APPLYING THE AVALANCHE DANGER SCALE: CONSISTENCY DEPENDS MORE ON THE SITUATION THAN THE FORECASTER

Brian Lazar[1]*, Dylan Craaybeek[1], Frank Techel[2], Simon Horton[3], Mike Cooperstein[1], Simon Trautman[4], Ethan Greene[1]

[1]*Colorado Avalanche Information Center, Denver, CO USA*
[2] *WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland*
[3] *Avalanche Canada, Revelstoke, BC, Canada*
[4] *American Avalanche Association, Jackson, WY USA*

ABSTRACT: Public avalanche forecasts compress complex hazard assessments into a single danger level to guide backcountry recreationists' terrain choices. Danger increases exponentially between levels, so a one-level difference carries significant behavioral and public safety consequences. A decade after Lazar et al. (2016) demonstrated substantial forecaster disagreement, we repeated the exercise with 144 professional public avalanche forecasters across North America, Europe, and New Zealand, each evaluating ten simulated mountain-range-scale scenarios. Two randomly selected forecasters agreed on the exact danger level only 54% of the time (Krippendorff's ordinal α = 0.65, 95% CI: 0.33 to 0.78), though 96% remained within one level. Agreement did not differ clearly by region or experience. Instead, consistency was dominated by scenario complexity (exact agreement ranged from 41% to 75%). In contrast, agreement on the primary avalanche problem was substantially higher, exceeding 80% consensus on half the scenarios, suggesting that disagreement may stem from danger-scale threshold decisions rather than hazard identification. When respondents deviated from the reference rating, they compressed toward the middle of the scale, over-forecasting low-hazard conditions and under-forecasting high-hazard conditions. Pooling independent forecasters and utilizing the group median significantly raised inter-panel reproducibility. A panel of three increased agreement by 13 percentage points (to 67%), with the first two added forecasters improving agreement more than six times as much per forecaster as subsequent additions. However, group aggregation only reduces random individual noise, leaving shared systematic biases intact. We recommend drafting your ratings independently first, and then discuss them as a group. If you are forecasting solo, an external opinion is valuable, whether from a neighboring office or a model, because a team cannot detect a bias its members share.

KEYWORDS: avalanche danger scale, forecaster consistency, inter-rater reliability, judgment aggregation, avalanche problems, public avalanche forecasting

1. INTRODUCTION

The public avalanche danger rating is the most important piece of information in a public forecast, and it impacts where recreationists travel and what terrain they choose (Mannberg et al., 2026). Risk climbs exponentially between levels, so a one-level difference carries major behavioral and public safety consequences (Winkler et al., 2021). Forecasters work from shared frameworks, the Conceptual Model of Avalanche Hazard (CMAH) in North America (Statham et al., 2018a) and the European Avalanche Warning Services (EAWS) guidelines in Europe (Müller et al., 2025), and they still interpret and apply danger definitions differently (Horton et al., 2023; Techel et al., 2026). Lazar et al. (2016) tested that directly, giving professional forecasters identical scenarios and finding substantial disagreement in the ratings they assigned. They concluded that standard descriptors alone were not enough to improve forecaster consistency.

Ten years have passed since the Lazar et al. (2016) study, and operations have since adopted supplemental guidance such as the National Avalanche Center/Colorado Avalanche Information Center (NAC/CAIC) guidance in the United States (bit.ly/4yPnVXu) and the EAWS Matrix in Europe. We repeated the exercise with 144 professional forecasters from North America, Europe, and New Zealand, each evaluating ten simulated mountain-range-scale scenarios, and added agreement on the primary avalanche problem to the 2016 design. We ask whether the disagreement Lazar et al. found has persisted, and what drives it.

* Corresponding author address:
Brian Lazar, Colorado Avalanche Information Center, 1313 Sherman St, Rm. 718 Denver CO 80203
tel: +1 303-499-9650
email: brian.lazar@state.co.us

2. DATA AND METHODS

2.1 Participants

We recruited 144 professional backcountry avalanche forecasters (mean experience 9 years, range 1–30, representing 38 operations), 79 were from the United States (mean experience 8 years; 94% using NAC/CAIC guidelines), 29 from Europe (mean experience 10 years; 100% using the EAWS Matrix), 27 from Canada (mean experience 10 years; 100% using no supplemental guidance), and 9 from New Zealand (mean experience 11 years, 66.7% no guidance; 22.2% NAC/CAIC; 11.1% EAWS).

