variation between participants (like a persons' body height). In other words, there is considerable overlap between these three studies. Unfortunately, these studies are either not referred to (Sedon, 2021), or are only briefly mentioned (Griesser et al., 2023, L112-114). However, introducing these previous studies in greater detail is necessary to specify the research gap, and hence the research questions, and to evaluate the presented findings with regard to the mentioned previous research.

As I said above, I feel the authors could turn the similarity between studies into a strength of this submission. I provide more recommendations below.

1.1.2 Title

The title assumes that the reader is part of the "we"? How about something like: "Addressing the (un)reliability of force applications during snow stability tests".

1.1.3 Introduction - Section 1

The Introduction seems to assume that the reader is already fairly familiar with stability tests, what they are and how they work. However, most NHESS readers are likely not familiar with these tests. Therefore the Introduction requires more general background.

For instance, the introduction starts with the definition of snowpack stability (L22) rather than by introducing avalanche hazard in a more general context, leading to the factors, which define avalanche hazard (like snow stability), followed by what

snow stability is, and how it can be assessed. The latter point is addressed on L36-41. In contrast, I am not sure whether introducing the four classes of snowpack stability, the matrix used by forecasters to assess the danger level, and an example for using stability classes in the matrix (L26-34) is really relevant for the topic of increasing the reliability of hand tap tests.

Two stability tests, the Compression Test (L49) and Extended Column Test (L54) are introduced, without explaining what these tests aim to detect, and how this is being achieved. Regarding the first of these two points, Birkeland et al. (2023) provides a nice summary of the questions being addressed when performing these tests, which may be useful: «1) Is there a slab over a weak layer?, 2) Can we initiate a failure in the weak layer?, and 3) Will the crack propagate?» Moreover, a description and figure displaying the dimensions of these two tests would be helpful before listing the tapping instructions. Such a figure could also facilitate explaining the compression of snow when tapping, fracture initiation in a weak layer, and fracture propagation.

L48: please make the distinction between Rutschblock test and CT/ECT clearer (one is loaded by the weight of a human, the others are hand tap tests).

L63-83: Three examples for tapping instructions are provided. I suggest mentioning that these are examples, as instructions may vary even more, as, for instance in Switzerland, where the instructions are simply: «The blade of the avalanche shovel is placed on the block on one side and successively loaded with 10 hits each from the wrist (01-10), the elbow (11-20) and the shoulder (21-30).» (Dürr and Darms, 2016, p. 46). Consider shortening the instructions, as the American description is included to 100% in the Canadian instruction, though the latter contains an additional sentence for elbow and shoulder taps (which you could highlight using italics, if you like).

L84-102: It is certainly helpful to provide a brief summary of current approaches to interpret test results. However, I wonder whether the level of detail regarding the interpretation schemes proposed by Winkler and Schweizer (2009) and Techel et al. (2020b) is necessary for this study, or whether this could be summarized with fewer words. Again, it may be helpful to refer to the recent summary of stability tests by Birkeland et al. (2023). There are also some studies, which explored the repeatability of obtaining a similar test result at the scale of a snow pit, as for instance for the CT, which is most related to your setup (Schweizer and Bellaire, 2010).

L109-114: Please provide more detail on this previous research as it is highly relevant for your study, permitting you to address open questions and or deficiencies in these studies, and, thus, specifying the research gap and, consequently, your research questions.

L116-119: I suggest rephrasing the objectives of your study given my previous comment.

L118-119, L121-126: When reading the manuscript, it was rather surprising that you suddenly address the objective of providing data for mathematical modeling of stability tests, which was not mentioned before. If this is really an objective please be more specific why it would be necessary to model stability tests, and where the research gap is, which you are trying to address. Introduce the research gap before you bring this objective.

1.2 Methods - Section 2

Section 2.1 The device: tap-o-meter introduces the technical details regarding the developed device. I wonder whether L308-317 could be moved from the Discussion and could be integrated into the respective paragraphs in this section. I feel this applies also to L298-306, which, however, could be shortened considerably.

