COMPREHENSION VALIDATION OF THE REVISED EAWS DANGER SCALE DESCRIPTORS: LESSONS FROM A LARGE-SCALE EUROPEAN END USER SURVEY ON AVALANCHE RISK PERCEPTION

Philip A. Ebert[1*], Frank Techel[2], Matthias Walcher[3], Benjamin Reuter[4], Igor Chiambretti[5], Lorenzo Bertranda[6], Montse Bacardit Peñarroya[7], Jaka Ortar[8], Mark Diggins[9], Karsten Müller[10], Gerit Pfuhl[11]

[1] *University of Salzburg, Salzburg, Austria*
[2] *WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland*
[3] *Avalanche Warning Service Tyrol, Innsbruck, Austria*
[4] *Météo France, Briançon, France*
[5] *AINEVA, Trento, Italy*
[6] *Meteomont Carabinieri, Italy*
[7] *Centre de Lauegi – Conselh Generau d'Aran, Vielha e Mijaran, Spain*
[8] *Slovenian Environment Agency, Slovenia*
[9] *Scottish Avalanche Information Service, Aviemore, Scotland*
[10] *Norwegian Water Resources and Energy Directorate, Oslo, Norway*
[11] *Norwegian University of Science and Technology, Trondheim, Norway*
** Indicates the presenting author*

ABSTRACT: The descriptors of the European avalanche danger scale, in use since 1993, are being revised. In June 2025 EAWS forecasters agreed on a new provisional wording and, in doing so, provided its content validation. Because the descriptors are first of all a public communication tool, we ran a comprehension validation: a survey testing whether end users read the wording as forecasters intend. It ran online from 3 February to 8 April 2026 in six languages and retained 17,335 responses after cleaning, of which 15,749 were completed. All wording comparisons were randomised between subjects. We evaluated each proposal against three conditions: correctness (do users recover the order forecasters intend?), simplicity (how easy is the task?), and increase (does the wording convey how much the danger increases from level to level?). For avalanche occurrence the main finding is reassuring: all four wording variants were ranked accurately (83–92 % correct), and most differences that reach significance in the pooled data do not survive within individual languages. Only the variant that drops the distribution wording is clearly weaker. On the increase condition the variants do separate. Respondents placed the four statements on a continuous danger scale; wording level 2 as "in some places" rather than "in a few places" moves it closer to level 3 and shrinks the perceived difference between moderate to considerable from 36 % to 31 % of the distance between levels 1 and 4 (Cohen's $d = 0.39$, $p < 0.001$). For size and impact the wording effect is large: adding size class words (small, medium, large, very large) to the impact wording raised accuracy from 61 % to 95 % and self-reported ease from 4.7 to 6.1 (Cohen's $d = 1.04$). Most of that failure comes from using "bury and kill" at both levels 1 and 2, which makes them hard to tell apart. We also report how users read seven phrases that forecasters already use. For avalanche occurrence, the largest differences are between languages rather than between wordings.

KEYWORDS: Avalanche danger scale, risk communication, risk perception, descriptors, multilingual communication

1. INTRODUCTION

The five-level European avalanche danger scale was first introduced in 1993. (For an intriguing history of its creation, see Mitterer and Mitterer, 2018). The forecasted avalanche danger level is derived from snowpack stability, its distribution and the expected avalanche size (Müller et al., 2025; Techel et al., 2026). The descriptors provide additional information that is aimed at professional and recreational users specifying how the avalanche danger is to be interpreted in terms of the ease with which avalanches can be triggered and their expected size (amongst others). The danger scale has stood the test of time as the core instrument for public avalanche risk communication in Europe (Schweizer et al., 2020). However, progress in forecasting, and the introduction of the so-called European Avalanche Warning Services (EAWS) Matrix (Müller et al., 2025) — a tool to standardize forecasting amongst all European forecasting agencies — led to a desire to revise the original descriptors of the avalanche danger scale while keeping the five levels fixed. In June 2025, EAWS forecasters agreed on revisions to the original scale descriptors that included substantial changes to existing columns and the

* *Corresponding author address:*
Philip A. Ebert, Department of Philosophy (GW), University of Salzburg, Franziskanergasse 1, 5020 Salzburg, Austria; email: Philip.Ebert@plus.ac.at

addition of new ones. As such, European forecasters provided a *content validation* for the new descriptors (Müller et al., 2026). However, since the new descriptors are used mainly as a public communication tool, we deemed it necessary to test whether the wider public understood the new descriptors as intended, i.e. to conduct a *comprehension validation*. The paper presents the main insights of our survey.