2.2 Scenarios

Participants assessed ten scenarios representing diverse mountain-range-scale avalanche conditions (Table 1). The scenarios span all five danger levels and avalanche problem types, excluding glide avalanche and cornice-fall situations. Following Lazar et al. (2016), the scenarios were adapted from real operational forecasts across North America with locations anonymized to reduce geographic bias. Each scenario provided forecasters the same three information components: weather observations and forecasts, snowpack summaries and profiles, and recent avalanche activity. Scenario order was randomized to reduce item-order effects.

2.3 Data collection

For each scenario, participants assigned the highest expected 24-hour danger rating and identified the primary avalanche problem type driving their assessment, using either CMAH or EAWS definitions. The survey also collected the respondent's primary forecasting region, years of experience, and the supplemental guidance used operationally. One respondent did not rate S5, giving 1,439 ratings in total.

2.4 Statistical analysis

Danger ratings are ordinal and problem types nominal, so all analyses are non-parametric. Throughout this paper, "consistency" and "agreement" are used synonymously to refer to the concept of forecasters choosing the same rating when given identical information, while "reliability" refers to the formal, chance-corrected statistic (α).

We assess agreement among forecasters in two ways: raw rates and chance-corrected reliability.

Exact pairwise agreement (P_exact) is the probability that two randomly selected forecasters assign the same danger rating to a scenario. This is calculated across all possible forecaster pairs for each scenario and then averaged with equal weight across the ten scenarios. Within-One Agreement (P_within1) is the probability that two randomly chosen forecasters assign danger ratings that differ by no more than one level (e.g., Moderate and Considerable).

Table 1. Summary of the forecast scenarios.

#	Avalanche problem type(s) CMAH / EAWS	Issued danger rating	Brief description
S1	Persistent Slab / Persistent Weak Layers	3 - CON	Moderate loading on a sensitive pwl
S2	Deep, Persistent Slab / Persistent Weak Layers	3 - CON	Incremental loading on an old pwl; slides slowly increasing in size
S3	Storm slabs, Loose Dry/ New snow	1 - LOW	New snow over a frozen or wind-hardened old surface
S4	Loose Wet, Wet Slab / Wet Snow	4 - HIGH	Rain on substantial recent cold, dry snow
S5	Persistent slabs, Storm slabs / Persistent Weak Layers, New Snow	5 - EXT	90 cm storm snow over a facet/crust matrix
S6	Persistent Slab / Persistent Weak Layers	3 - CON	Light snow; sporadic remote triggering of large slabs
S7	Wet Slab / Wet Snow	3 - CON	Sporadic natural wet slabs to size 2–3
S8	Storm, Wind slabs and Loose Dry / New Snow, Wind-drifted snow	2 - MOD	Worrisome structure, only small storm avalanches
S9	Loose Wet / Wet Snow	2 - MOD	Melt-induced wet activity, mostly size 2
S10	Persistent Slab / Persistent Weak Layers	2- MOD	Natural on pwls over; human triggering continues

Note: pwl=persistent weak layer

P_exact and P_within1 count every match, including those that occur by chance. With only five danger levels and most scenarios sitting in the middle of the scale, two forecasters guessing at random would still achieve some level of agreement. To control for this, we calculated

Krippendorff's ordinal alpha (α) (Krippendorff, 2004) to measure inter-rater reliability, utilizing the scenarios as items and the forecasters as raters. This index subtracts chance agreement: an α of 1.0 indicates perfect agreement, 0.0 indicates agreement consistent with random chance, and values below 0.0 indicate systematic disagreement.

To evaluate the effect of group forecasting, we simulated virtual "panels" of size k (odd numbers 1 to 21) using Monte Carlo simulations (60,000 runs per scenario). For each run, we drew respondents without replacement and split them into two independent panels of size k. For each panel, we calculated the panel's median rating. We then measured inter-panel consistency (how often their ratings agreed) and reference agreement (how often they matched the published rating). Respondent pools were treated as fixed, and scenario-specific noise was paired across cohorts.