Section 2.2 Data collection process L163-166. Please provide further details: Did you provide instructions prior to tapping experiments? If so, what were these instructions? Please mention the date and number of participants at each of the events, maybe in a small table. Show the total number of participants (if I am not mistaken, you only mention this number in the abstract?). After all, the sample size is a strength of your study.

Section 2.2 Data processing Mention in Figure 2 that this is an example. Consider removing the second and third sentence from the caption, as this information is provided in the text (L208-209).

I feel that *Section 4.1.3 Metric Selection* could be integrated into *Section 2.3*, as it describes why peak force and loading rate are chosen and not other metrics, which are introduced in this section.

I suggest adding another subsection titled *Statistical analysis*, or similar, where the modeling approach could be described (L255). In this section you could also introduce the statistical test you are using, and what p-value is considered as significant.

1.3 Results - Section 3

1.3.1 Tapping force

Currently, in the text (L220-240) you repeat most of the results, which are also shown in Tables 1 and 2. As a reader, I feel this is not very informative. Instead, I propose to select and highlight the key findings like that the median force doubles from one class to the next. This doubling of force values from one class to the next is an interesting result. It is also fully in line with Sedon (2021) and Griesser et al. (2023). I would make this much more obvious to the reader.

I propose to combine L220-240 with Section 3.1, as it all relates to force measurements.

Regarding these results, which I believe represent the key findings of the study, I allow myself to propose an alternative way to present the results (Figure 1). Maybe a Figure similar to this could assist in summarizing the results, when proposing a «normal» or typical force for each tapping level, and, may thus help to highlight deviations from the «normal» , and what such deviations may mean: You have a data set comprising about 2800 taps per tapping class, from avalanche professionals from different countries (which I would emphasize as an additional strength of the data set). Given your data, you could define a «tapping norm» according to the force applied by the majority (you already mention something like this on L359), by choosing optimal thresholds between classes *wrist*, *elbow*, and *shoulder*. In your case, these may be close to the IQR shown in Table 1 and Figure 4. In addition, there could be two classes with taps lower than the desired minimal value for *wrist* taps (<*wrist*) and higher than the desired maximal value for *shoulder* taps (>*shoulder*). You could then show the proportions of observed cases in each bin. - This would be a very visual way to show the distribution of the data, the frequency distribution and magnitude of «errors» (or variation) in terms of their correspondence with a majority force «norm». Moreover, you could use such a figure to describe the proportion of participants who had all their taps lower (i.e., cells with red cross) or higher (cells with blue

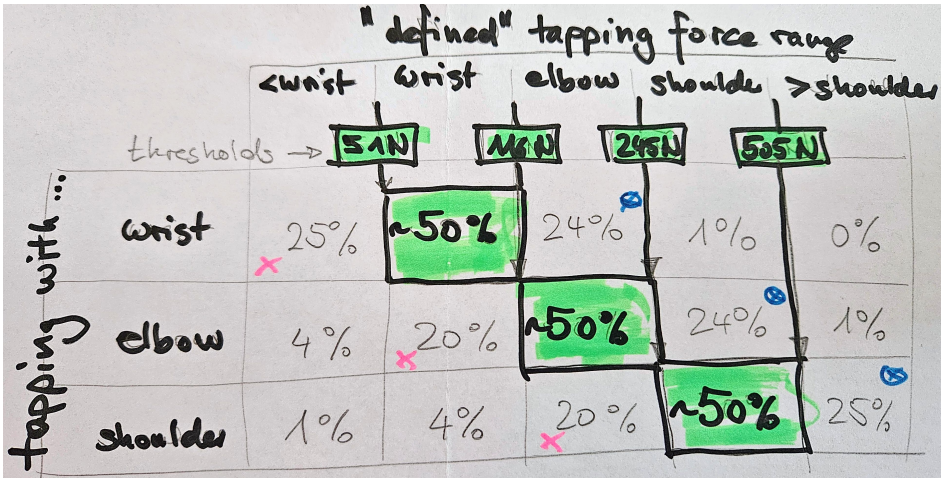

Figure 1. Proposition to visualize, analyze and discuss results on tapping force. Cells could be highlighted to mark the force, which should ideally be applied.

circle) than what you would define «normal». In other words, participants tapping with a force falling into the red-cross cells, essentially lack an entire level of force values when performing the test, while participants tapping with force values in the blue-circle cells, add ten taps harder than the force normal for shoulder taps. (maybe again twice as hard?). This will obviously impact snow compaction, and, hence, the force exerted on potential weak layers.