The focus of the survey was to test two new columns that together describe avalanche danger at each of the five levels: The new avalanche occurrence column eschews the explicitly probabilistic terms of the original scale ("triggering is possible / likely"), which end users and forecasters interpret very differently (Ebert et al., 2025). Instead, it describes how *easily* avalanches can be triggered and the distribution of potential trigger points, mirroring the EAWS matrix guidelines (Müller et al., 2025). The *size and impact* column describes the potential size of an avalanche at that danger level and uses impact language to explain its destructive potential.

End-user comprehension of avalanche forecasts has been studied before: Engeset et al. (2018) assessed how different modes of communication affect understanding, St. Clair et al. (2021) outlined the different ways user groups draw on forecasts' content, Morgan et al. (2023) examined how North American users interpret their danger scale (linear versus exponential increase), and Lucas et al. (2023) investigated user understanding of sublevels in Switzerland. Our overall question was whether the newly proposed descriptors are fit for purpose. Avalanche risk communication should, in the first instance, provide accurate and clear information that aims to strengthen the competence of the people who use it (for an argument in that direction, see Ebert, 2026). This places three demands on the forecaster: the information must be interpreted as intended (correctness), be easy enough for the average end user to comprehend (simplicity), and differentiate the danger levels well enough to convey how the danger grows (increase). We pursued this through three aims: a) a comprehension validation of the originally proposed descriptors; b) a test of specific variations of the proposed descriptors, as a robustness check and to provide additional evidence of how end users understand the descriptors and their variations; c) a test of how end users interpret phrases often used in forecasts, addressing concerns raised by forecasters about whether their own language is clear. The survey was presented in six languages, with English as base survey from which forecasters translated the other versions. As we will see, some of the largest differences we report are between languages rather than between different wordings of the descriptors.

2. THE SURVEY

2.1. What we tested for

A descriptor can fail in more than one way, so we evaluated each proposal against the following three criteria. We focused our survey on testing the descriptors for levels 1 to 4 only.

Correctness. Whether end users recover the ordering of the descriptors as set by the forecasters. The benchmark is deliberately internal. Naturally, a descriptor could describe avalanche conditions poorly and still be ranked correctly, and an excellent description could be misread; this criterion asks only whether end users can intuitively rank the descriptors by increasing avalanche danger, with nothing else to go on. We report this as ranking accuracy: the percentage who recover the intended order.

Simplicity. Whether users find ranking the statements easy. Testing for perceived ease of ranking also offered another robustness check on the correctness condition: a statement ranked correctly but only with effort is fragile, since the effort will not always be made. We measured it by self-reported ease immediately after each ranking. (Confidence was also collected but is not reported as it tracks perceived ease.)

Increase. Whether the wording conveys how much the danger increases from one level to the next. Getting the order right is a weak requirement: it says only that level 3 (considerable) is read as more dangerous than level 2 (moderate), not by how much. Real risk rises steeply with the danger level – by a factor of roughly four to eight per level, depending on method (Winkler et al., 2021; Degraeuwe et al., 2024) – and the step from *moderate* to *considerable* is the one that matters a lot: considerable is forecast less than 30 % of the time but accounts for roughly 50–80 % of fatalities, depending on activity (Winkler et al., 2021). A wording that is ranked correctly but presents the levels as evenly spaced may therefore be less suitable than one that gives more room to the increase from *moderate* to *considerable*. We measured it by asking respondents to place the statements on a continuous danger scale using a slider, which shows not only whether they separate the levels but also how they distribute the increase across them. This criterion was applied only to the occurrence column.

The three criteria can point in different directions, and for the occurrence column they do. Where that happens, we say so rather than collapsing them into a single score.

2.2. Participants

The survey ran online using the Qualtrics platform from 3 February to 8 April 2026. Participating warning services distributed it through their websites, apps, blog posts, and social media. Of the 26,768 responses received, we dropped those flagged by Qualtrics as duplicates or ballot-box stuffing, those with a reCAPTCHA score below 0.5, and those with a total duration below 180 s. This left 17,335 responses, of which 15,749 (90.9 %) were completed; the median completion time was 8.9 minutes. The language split is uneven: German 9,604, Italian 3,761, French 1,939, English 1,782, Slovenian 155 and Spanish 94. Slovenian and Spanish are too small for per-language inference. The survey language was set by the user's operating system, with English as the default and a drop-down menu to change it.