3. RESULTS

3.1 Agreement and reference rating

Two randomly selected forecasters assigned the same danger level 54% of the time (95% CI: 48–62%). When we consider the levels adjacent to the most selected level, 96% (CI: 94–98%) of the forecasters selected one of these three levels (Fig. 1). Chance-corrected reliability was weak (α = 0.65 [0.33, 0.78]). No scenario reached consensus, and all ten scenarios drew responses spanning at least three danger levels. Across all 1,439 ratings, 2.9% differed by two or more levels from the issued reference rating. Even among colleagues working within the same operation, 86% of centers with four or more respondents had at least one scenario where assessments spanned three levels. Individual assessments matched the published reference rating 54% of the time (784 of 1,439), with the median response matching the reference in six of ten scenarios.

Disagreement with this reference was directional, compressing toward the middle of the scale. Respondents rated Low/Moderate scenarios higher by 0.35 of a level on average, and High/Extreme scenarios lower by 0.75 of a level. While the reference rating is itself a single judgment rather than objective ground truth, this systematic shift highlights a shared tendency to avoid the ends of the danger scale.

3.2 Scenario difficulty

Two forecasters evaluating the same scenario selected the exact same danger level between 41% and 75% of the time, depending entirely on the scenario (Fig. 2). In contrast, no forecaster characteristic (where they worked, years of experience, which guidance they used) produced a statistically significant difference in agreement. The specific scenario, not the individual forecaster, drove the disagreement.

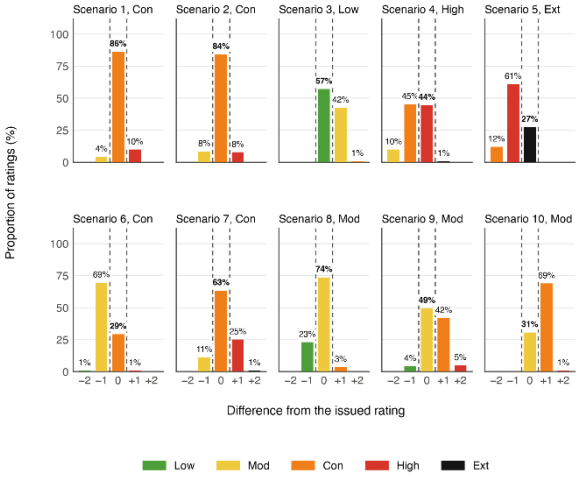

Figure 1. Distribution of danger ratings relative to the reference rating for each of the ten 2026 scenarios (n = 144; n = 143 for S5). Bars show the proportion of respondents at each rating difference. Zero denotes agreement with the reference rating and is framed by dashed lines; its proportion is shown in bold. Negative values indicate lower ratings and positive values higher ratings, with differences of two or more levels grouped at −2 and +2. Colors indicate the danger rating selected (1 - Low to 5 - Extreme), and labels above non-zero bars show the corresponding proportions.

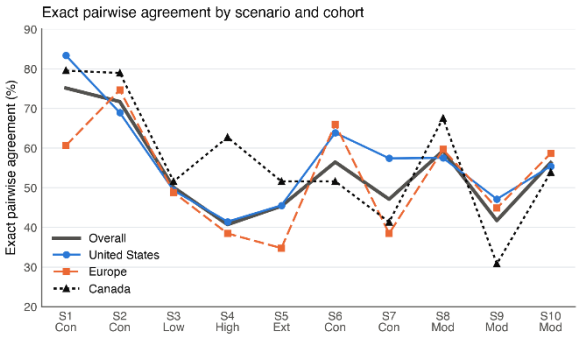

Figure 2. Exact pairwise agreement (P_exact) for each scenario, for all respondents and for the three largest geographic cohorts. The issued danger level appears beneath each scenario label. New Zealand (n = 9) is omitted.

