Cold Regions Science and Technology 59 (2009) 217–226

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

Comparison of snow stability tests: Extended column test, rutschblock test and compression test

Kurt Winkler *, Jürg Schweizer

WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos Dorf, Switzerland

ARTICLE INFO

Article history:
Received 5 November 2008
Accepted 11 May 2009

Keywords:
Snow cover stability
Snow stability evaluation
Stability test
Avalanche forecasting

ABSTRACT

Several field tests have been proposed in the past for evaluating snow stability. However, few comparative studies have been performed so that presently the advantages and disadvantages of the various tests are partly unclear. During winter 2007–2008 we have collected a dataset of 146 snow profiles, consisting of snow stratigraphy, a rutschblock test (RB), one to two extended column tests (ECT) and in most of the cases also one to two compression tests (CT). Study slopes were classified in regard to stability as either rather stable or rather unstable, based on signs of instability or profile classification. We then studied whether the various tests were able to predict the slope stability class. The CT had an almost perfect probability of detection, but — as the structural stability index (threshold sum) — the CT largely overestimated instability (high proportion of false alarms). Of the small scale tests the ECT was best suited to differentiate between stable and unstable situations. By including the ECT score (number of taps), the number of false alarms was slightly reduced. The performance was similar to the RB which is, however, not independent of the stability classification we used. With two adjoining ECTs it was possible to classify 87% of our test slopes with an accuracy of about 90% into rather stable or rather unstable. Comparing two adjacent stability test results showed that only in about half of the pairs the same weak layer showed up as the most critical one. The snowpack properties (weak layer and slab) that favoured unstable test results for the ECT were associated with whole block releases in the rutschblock test. Thus, the two tests seem to provide similar information possibly related to fracture propagation propensity.

1. Introduction

Forecasting dry-snow slab avalanches relies on snow stability information. In the absence of signs of instability such as collapsing sounds ("whumpfs") (e.g. Seligman, 1936, p. 426), shooting cracks or avalanche activity, snow stability is assessed with the help of stability tests. Several stability tests are in use and new tests have been proposed recently, however, it is presently unclear which test is best suited — given certain snow conditions — to provide information on the ease of fracture initiation as well as on fracture propagation propensity (Schweizer et al., 2003).

The two tests most widely used are the rutschblock test (Föhn, 1987) and the compression test (Jamieson, 1999). For both tests it has been shown that the score is related to skier-triggered avalanche activity (Föhn, 1987; Jamieson and Johnston, 1995; Jamieson, 1999), but also that the test score can be highly variable. The usefulness of both tests has been improved by noting the type of fracture: the release type for the RB (area of the block that releases) (Schweizer and Wiesinger, 2001) and the fracture character for the CT (van Herwijnen and Jamieson, 2007). Whereas the stability test score depends on the weak layer strength, and hence should be related to fracture initiation, the fracture type relates to fracture propagation propensity and depends, among other things, on the slab properties (Schweizer et al., 2008).

Critical weak layers often have typical properties compared with their surrounding layers. Consequently, structural instability indices (threshold sum) were developed (e.g. Schweizer and Jamieson, 2007). They are best known as lemons (McCammon and Schweizer, 2002) or yellow flags (Jamieson and Schweizer, 2005).

Recently, beam-type fracture tests were proposed (Gauthier and Jamieson, 2006; Sigrist and Schweizer, 2007). In a fracture test an artificial fracture is introduced and the resistance to fracture is tested by either increasing the load to failure for a given crack length or increasing the crack length until failure occurs for a given load. Gauthier and Jamieson (2008a) have further developed the test into a fracture propagation field test (the so-called propagation saw test: PST) and showed that the PST was able to correctly predict slope-scale fracture propagation in about 75% of the cases (Gauthier and Jamieson, 2008b).

Simultaneously, Simenhois and Birkeland (2006, 2007) proposed and validated a new test of similar design, the so-called extended column test (ECT). Validation results suggest that the ECT is a very good predictor of instability (Birkeland and Simenhois, 2008).