1.3.2 Survey

In Section 3.2, the method of multivariate regression is mentioned (L255). Please introduce this method with sufficient detail in the Methods section.

On L361-370, some results are described, e.g. «... weight and height are significantly and positively correlated with tap force...». However, no numbers are presented. If possible, please provide more detail on the results.

As the model(s) seemed to have little explanatory power (though again no numbers are shown) (L366), a bi-variate analysis was performed. I suggest to compile the results from this analysis in a table, showing, for instance, median values for each analysed group and tap level, and whether these were significantly different. This table could either be presented in the main part of the paper, or in the Appendix. Presenting these findings in a more accessible manner will also distinguish your study from Sedon (2021) and Griesser et al. (2023), who also primarily mentioned these results briefly in the text. Moreover, it will be easier to refer to specific results when taken them up and comparing them with the respective findings presented by Sedon (2021) and Griesser et al. (2023) in the Discussion section.

1.4 Discussion - Section 4

1.4.1 Discussion of limitations due to force measuring device

In Section 4.1, the force measuring device is being discussed. Several paragraphs could probably be moved to the Methods section, as proposed before. Instead, it could be discussed that the use of different devices likely leads to differences in absolute force values in the three studies.

To decrease variability in applied force during tapping, you propose that a training device could be developed, providing participants with feedback on their tapping during training sessions (L406-408). This is a useful proposition. However, as there is no standard for building such a device, potentially leading to variations in measured force due to the device, it would be valuable if you could provide appropriate drop heights using a specific weight to obtain the median force values obtained for the three tapping classes. If new devices were being built, their sensitivity to force measurements could be compared to your device, which would allow integrating your recommendations on typical force values associated with *wrist*, *elbow*, and *shoulder* taps.

To achieve more consistent tapping in the field, you briefly mention that known weights could be used (L408-409). This is also in line with the proposition by Sedon (2021). Again, it may be worthwhile to provide specific recommendations, like the actual drop height of a ski pole, needed to achieve a desired impact force.

1.4.2 Comparison with findings in other studies

I suggest to include a subsection where you compare your results with the findings in Sedon (2021) and Griesser et al. (2023). For instance, you could show a small table with the median force values from the three studies. As all three studies show an approximate doubling of the force from one tapping class to the next, you could propose an addition to the instructions (L66-82) by a statement like «Tapping should be about twice as hard for elbow taps as for wrist taps.»Ṫhis would be a very practical advice, and, potentially, be more useful than the Canadian description cited on L72-73.

I also suggest to briefly summarize and compare the findings from the corresponding surveys relating age, body size, ... to tapping force. Again, this may potentially lead to some recommendations for practice.

You introduce the instructions taken from the observation guidelines in detail, which had me expect that there are differences between U.S., Canadian, and Norwegian avalanche professionals. However, this didn't seem to be the case. Please take up and discuss this finding. What does this apply for the interpretation of test results?

On L15-17, L370-371, and at a few more places, you emphasize that the variability of tapping force between participants questions the reliability of these results. Please discuss this in more detail. Please also provide more detail on why your interpretation of the results is different to Griesser et al. (2023, p. 6) who obtained very similar results but concluded: «We could show that the differences between different test persons was surprisingly narrow, ...» I feel this different view on results is particularly interesting as the main author participated in both studies.

1.4.3 Discussion of impact force

L355-357: You mention that in Canada, there is no advice on how hard one should tap. However, neither of the other two instructions shown on L63-83 state these. Please rephrase.

L358-359: As outlined above, instead of mentioning that such an approach could be undertaken, I propose to do so, and show the results. I feel this would considerably strengthen your manuscript.

1.4.4 Ways forward

To me, *Section 5.1 Calls to Action*, currently in the Conclusions, should be moved to the Discussion section.