The two scenarios with the highest agreement rates on the danger level (P_exact 72%–75%; S1, S2) featured persistent slab situations, with respondents agreeing strongly on the problem (P_exact 76-82%). In contrast, we observed

lower agreement rates in scenarios S4, S5, S7, and S9 (P_exact 41%-47%), with responses being split almost equally between two adjacent danger ratings. These involved wet snow situations or complex multi-problem conditions.

3.3 Geographic cohorts

We found no statistically significant difference between geographic regions. Exact pairwise agreement (P_exact) and chance-corrected reliability (α) did not differ clearly by region (United States: 57%, α = 0.70; Canada: 57%, α = 0.69; Europe: 53%, α = 0.60; New Zealand: 50%, α = 0.55; all respondents: 54%, α = 0.65).

As noted in Section 2.1, operational guidance and geographic region are almost perfectly confounded in our sample, preventing a clean separation of these variables. Neither geographic cohort, experience, nor guidance framework accounted for the observed variation in danger rating consistency.

3.4 Pooling forecasters

Pooling independent assessments reveals a stark divergence of reproducibility between independent panels, and agreement between those panels and the published reference rating.

Inter-panel agreement rises continuously with panel size, but the benefits are front-loaded. Adding the first two forecasters increases agreement by 6.5% per forecaster, about six times more than adding subsequent forecasters.

3.5 Avalanche problem types

Agreement on the primary avalanche problem was substantially higher than agreement on the danger level (Fig. 3). Respondents using the nine CMAH problem types agreed on the primary problem 63% of the time, against 68% among respondents using the five EAWS categories. Mapping the CMAH responses onto the EAWS categories raised agreement to 73%. Chance-corrected, the sets are really close (α = 0.50, 0.52 and 0.59), indicating that the apparent difference reflects the number of categories rather than genuine disagreement. Both CMAH and EAWS problem type frameworks exceeded 80% consensus on half of the scenarios, reaching near-unanimity on straightforward persistent slab (S10) and wet loose snow (S9) situations. This indicates that forecasters generally agree on which specific avalanche problem is decisive, even when they disagree on the associated danger level. The low-danger, new-snow case (S3) was the exception. On S3, CMAH responses spread their selections across eight distinct problem types (with No Distinct Problem leading at only 31%), while EAWS responses split almost evenly between Persistent Weak Layers, No Distinct Problem, and Wind Slab (25% each). It is the one scenario where forecasters disagreed about what they were looking at rather than about how serious it was.

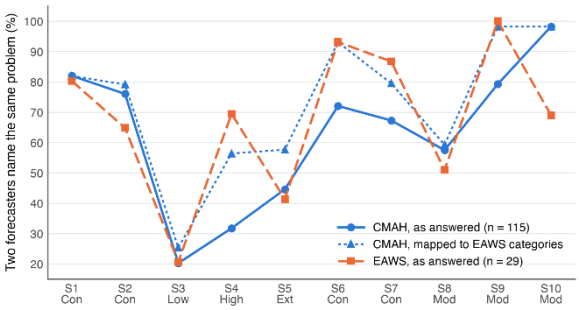

Figure 3. Exact pairwise agreement (P_exact) for each scenario, for all respondents and for the three problem categories. The issued danger level appears beneath each scenario label.

4. DISCUSSION

To contextualize our findings and translate them into actionable changes for public forecasting operations, we summarize the core themes, statistical evidence, and practical takeaways of this study in Table 2.

Table 2. Synthesis of key findings, empirical evidence, and operational takeaways for regional avalanche forecasting services

Core Theme	Key Finding	Empirical Evidence	Operational Takeaway
Situational Variance	Specific scenarios drive disagreement more than regional cohort, guidance type, or experience.	Exact agreement ranged from 41% to 75% across scenarios.	Target calibration by avalanche situation
Judgment Aggregation	Pooling independent assessments reduces individual noise but leaves shared bias.	Pooling 3 forecasters raised agreement from 54% to 67%.	Operations deciding how to staff a shift, 3 is the number that matters.
Compression Bias	Ratings consistently compress toward the middle of the scale.	Low/Mod rated 0.35 higher; High/Ext rated 0.75 lower.	Train for and measure bias compression of 5-tier scales
Dual-Pathway Model	Guidance and collaboration address different parts of the problem.	Problem consensus is strong (>80%), but danger consensus is weaker (54%).	Combine guidance with teamwork.