* Corresponding author. Tel.: +41 81 4170127; fax: +41 81 4170110.
E-mail address: winkler@slf.ch (K. Winkler).

0165-232X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.coldregions.2009.05.003

The newly developed tests were compared with each other (Birkeland and Simenhois, 2008; Simenhois and Birkeland, 2009-this issue) and some of the well-established tests (Gauthier and Jamieson, 2008c) partly addressing the question of which test performs best under which conditions. Previously, Jamieson (1999) had compared the CT with the RB.

The aim of this study was to compare the ECT with other well-established tests (RB and CT) and the structural instability index, and assess its performance for the snow conditions of the Swiss Alps.

2. Methods

2.1. Observations

On each study slope we performed a full snow profile in conjunction with a number of stability tests: a rutschblock test (RB), one to two extended column tests (ECT) and in most of the cases also one to two compression tests (CT). The different tests were arranged as closely together as possible (Fig. 1). The rearmost wall of all the columns and the RB were cut with a cord. Occasionally, observations were combined with snow micro-penetrometer measurements (SMP) (Pielmeier and Marshall, 2009-this issue). The snow profile observations included grain type and size as well as hand hardness index of each layer in the snowpack, snow temperature, and in 68% of the profiles also ram hardness, all corresponding to standard methods (e.g. CAA, 2002; Greene, 2004).

2.2. Tests and their interpretation

Below we describe the various predictors of instability and how we define how the test results are interpreted in regard to stability.

Based on observed snow stratigraphy, we calculated the threshold sum as indicator of structural instability using the threshold values as described by Schweizer and Jamieson (2007). A threshold sum of ≤ 4 indicated rather stable, ≥ 5 rather unstable conditions.

For the rutschblock test (RB), performed as described in Schweizer (2002), the score and the release type were recorded. For the RB score, values ≤ 3 indicated rather unstable, higher values rather stable conditions. Only the release type "whole block" indicated unstable conditions (e.g. Schweizer et al., 2008). The results of the RB score and the RB release type were combined with the threshold sum (Schweizer et al., 2008).

For the compression test (CT), the number of taps (score) and the fracture character were recorded according to van Herwijnen and Jamieson (2007). A CT score ≤ 13 indicated rather unstable, >13 rather stable conditions. The fracture characters "Sudden Collapse" (SC) and "Sudden Planar" (SP) were assumed to be related to unstable slopes; sudden fractures are equivalent to shear quality Q1 (Birkeland and Johnson, 1999). The fracture characters "Resistant Planar" (RP), "Progressive Compression" (PC), "Non-planar Break" (B) and no fracture were assumed to indicate rather stable conditions. When analyzing the combination of CT score and fracture character, we either required that the unstable test result had to be found for the same failure layer, or allowed the unstable test results to be from different failure layers at different depths. As for the RB, we combined the CT test results with the threshold sum.

The extended column test (ECT) was performed according to Simenhois and Birkeland (2006). The slope was assumed to be rather unstable when a fracture crossed the entire column in one layer and during the same or the next loading tap when the fracture initiated.

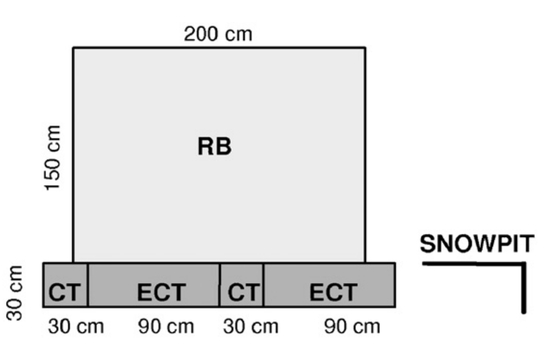

Fig. 1. Set-up for slope observations. Either one or two ECT and CT were done adjacent in front of the RB.

Table 1
Criteria for classifying study slopes as rather unstable.