Respondents had a median age of 45 years (IQR 34–56), 77% identified as male, and most lived in Italy (25%), Austria (24%), Switzerland (16%), Germany (15%), France (9%) or the United Kingdom (4%). This is a committed audience rather than a general public sample: 61% had more than ten winters of backcountry experience, 68% spend at least 11 days per winter in the backcountry, and 84% had done some formal avalanche training.

2.3. Design and analysis

Every wording comparison was randomised between subjects, and the randomisers were independent, so the wordings are compared with independent samples. In this paper we report the following: Each respondent ranked one of eight occurrence cells from lowest to highest danger level (four wording variants crossed with two word orders, $n \approx 1{,}880$ per cell; Table 1), presented initially in random order, and then one of four size and impact descriptor sets, again in random order. After each ranking, they rated its ease and their confidence on 1 (difficult) – 7 (easy) scales. To test the increase condition, respondents placed the four statements of one occurrence variant on a continuous scale from "no avalanche danger" (0) to "extreme avalanche danger" (1). Finally, each respondent answered seven short questions about phrases commonly used by forecasters.

We score a ranking as ordered correctly if the respondent placed all four statements in the intended order or in its exact reverse. Crediting the full reversal needs a justification. Its frequency depends strongly on language (German 20.4 %, Italian 11.4 %, French 10.7 %, English 8.3 %) and much less on wording: among respondents who produced either the intended order or its exact reverse (n = 13,249), language contributes a likelihood-ratio χ^2 of 288.9 against 38.4 for the wording variant and 0.3 for word order. Further checks point the same way: reversers placed the same levels on the labelled danger slider in the correct order as often as those who ranked correctly (90.6 % against 89.9 %, compared with 68.2 % of partial-error respondents), reported the task as easy and were as confident, and their rate does not fall with avalanche training or experience. The most plausible reading is that someone who reverses the whole list has understood which statement is most dangerous but mapped that ordering to the opposite end of the response scale. Treating reversals as errors lowers every cell by a similar amount and leaves the two best occurrence versions in the same order, but it swaps the two worst.

For the increase criterion, we express each step as a share of that respondent's own total range, which removes individual differences in how widely people spread the sliders. The middle-gap ratio is (L3 − L2) / (L4 − L1); applied to all three steps, it decomposes the perceived range into shares that sum to one. It is defined only for strictly descending responses, so 12.6 % are excluded. We used logistic regression for ranking accuracy, linear models for ease, and Welch t-tests with Holm correction for pairwise comparisons. Given the sample size, very small differences reach significance, so we report effect sizes throughout. We also performed two robustness checks: adding age, gender and avalanche training as covariates shifts the pairwise version odds ratios by at most 5 % and the predicted accuracy in any cell by at most 1.4 points, so these variables predict accuracy but do not confound the wording effect; and relaxing or removing the filters of acceptable responses shifts the main accuracy figures by at most one percentage point.

3. AVALANCHE OCCURRENCE

Table 1 shows the four wording variants. vJune, the original June 2025 proposal, served as the base. vJune_mod changes level 2 from "a few places" to "some places" and adds "only" at level 1; vAUT replaces the distribution words with "specific", "several" and "isolated" and adds "very easily" at level 4; vJune_simp drops the distribution wording completely and refers to triggering "by an individual". The rationale for each variant arose from internal discussions, reflecting the preferences of some forecasting agencies and, in the vJune versus vJune_simp pair in particular, a wish to test whether giving both trigger sensitivity and distribution is too complicated.

3.1. Correctness and simplicity conditions

All four versions were ranked accurately by most users (Fig. 1a), and the differences between them

Table 1: The four tested avalanche occurrence wording variants (English); every statement begins with "Avalanches …". Each was also tested in a second word order, with the adverb following the verb ("can be triggered easily in many places"); this manipulation survives translation only in English, Italian and Slovenian.