4.1 *Comparison with 2016 and other studies*

The 2026 scenarios were designed to represent situations similar to those used in 2016, but they were not identical, so the two surveys are independent. Nevertheless, the main patterns were remarkably similar. Exact pairwise danger level agreement rose from 48% in 2016 to 54% in 2026, a difference that the bootstrap cannot separate from zero (+6 percentage points, 95% CI: −5 to +17), and chance-corrected reliability was essentially unchanged (α = 0.65 vs. 0.66).

The primary difference was that the group median matched the issued reference rating less frequently in 2026 (6 out of 10) than in 2016 (9 out of 10), which reduces the impact of pooling on reference agreement. This difference likely reflects danger-level threshold decisions for specific scenarios rather than a shift in underlying forecaster consistency, though our data cannot isolate whether individual deviation from the reference also increased.

Our 54% survey baseline is substantially lower than the 75% to 87% agreement rates observed in active forecasting operations (e.g., Techel et al., 2024; Durbach et al., 2026). This gap highlights a well-documented characteristic of survey-based assessments. Across diverse fields such as medical diagnostics (Peabody et al., 2000), naturalistic decision-making (Kahneman & Klein, 2009), and scenario planning (Wright & Goodwin, 2009) survey evaluations routinely produce lower agreement than live operational practice. In our survey, participants worked in isolation without real-time field observations, seasonal snowpack context, or peer discussion. In operational forecasting, these sources of information and interaction provide opportunities to reconcile individual assessments.

Operational settings show a similar pattern. Forecasters working under the same roof often align well, but that breaks down when comparing across operations. For example, in the European Alps, adjacent regions managed by the same forecasting center showed about 90% agreement on danger levels, but agreement dropped to 60% when crossing organizational or national borders (Techel et al., 2018). Similarly, in western Canada, different agencies often diverged on how they identified and prioritized avalanche problems (Statham et al., 2018b). These studies suggest that organizational context can increase local consistency, while differences remain across organizational boundaries.

4.2 *Possible causes of inconsistency*

Three main factors may explain why forecasters disagree.

First, disagreement was concentrated near danger-level boundaries, with responses often straddling two adjacent danger levels rather than spreading across the scale, suggesting that forecasters largely interpreted the situation similarly but differed about where they placed the threshold. The danger scale presents five distinct danger choices, and we cannot distinguish confident disagreement from boundary-line uncertainty. Some warning services address the continuous nature of avalanche hazard by expressing tendencies within a level, or when assessing the

input factors to decision tools like the EAWS Matrix (Techel et al., 2026). Making a tendency explicit leaves the underlying danger scale unchanged.

Second, disagreement was directional. Ratings consistently compressed toward the middle of the scale, mirroring a central tendency bias in five-tier scales that has been well documented in other survey data (e.g. Garland (1991), Douven (2018)). This middle-of-the-scale bias is a shared behavior under survey conditions. Pooling responses only cancels error that varies independently between people, and averaging forecasters removes noise; it does not remove a tendency they all have.

Third, the structure of our assessment frameworks forces a trade-off between detail and consensus. Previous studies have demonstrated that providing more decision categories increases variance across independent estimators (e.g. Regier et al., 2007; Preston & Colman, 2000). Higher-resolution systems like the nine-category CMAH problem framework allow forecasters to choose problems with more precision, but naturally lower consensus. Lower-resolution systems like the five-category European framework increase agreement simply by grouping distinct physical situations under broader categories, but reduce precision.

4.3 Practical recommendations

Our results demonstrate that aggregating independent individual assessments is an effective tool for reducing forecaster noise. To maximize the benefits, forecasting operations should consider the following:

Where resources allow, pooling two or three independent forecasters reduces individual noise and raises the reproducibility of the group rating, with the first additions contributing most. If agreement between individual ratings is low, moving from a single forecaster to a panel of two or three raises exact pairwise consistency, in our case from 54% to 67% (+13 percentage points). Higher initial agreement would mute this impact.