Criterion	Number of slopes for which criterion was fulfilled
Whumpf on study slope	17 (2)
Crack on study slope	10 (2)
Whumpf or crack on study slope	20 (4)
Recent natural or human-triggered avalanche on nearby slope	12 (3)
Profile classified as "poor" or "very poor" according to Schweizer and Wiesinger (2001)	29 (9)
Total number of unstable slopes (at least one criterion fulfilled)	36

Values in brackets give the number of study slopes where only this sign of instability was present (no other criterion of instability fulfilled).

2.3. Classification of study slopes

Study slopes were classified with regard to stability. A slope was considered as rather unstable (Table 1), when at least one of the following criteria was satisfied:

1) Signs of instability such as whumpfs or cracks were observed on the study slope.
2) Recent (not more than one day old) natural or human triggered avalanches on nearby slopes.
3) The profile was classified as "poor" or "very poor" according to Schweizer and Wiesinger (2001). This criterion takes into account — among other things — the stratigraphy, the properties of the individual layers, and the stability test result (RB score and release type).

As the RB test results were used for the profile classification, the third criterion is not independent of the RB test and hence our classification of study slopes is partly biased towards the RB test.

With the first two criteria, obvious signs of instability were present and we can assume that the slope might be unstable without further investigation. The real value of stability tests is to detect unstable conditions even in the absence of obvious signs of instability. To classify these slopes, the third criterion is essential — even though the classification is somewhat subjective. However, analyzing the snow profile data used by Schweizer et al. (2008) collected on slopes that were either skier-triggered or skier tested showed that this classification had an unweighted average accuracy of about 75%.

For analyzing the reproducibility of the stability tests and the most critical weak layer, we subdivided the class of rather stable slopes into either "fair" if the RB score was ≤ 3 or the RB release type was "whole block", and "good" otherwise.

2.4. Data analysis

For each observation on a slope, the slope stability estimate was compared to the results of the various stability tests. For this purpose, only slabs in the range of 0.13 to 0.89 m were considered (Schweizer and Jamieson, 2007). Whereas thinner slabs are less critical, as they cause mostly small avalanches, thicker slabs are rarely triggered by people (Schweizer and Lütschg, 2001). This restriction only marginally influenced our test results.

On most test slopes, several potential weak layers existed in the snowpack — each of them with an unknown stability. On all slopes where there were no signs of instability, we cannot know a priori which of the potential weak layers is the most critical one — and it would be arbitrary to designate one of the layers as the most critical or decisive weak layer. Thus, for a given test location different layers may have been considered as the most critical one depending on the results from the various tests.

The most critical (or decisive) failure layer was determined as described below:

– For the scores, the first fracture was assumed to be decisive (regardless of release type or fracture character). If multiple fractures occurred at the same tap, we chose in following order: (1) the fracture that initiated first, (2) the fracture with the more critical release type or fracture character.
– For the RB release type as well as for the CT fracture character, we selected the failure layer depth with the most critical fracture type. If a critical fracture type was observed at various depth, we selected the most critical one as follows: (1) the layer with the lower score, (2) the layer that was observed to fracture first.
– For the threshold sum, the layer with the highest score in the profile was considered as the critical one.
– For the ECT, the layer where the fracture propagated furthest was assumed to be the critical failure layer. If more than one fracture crossed the entire column, the layer was selected for which the number of additional taps it took for the fracture to cross the entire column, was lowest. In case of ties, the critical layer was selected as follows: (1) the layer where the fracture started at the lowest score, (2) the layer that fractured first.

If finally still multiple fractures remained in a single stability test, all of them were assumed to be critical.

To determine whether the critical layers found in the various tests agreed between tests, the analyses were done separately for each CT or ECT test in a pair.

When analyzing the reproducibility of a specific test, we checked for each pair whether the critical layer found in the first test, coincided with the critical layer found in the second test.