Level	*vJune (base)*	*vJune_mod*	*vAUT*	*vJune_simp*
4 high	can easily be triggered in many places.	can easily be triggered in many places.	can very easily be triggered in many places.	can very easily be triggered by an individual.
3 considerable	can easily be triggered in some places.	can easily be triggered in some places.	can easily be triggered in several places.	can easily be triggered by an individual.
2 moderate	can be triggered in a few places.	can be triggered in some places.	can be triggered in specific places.	can be triggered by an individual.
1 low	can be triggered in rare circumstances.	can only be triggered in rare circumstances.	can rarely be triggered and only in isolated places.	can only be triggered in rare circumstances by an individual.

are small. Pooled across the four main languages, the ordering is vJune_mod 92.0 %, vJune 89.9 %, vAUT 87.0 % and vJune_simp 83.2 %, and all six pairwise contrasts survive Holm correction. Adverb placement made no practical difference, but the comparison is only meaningful in some languages. The two word orders are word-for-word identical in German and Spanish, and differ in only 4 of the 16 statements in French: German syntax leaves the adverb where it is, so for the largest language group the two cells contained the same text. Where the manipulation is real — English (11 of 16 statements) and Italian (14 of 16) — base and adverb cells are indistinguishable (85.1 % against 85.1 %). The one large gap, the French vJune_simp adverb cell at 48.8 %, is a translation error we return to below. We would in any case keep the adverb in front of the verb: a manipulation that does not survive translation into two of six languages is not worth carrying through a multilingual scale. Self-reported ease separates the versions much less: the four pooled means lie between 5.67 and 5.79 (Fig. 1b). The wording variant effect is significant (F(24, 14,690) = 3.7, p = 2×10^{-9}) but spans only 0.13 of a 7-point scale, which is not operationally relevant.

3.2. *Why the pooled numbers need care*

With about 15,000 occurrence rankings, two percentage points are significant, but within languages the picture is less clear. Wording matters within German (likelihood-ratio $\chi^2(3) = 137.2$, Cramér's V = 0.13, 84.4–94.4 %) and within French ($\chi^2(3) = 106.7$, V = 0.25, 71.2–94.2 %), but not within English or Italian once the four tests are Holm-corrected (both p = 0.084). The French effect rests on the faulty adverb cell of vJune_simp; without it French too drops below significance ($\chi^2(3) = 6.3$, p = 0.098). And where wording does matter, it is not the choice between the two strongest versions overall: vJune_mod against vJune is not significant in any single language after correction (German 94.4 vs 92.6 %, OR = 1.36, p = 0.088; French 94.2 vs 89.9 %, OR = 1.83, p = 0.088; English 93.0 vs 92.1 %; Italian 84.2 vs 81.8 %). German alone contributes 8,321 of the 15,062 occurrence rankings, so the pooled result is weighted towards German.

There is a second reason for care. Three of the most striking cell-level results turn out to be translation errors, found only by going back to the questionnaire file. In the French vJune_simp adverb cell, the statements for levels 2 and 3 were translated identically (or the result of a pasting error), so those respondents were asked to rank two identical options which accounts for the full 44-point drop in that cell and also drives the French omnibus test. In the Italian vJune adverb cell the translation used the "by an individual" wording instead of the distribution wording, and in the Italian vAUT adverb cell "specific places" became "pochi luoghi" (a few places). None of this changes the pooled ranking, but each would have been reported as a wording effect had we not checked. These errors are unfortunate but not surprising given the survey's scale.

3.3. *Language differences*

Correctness differs more between languages than between wordings. Pooled across the four versions, English respondents ordered the occurrence statements correctly 93.0 % of the time, German 90.0 %, French 86.9 % and Italian 81.3 % (English versus Italian OR = 3.04, Cohen's h = 0.36). Italian is the weakest language on all four versions, and the picture within languages is just as striking. English respondents show almost no spread across the four wordings — 91.0 % to 95.8 % — and neither the English nor the Italian version effect survives correction across the four languages (p = 0.084 in both); the Italian versions sit between 79.1 % and 84.2 %. The strong version effects are German (84.4–94.4 %) and French (71.2–94.2 %), and the French range is almost entirely due to the