Have each forecaster record a rating before the group discussion. The gain depends on those inputs being independent, and assessments made after discussion are reached through discussion. In Switzerland, the two or three forecasters on duty each assess alone, and those ratings are combined into a starting value for the discussion that follows (Winkler et al., 2024). When ratings disagree, resolve the difference against data and definitions rather than by deferring to the majority, since respondents compressed toward the middle of the scale. We did not test discussion directly; every assessment in this study was made alone.

A single-office team can reinforce a shared bias that aggregation cannot filter out. Operations can seek outside inputs as a remedy. When solo forecasting is unavoidable, secure outside inputs such as peer reviews from adjacent regions, weather-analog days, or objective models (Winkler et al., 2024). To be effective, these assessments must use a different logical path. A tool or workflow that merely reproduces the forecaster's own subjective reasoning adds confidence without adding new information.

Forecaster inconsistency has persisted across a decade in which supplemental guidance was widely adopted. Rather than reading this as evidence that guidance does not work, we interpret it as a sign that guidance and collaboration address different parts of the problem, and that neither alone is sufficient.

Frameworks like the EAWS Matrix or CMAH standardize what forecasters evaluate. These resources establish a shared vocabulary for key ideas and physical inputs (stability, likelihood of triggering, and expected avalanche size), ensuring different offices are at least assessing the same parameters. But a shared vocabulary is not enough. Consistently translating evidence into a single danger level also requires a shared understanding of how to interpret that evidence, how to weigh contributing factors, and where to place danger-level thresholds. This kind of alignment develops through use and through regular exchange among forecasters, not from the framework alone.

It is important to understand that these tools are not meant to be used as algorithms. Techel et al., 2024, 2026 showed that feeding subjective assessments such as stability or avalanche size inputs through a rigid matrix can propagate and amplify small differences in individual judgment rather than resolve them.

4.4 Limitations and future work

Geographic region and operational guidance are perfectly confounded in our sample. Isolating the statistical efficacy of supplemental guidance (like the EAWS Matrix or NAC/CAIC descriptive guidance) from regional forecasting influences would require operations within the same geographic area to use different frameworks, or a longitudinal "before-and-after" design like the Scottish transition (Durbach et al., 2026).

We evaluated consensus and alignment with a published reference rating rather than absolute correctness. The reference rating itself is an expert forecast made under uncertainty rather than verified ground truth. Our results measure consistency rather than accuracy. Future studies could build on retrospective field verification, but even then, a regional danger rating remains a subjective estimate, relying on the same judgmental process as the forecast did.

5. CONCLUSIONS

A decade after our last investigation, professional avalanche forecasters, given identical weather, snowpack, and avalanche information, still choose the same danger rating only 54% of the time, though 96% remain within one level. Inconsistency seems to be related to specific avalanche conditions or situations bordering two danger levels, with wet avalanche situations and complex, multi-problem conditions presenting the most challenging situations. Inconsistency remains despite the widespread adoption of decision-support tools like supplemental guidance. Individual expert disagreement is therefore normal, expected, and unavoidable. The practical question is not how to eliminate it but how to manage it. Operations can measure how much disagreement exists within their own team, combine independent assessments to cancel individual noise, and pair codified guidance with collaboration to align not just the criteria but their interpretation. Disagreement cannot be removed, but it can be understood and reduced.

ACKNOWLEDGEMENTS

We thank all the forecasters who participated in this survey. Without their willingness to complete the exercise thoughtfully, this study would not have been possible. Karl Birkeland also provided helpful review and feedback for the 2016 study.

REFERENCES

Douven, I. (2018). A Bayesian perspective on Likert scales and central tendency. *Physica A: Statistical Mechanics and its Applications*, *490*, 1494–1502. doi.org

Durbach, I. N., Fyffe, B., Morrison, B., Diggins, M., and Ebert, P. A.: Avalanche Forecast Verification and Occurrence in Scotland, 2007–2025, Proceedings of the International Snow Science Workshop, Whistler, Canada, 2026.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, *2*(1), 66–70.