Assuming that you show more clearly the limitations due to variability in tapping force, I agree that it is important to propose result-driven ways forward. Consider whether making it more clear that the proposition of reducing tapping force variability through the use of training and an appropriate tool (Sect. 5.1.1) is the more relevant proposition compared to the second way forward. As mentioned before, I also suggest taking up the proposition by Sedon (2021), who proposed dropping a ski pole or another piece of equipment normally carried in the field from a certain height to achieve more reliable tapping.

I believe that your data, combined with the two other studies, allows you to make recommendations to the tapping instructions. Doing so would make a nice link from science to practice.

Regarding your second proposition, - reducing the interpretation of test results by excluding the tapping force (L410-419), please provide a more thorough discussion. Please discuss why this proposition may be warranted, and why not. For instance, in the two cited studies (Winkler and Schweizer, 2009; Techel et al., 2020b), which relied on Swiss data (dozens of different observers) but also in Techel et al. (2020a), which in addition made use of North American data (probably hundreds of different observers), tapping force showed to be a relevant discriminator, though (clearly) at a much lower level than fracture propagation. Moreover, reducing the interpretation of ECT test results to fracture propagation is prone to misinterpretations too: for instance, Techel et al. (2016) showed that side-by-side ECT showed contradictory propagation results 20% of the time. - For these reasons, please provide a more in-depth discussion of your proposition. And lastly, along this line, it may be worth repeating, what Birkeland et al. (2023) wrote: «Stability tests provide important data for stability evaluations during times of conditional stability. However, no test provides a definitive *go/no go* result. With accuracies of around 80%, tests are obviously not reliable enough to bet your life on them.»

1.4.5 Limitations

I suggest to shorten and move several parts of the Discussion to a Limitations section. This may include, for instance, L326-332.

1.5 Abstract and Conclusions

I suggest rephrasing Abstract and Conclusions after revising the manuscript.

References

Birkeland, K. W., van Herwijnen, A., Techel, F., Bair, E. H., Reuter, B., Simenhois, R., Jamieson, B., Marienthal, A., Chabot, D., and Schweizer, J.: Comparing stability tests and understanding their limitations, in: Proceedings International Snow Science Workshop, Bend, OR, USA, 2023.

Dürr, L. and Darms, G.: SLF-Beobachterhandbuch (Observation guidelines), WSL Institute for Snow and Avalanche Research SLF, Davos, https://www.slf.ch/fileadmin/user_upload/WSL/Publikationen/Sonderformate/pdf/SLF-Beobachterhandbuch.pdf, 2016.

Griesser, S., Pielmeier, C., Boutera Toft, H., and Reiweger, I.: Stress measurements in the weak layer during snow stability tests, Annals of Glaciology, p. 1–7, https://doi.org/10.1017/aog.2023.49, 2023.

Schweizer, J. and Bellaire, S.: On stability sampling strategy at the slope scale, Cold Regions Science and Technology, 64, 104–109, https://doi.org/10.1016/j.coldregions.2010.02.013, 2010.

Sedon, M.: Evaluating forces for extended column tests and compression tests, 127, 39–41, 2021.

Techel, F., Walcher, M., and Winkler, K.: Extended Column Test: repeatability and comparison to slope stability and the Rutschblock, in: Proceedings ISSW 2016. International Snow Science Workshop, 2–7 October 2016, Breckenridge, Co., pp. 1203–1208, 2016.

Techel, F., Birkeland, K., Chabot, D., Earl, J., Moner, I., and Simenhois, R.: Comparing Extended Column Test results to signs of instability in the surrounding slopes - exploring a large, international data set, The Avalanche Review, 39, 24–25, 2020a.

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A., and Schweizer, J.: On snow stability interpretation of extended column test results, Natural Hazards Earth System Sciences, 20, 1941–1953, https://doi.org/10.5194/nhess-2020-50, 2020b.

Winkler, K. and Schweizer, J.: Comparison of snow stability tests: Extended Column Test, Rutschblock test and Compression Test, Cold Regions Science and Technology, 59, 217–226, https://doi.org/10.1016/j.coldregions.2009.05.003, 2009.