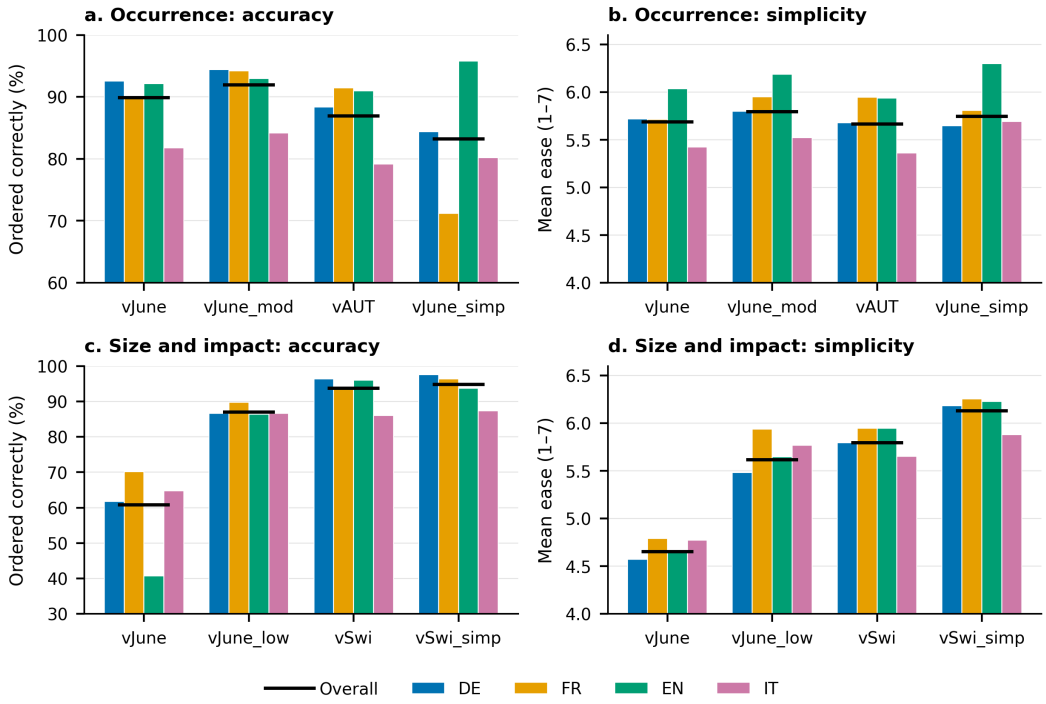

Figure 1: Ranking accuracy (correctness condition) and self-reported simplicity, for avalanche occurrence (a, b) and size and impact (c, d). Accuracy is the percentage of respondents who placed all four statements in the intended danger order or its exact reverse; simplicity is mean self-reported ease (1 = very hard, 7 = very easy). Bars are languages; the black line is the value pooled across the four main languages.

mistranslated vJune_simp cell.

Errors are uncommon, and their shape is more a language than a wording matter: 11.9 % of respondents made a partial error, from 7.0 % in English to 18.7 % in Italian. Among those, the most frequently inverted pair is levels 2 and 3 in German (55.5 %), French (57.8 %) and English (50.0 %), but levels 3 and 4 in Italian (63.4 %). Wording does matter at that boundary — in absolute terms 3.6 % of vJune_mod, 4.2 % of vAUT, 5.2 % of vJune and 12.2 % of vJune_simp respondents placed level 3 below level 2 ($\chi^2(3) = 301$, $p < 10^{-60}$) — but the difference is almost entirely vJune_simp, the one version in which the two middle levels differ by nothing except “easily”. For the other three the rates are low and close together.

3.4. *Increase condition*

This is where the versions separate. Fig. 2 splits the perceived total range into its three steps. vJune, vJune_L1only (vJune with “only” added at level 1) and vAUT all give the moderate-to-considerable step the largest share (0.357, 0.358 and 0.356). vJune_mod gives it the smallest (0.313) and widens the low-to-moderate step instead. vJune_mod sits significantly below the other three in German, French and Italian, with d between 0.42 and 0.58; in English the direction is the same but only the contrast with vAUT survives Holm correction. Unlike correctness, this result is not language-specific: the finding runs in the same direction in all four languages and differs only in size. The shape of the three steps is also informative. For vJune and vJune_L1only the top step exceeds the bottom step by 0.048 and 0.046, shifting weight towards the upper end as an exponential reading requires, while vAUT compresses the top step (-0.034) and vJune_mod is U-shaped, with the middle as the trough (-0.013). No version is read as strictly exponential, which is consistent with users reading the scale close to linearly (Ebert et al., 2025; Morgan et al., 2023), but vJune leans furthest in the intended direction.