To describe the performance of the different tests, the following measures for categorical forecasts were used (Doswell et al., 1990). With the definitions used in contingency tables (Table 2), the measures are calculated as follows:

$$\text{Probability to detect a stable slope or specificity: PON} = \frac{a}{a+c} \quad (1)$$

$$\text{Probability to detect an unstable slope or sensitivity: POD} = \frac{d}{b+d} \quad (2)$$

$$\text{Unweighted average accuracy: RPC} = 0.5\left(\frac{a}{a+c} + \frac{d}{b+d}\right) \quad (3)$$

With only about 25% unstable observations (Table 3) our dataset is not balanced and the unweighted average accuracy must be used instead of the overall accuracy (or hit rate).

Table 2
Contingency table.

		Observed stability	
		Stable	Unstable
Result of the stability test	Stable	a	b
	Unstable	c	d

Total of samples: $N = a + b + c + d$.

Table 3
Dataset and proportion of unstable slopes.

Stability test	Number of samples	Proportion unstable
RB	146	0.25
ECT	225	0.25
	(67 profiles with 1 test)	(0.24)
	(79 profiles with 2 tests)	(0.25)
CT	240	0.27
	(32 profiles with 1 test)	(0.16)
	(104 profiles with 2 tests)	(0.29)
Threshold sum	146	0.25

Further of relevance are the false alarm ratio FAR $= \frac{c}{c+d}$ and the frequency of misses (so-called false-stable predictions) FOM $= 1 - \text{POD} = \frac{b}{b+d}$. Although stability tests are only one factor considered in avalanche forecasting, false-stable predictions can have more serious consequences than false alarms. In the following we will mainly report the specificity, the sensitivity and their average i.e. the unweighted average accuracy.

If at a single snow pit location two ECT or CT were made, we did not consider the two test results independently for the analysis. As a mean value cannot be derived, we did all the comparisons to other tests twice by selecting first the first test of each pair for the statistical analysis, and then repeated the analysis with the second test of each pair, and calculated the mean of the two analyses.

For investigating the influence of weak layer and slab properties on the stability test results (indicating either rather unstable or rather stable conditions), we analyzed the snow properties for all failure layers found with a given test (ECT, RB score and RB release type) that were at least 13 cm below the snow surface. The failure layer was defined as follows (according to Schweizer and Jamieson, 2003). (a) If the failure was recorded at the boundary between two layers, the softer of the two layers was considered as the failure layer. In case of no difference in hardness between the upper and lower layer, the layer with the larger grain size was considered as the failure layer, and if no grain size difference existed, we chose the upper layer as the failure layer (and the lower one as the adjacent layer. (b) If the failure was recorded by the observers within a layer, this layer was considered as the failure layer. If the observed failure was close to either the upper or lower interface (≤1 cm), that interface was considered as the failure interface and the layer across the interface was considered as the adjacent layer. If the observed failure was within 1 to 5 cm of either the upper or lower interface, we assume the failure interface to be the boundary with the larger hardness or grain size difference, or in case of no difference, we chose the lower layer as adjacent layer. (c) Occasionally, a failure was observed within a thick (>10 cm) homogenous layer. We assumed a layer boundary at the depth where the failure was observed. Consequently, the failure layer (upper layer) and the adjacent layer (lower layer) had identical snow properties.

The following weak layer and slab properties were used to study their influence on the stability test results:

– location of the fracture in the failure layer: either top, base or within; fractures within were those that were >1 cm away from both layer boundaries
– hand hardness index of the failure layer, the adjacent layer, and of the layers directly below and above the failure layer
– difference in hand hardness index between the failure layer and the adjacent layer (across the failure interface)
– the maximum and the average hand hardness index of the slab weighted by thickness
– the slab thickness
– the slab thickness multiplied with the weighted average slab hardness (called "bridging" parameter by Schweizer and Jamieson, 2003).