Horton, S., Haegeli, P., Statham, G., Shandro, B., Clark, T., Nowak, S., Towell, M., Hordowick, H., & Herla, F. (2023). Is it a problem? Takeaways from research into the use and effectiveness of avalanche problems. Proceedings of the International Snow Science Workshop, Bend, Oregon, 26-31.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. doi.org

Krippendorff, K. (2004). Content Analysis: An Introduction to Its Methodology (2nd ed.). Thousand Oaks, CA: Sage.

Lazar, B., Trautman, S., Cooperstein, M., Greene, E., & Birkeland, K. (2016). North American Avalanche Danger Scale: Do backcountry forecasters apply it consistently? *Proceedings of the International Snow Science Workshop*, Breckenridge, CO, 457-465.

Mannberg, A., Toft, H. B., Stefan, M. A., Aase, M. J., and Hetland, A., 2026. Do backcountry skiers react to the avalanche forecast? Measuring avalanche terrain exposure using a continuous score, *Cold Regions Sci. Technol.*, 105039, https://doi.org/10.1016/j.coldregions.2026.105039.

Müller, K., Techel, F., and Mitterer, C.: (2025) The EAWS matrix, a decision support tool to determine the regional avalanche danger level (Part A): conceptual development, Nat. Hazards Earth Syst. Sci., 25, 4503–4525, https://doi.org/10.5194/nhess-25-4503-2025

Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: A prospective validation study of quality of care and clinical practice variation. *JAMA*, 283(13), 1715–1722. https://doi.org/10.1001/jama.283.13.1715

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. Acta Psychologica, 104(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partition of color space. Proceedings of the National Academy of Sciences, 104(4), 1436–1441. https://doi.org/10.1073/pnas.0610341104

Statham, G., Haegeli, P., Greene, E., Birkeland, K., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., & Kelly, J. (2018a). A conceptual model of avalanche hazard. *Natural Hazards*, 90, 663-691.

Statham, G., Holeczi, S., & Shandro, B. (2018b). Consistency and accuracy of public avalanche forecasts in Western Canada. *Proceedings of the International Snow Science Workshop*, Innsbruck, Austria, 1492-1495.

Techel, F., Mitterer, C., Ceaglio, E., Coléou, C., Morin, S., Rastelli, F., and Purves, R. S. (2018). Spatial consistency and bias in avalanche forecasts – a case study in the European Alps. *Natural Hazards and Earth System Sciences*, 18(10), 2697–2716. https://doi.org/10.5194/nhess-18-2697-2018

Techel, F., Lucas, C., Müller, K., Pielmeier, C., and Morreau, M. (2024). Unreliability in expert estimates of factors determining avalanche danger and impact on danger level estimation using the Matrix. *Proceedings of the International Snow Science Workshop*, Tromsø, Norway. https://arc.lib.montana.edu/snow-science/item.php?id=3144

Techel, F., Müller, K., Marquardt, C., and Mitterer, C.: (2026) The EAWS matrix, a decision support tool to determine the regional avalanche danger level (Part B): operational testing and use, Nat. Hazards Earth Syst. Sci., 26, 1161–1181, https://doi.org/10.5194/nhess-26-1161-2026

Winkler, K., Trachsel, J., Knerr, J., Niederer, U., Weiss, G., Ruesch, M., and Techel, F. (2024) SAFE – a layer-based avalanche forecast editor for better integration of model predictions. *Proceedings of the International Snow Science Workshop, Tromsø, Norway, 23–29 September 2024*, pp. 124–131. https://arc.lib.montana.edu/snow-science/item.php?id=3123

Winkler, K., Schmudlach, G., Degraeuwe, B., and Techel, F. (2021) On the correlation between the forecast avalanche danger and avalanche risk taken by backcountry skiers in Switzerland. *Cold Regions Science and Technology*, 188, 103299. https://doi.org/10.1016/j.coldregions.2021.103299

Wright, G., & Goodwin, P. (2009). Decision making and scenario planning: Exploring the limits of rationality. Technological Forecasting and Social Change, 76(6), 813–825. https://doi.org/10.1016/j.techfore.2008.02.005