4. AVALANCHE SIZE AND IMPACT

4.1. *Two modes of risk information*

The size descriptors we tested carry two pieces of information at each level: the size that most likely poses a threat (“avalanches are typically medium-sized”) and a reasonable worst case (“but some can be large: big enough to bury and kill groups of people”). The rationale was that users should know the typical size they would normally encounter if they triggered an avalanche, while the worst case helps explain differences

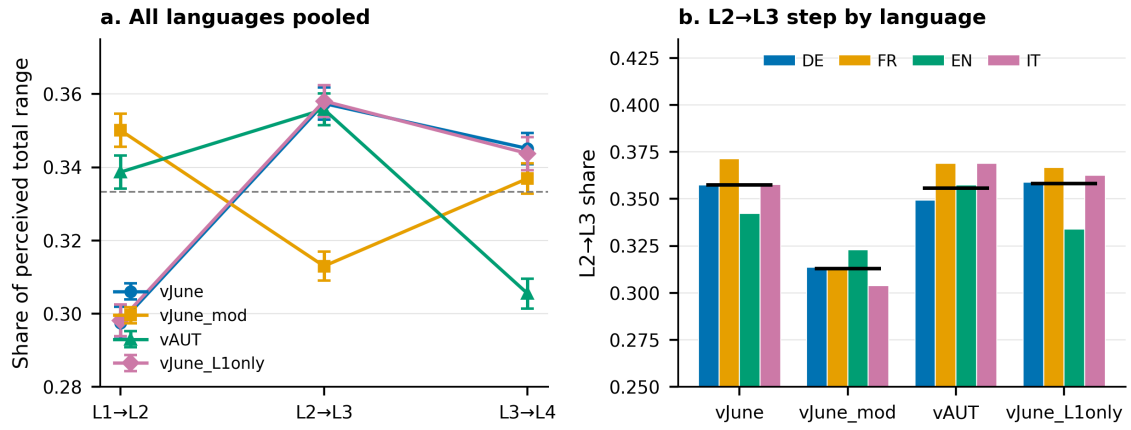

Figure 2: How the perceived increase in danger is distributed across the three steps of the scale. (a) Share of each respondent's own total range (level 4 minus level 1) that falls in each step, pooled across languages, with 95 % confidence intervals; the dashed line marks equal thirds. (b) The moderate-to-considerable share by language.

Table 2: Size and impact wordings with the lowest and highest ranking accuracy: vJune 60.8 % and vSwi_simp 94.8 %, pooled over the four main languages. Every statement begins "Avalanches …", omitted below. vJune_low differs from vJune only at level 1 ("…can be large enough to carry a person."). vSwi differs from vSwi_simp in using "mostly" instead of "typically" and in opening level 4 with "…are mostly medium-sized".

Level	***vJune (consequence only)***	***vSwi_simp (size class + consequence)***
4 high	can be large enough to kill people and can sometimes be large enough to destroy infrastructure.	can be large or very large: big enough to bury and kill groups of people or destroy infrastructure.
3 considerable	can be large enough to kill people and can sometimes be large enough to damage a small building.	are typically medium-sized: they can bury and kill people, but some can be large: big enough to bury and kill groups of people.
2 moderate	can be large enough to carry and sometimes be large enough to bury and kill a person.	are typically small- or medium-sized: they can carry or bury and kill people.
1 low	are expected to carry and can sometimes be large enough to bury and potentially kill a person.	are typically small-sized: they can carry a person.

between levels and matters more to risk-averse users. The main difference between the variants vJune and vJune_low on the one hand, and vSwi and vSwi_simp on the other, is that the former pair did not use size class words (small, medium, large, very large), whereas the latter does.

4.2. *A very large wording effect*

The size column produced by far the largest wording effect in the survey. Ranking accuracy was 94.8 % for vSwi_simp, 93.7 % for vSwi, 87.0 % for vJune_low and 60.8 % for vJune (Fig. 1c). vJune, the least accurately ranked version, was the worst in every language, most severely in English (40.8 %). The cause is a single statement: vJune_low replaces level 1 and nothing else, and accuracy rises from 60.8 to 87.0 %. In vJune, level 1 read "Avalanches are expected to carry and can sometimes be large enough to bury and potentially kill a person" and level 2 "Avalanches can be large enough to carry and sometimes be large enough to bury and kill a person". Both contain "kill", with otherwise little difference in avalanche size, so the two lowest levels are hard to tell apart. Changing only that statement (vJune_low: "Avalanches can be large enough to carry a person") raised accuracy to 87.0 % and cut the share of partial errors involving levels 1 and 2 from 69.4 % to 31.1 %. The problem was not the impact-only wording as such, but that the descriptors did not clearly differ, with high-impact words such as "kill" dominating the description.

4.3. *Size class words help*

The two versions that include a size class — small, medium, large, very large — in addition to describing the consequence reach about 94 %, roughly seven percentage points above the best impact-

only version, and they do so in every language (vSwi_simp: German 97.6 %, French 96.3 %, English 93.7 %, Italian 87.3 %). The same holds for simplicity, and here the effect is large by any standard: vSwi_simp was rated 6.13 against 4.65 for vJune, a difference of 1.48 points on the 7-point scale and Cohen's *d* = 1.04, with partial η^2 = 0.137 for version (Fig. 1d). Every pairwise contrast survives Holm correction and vSwi_simp is rated easiest in every language; its margin over vSwi is small but consistent on both measures (94.8 vs 93.7 %; 6.13 vs 5.80, *d* = 0.26).

4.4. Language differences

Language matters much less here than wording. Every language shows the same ordering and a very large spread between the versions — 22.6 points in Italian, 26.1 in French, 35.9 in German, 55.2 in English — with the version effect highly significant in each (LR χ^2 126 to 1,161). Pooled over versions, the languages sit closer together than for occurrence (French 87.2 %, German 85.8 %, Italian 81.6 %, English 77.7 %); the low English figure is entirely vJune, since English reaches 96.0 % and 93.7 % on the two size-class versions. Italian is again the weakest language overall in ranking accuracy (86.0–87.3 % against 93.4–97.6 % in German and French). Size errors have the opposite shape to occurrence errors: they concentrate at the ends of the scale — levels 1 and 2 (54.2 % of partial errors) and levels 3 and 4 (44.7 %) — while the middle is cleanest (26.1 %). Base rates matter here too: only 5.2 % of vSwi_simp respondents make a partial error, so its high within-error rate at levels 3 and 4 (67.6 %) amounts to 3.5 % of all its respondents — the lowest absolute rate of the four versions, against 13.0 % for vJune.

5. UNDERSTANDING OF FORECAST PHRASES

The survey closed with short interpretation questions about phrases forecasters tend to use. Each respondent saw one of seven, so the answers are independent samples of about 1,900 to 2,000 respondents each. These are the most direct operational results: they ask whether wording already in use means what it is meant to (Table 3).

5.1. What works

"Avalanches are big enough to bury a person" was understood to imply that they can also kill by 95.9 % of respondents, the only one of the seven questions with no difference at all between languages. "Generally safe avalanche conditions" survives scrutiny from three directions: 97.2 % chose "a very small chance of an avalanche" over "no risk", 86.6 % said it does not imply that there is no chance of an avalanche, and 89 % said it does not mean avalanches are impossible to trigger. It is read as low residual risk, not as an absence of avalanche risk.

5.2. What works less well

"Avalanches are usually small-sized: they can carry a person" was read as still potentially fatal by only 77.3 %, with 8.9 % unsure and 13.8 % answering no: about one user in seven does not connect a small avalanche described as carrying a person with a possible fatality. Users associated quite different avalanche danger levels with natural-avalanche phrases: "Natural avalanches are not expected to occur" is used in the new descriptor as a low-danger marker, but only 39.2 % of respondents associated it with low danger, while 56.6 % went to moderate and 4.3 % to no danger at all. "Natural avalanches can occur in a few places" split 47.2 % moderate against 41.9 % considerable, with a further 10.9 % at low, so it does not identify a level. Educational work is definitely needed and forecasters should remember that a negated expectation does not always work as reassurance!

6. CONCLUSIONS

The survey results are most relevant for the avalanche size column, where using size-class words and avoiding "bury and kill" at the lowest level raised ranking accuracy from 61 % to 95 %. For the avalanche occurrence column the news is largely good: end users ranked all four variants accurately, and they read the distribution words — a few, some, many — in the intended order. The three variants that give both trigger sensitivity and distribution did best, and only vJune_simp, which drops the distribution wording, is clearly weaker, and then only at the moderate-to-considerable boundary. The increase condition separates the variants clearly and in the same direction in all four languages. What remained contested among us is how much weight that criterion should carry against correctness and simplicity, and on that we did not reach consensus. Our operational conclusions: name the size class, to differentiate the levels; make sure the impact language differentiates at the lower levels, since signal words such as "kill" crowd out finer differences at the bottom end; and be aware of differences between languages — with samples this large, results should be reported within language, and for avalanche occurrence the differences between languages were larger than those between wordings. Finally, we strongly recommend involving end users in this process: several changes that seemed harmless turned out to have measurable costs, such as lifting level 2 to

Table 3: Interpretation of phrases used by forecasters, by language. Percentages are of respondents who answered the question. * Chi-squared test of independence between language and response. † The French version of this question could not be used because its response labels were corrupted, so the pooled figure excludes French.

Statement and question	*DE*	*FR*	*EN*	*IT*	*All*	*p**
"Big enough to bury a person." Can they kill? (% yes)	96	97	96	96	95.9	0.44
"Usually small-sized: they can carry a person." Can they be fatal? (% yes)	75	77	80	82	77.3	0.027
"Generally safe." No chance of an avalanche? (% no)	87	82	87	88	86.6	0.0027
"Generally safe." Impossible to trigger? (% no)	91	–	90	82	89†	4×10^{-6}
"Generally safe." A very small chance rather than no risk (%)	96	98	98	99	97.2	0.014
"Natural avalanches are not expected to occur." → low danger rating (%)	36	10	53	55	39.2	$< 10^{-10}$
"Natural avalanches can occur in a few places." → considerable danger rating (%)	49	24	36	37	41.9	$< 10^{-10}$
Respondents per question (*n*)	1,086–1,154	185–257	191–260	390–442	1,897–2,022	

"some places" or dropping the distribution wording. The wording finally adopted by the EAWS differs from the versions tested here, so our results reflect the June 2025 proposal and its variants rather than to the most recent proposal (Müller et al., 2025).

ACKNOWLEDGEMENTS

The survey was designed by PE and GP who wrote the first draft. FT, IG, KM contributed to the final version. All listed authors produced translations and helped with recruitment. Claude Opus 5.0 (Anthropic) helped write the original R analysis, produced the figures, and drafted and edited parts of the manuscript. The authors checked all results and take full responsibility. PE's research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/COE3]. For open access purposes, the author (PE) has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

REFERENCES

Degraeuwe, B., Schmudlach, G., Winkler, K., and Köhler, J., 2024. SLABS: An improved probabilistic method to assess the avalanche risk on backcountry ski tours, Cold Reg. Sci. Technol., 221, 104169.

Ebert, P. A., 2026. The ethics of communicating voluntary and personal risk, Synthese, 208, 57, https://doi.org/10.1007/s11229-026-05692-w.

Ebert, P. A., Miller, D. L., Comerford, D. A., and Diggins, M., 2025. End user and forecaster interpretations of the European Avalanche Danger Scale: a study of avalanche probability judgments in Scotland, Risk Anal., 45, 2285–2299, https://doi.org/10.1111/risa.70016.

Engeset, R. V., Pfuhl, G., Landrø, M., Mannberg, A., and Hetland, A., 2018. Communicating public avalanche warnings – what works?, Nat. Hazards Earth Syst. Sci., 18, 2537–2559.

Lucas, C., Grüter, S., Eberli, M., Trachsel, J., Winkler, K., and Techel, F. (2023). "Introducing sublevels in the Swiss avalanche forecast". In: Proceedings, International Snow Science Workshop. Bend, Oregon, USA, 240–247.

Mitterer, C. and Mitterer, L., 2018. 25 Jahre Europäische Lawinengefahrenstufenskala, bergundsteigen, 104, 67–76.

Morgan, A., Haegeli, P., Finn, H., and Mair, P., 2023. A user perspective on the avalanche danger scale – insights from North America, Nat. Hazards Earth Syst. Sci., 23, 1719–1742.

Müller, K., Techel, F., and Mitterer, C., 2025. The EAWS matrix (Part A): conceptual development, Nat. Hazards Earth Syst. Sci., 25, 4503–4525.

Müller, K., Techel, F., Mitterer, C., Ebert, P. A., Reuter, B., Krieger, V., Chiambretti, I., Dufour, A., Bellido, G. M., and Bertrand, L. (2026). "Revising the European Avalanche Danger Scale: Closing the gap between forecasting and public communication". In: Proceedings International Snow Science Workshop, Whistler, BC, Canada, 28 Sep–2 Oct 2026.

Schweizer, J., Mitterer, C., Techel, F., Stoffel, A., and Reuter, B., 2020. On the relation between avalanche occurrence and avalanche danger level, The Cryosphere, 14, 737–750.

St. Clair, A., Finn, H., and Haegeli, P., 2021. Where the rubber of the RISP model meets the road: contextualizing risk information seeking and processing with an avalanche bulletin user typology, Int. J. Disaster Risk Reduct., 66, 102626.

Techel, F., Müller, K., Marquardt, C., and Mitterer, C., 2026. The EAWS matrix (Part B): operational testing and use, Nat. Hazards Earth Syst. Sci., 26, 1161–1181.

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 Reg. Sci. Technol., 188, 103299.