– "slab strength index": As the hand hardness index (k) represents approximately a logarithmic scale of penetration resistance, we calculated the "slab strength index" (A) by multiplying the thickness (h) of each layer with the square of its hand hardness index and added these values up over the slab (unit: cm × hand hardness index^2):

$$A = \sum\nolimits_{\text{fracture}}^{\text{snow\_surface}} h_i k_i^2 \qquad (4)$$

– grain size in the failure layer, the adjacent layer and the grain size difference across the failure interface; the arithmetic mean of the reported average size and the average maximum grain size was used.
– grain type (either persistent or non-persistent); grain type was assumed to be persistent, if the primary reported type was either faceted grains (FC), depth hoar (DH) or buried surface hoar crystals (SH), and non-persistent otherwise.
– threshold sum (calculated according to Schweizer and Jamieson (2007)).

To check for differences in the performance of the various tests, the two-proportion Z-test (SYSTAT, 2007) was used. To contrast layer properties from tests which indicated rather stable conditions with those which indicated rather unstable conditions we used the non-parametric Mann–Whitney U-test. Observed differences were judged to be statistically significant where the level of significance is $p<0.05$. For comparing categorical variables such as grain type, the data were cross-tabulated and the Pearson χ^2 statistic was calculated. Classification tree statistics provided threshold values when contrasting stable with unstable test results (Breiman et al., 1998).

3. Data

Data from 146 profiles collected during winter 2007–2008 were analysed. All profiles were from the Alps, mainly from the Grisons region in Switzerland. The elevation at the profile site ranged from 1936 m to 3184 m a.s.l. with a median elevation of 2450 m a.s.l. Profiles were performed prevailingly on shady slopes (NW, N and NE) (Fig. 2) where more frequently poor snow stability can be found and a large part of the avalanche accidents occur. The mean slope angle was 35°, and the large majority (88%) of the profiles was taken on slopes >30°. For safety reasons, most of the profiles were taken on small slopes or near the top or the side of medium sized slopes.

The profile type was classified according to Schweizer and Wiesinger (2001) mostly based on the ram hardness (in 68% of the cases), otherwise based on the hand hardness. The dataset contained all different profile types; profile type 7 was found most frequently (Fig. 3). Most profiles during the winter 2007–2008 showed rather hard basal layers (62%), but snowpacks with weak basal layers were still well represented (38%).

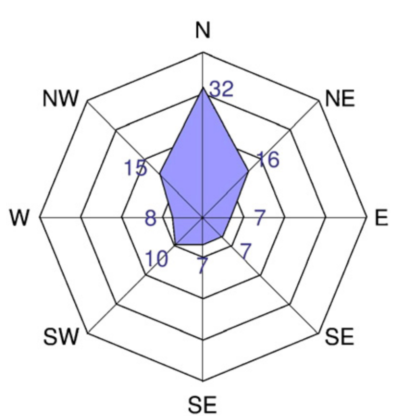

Fig. 2. Relative frequency of aspects (%), ($N = 146$).

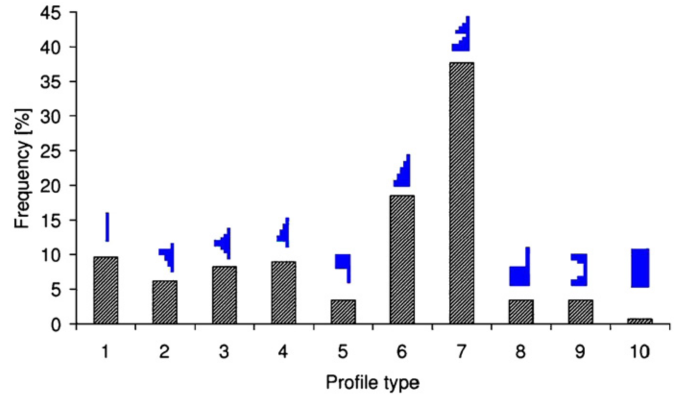

Fig. 3. Relative frequency of profile types (only full depth profiles to the ground were considered, $N = 137$).

The dataset contained several new snow instabilities, but 76% of the unstable ECT fractures originated from persistent weak layers with mostly faceted crystals or depth hoar.

4. Results

4.1. Classification

Table 4 shows the performance of the different stability tests. Of all tests, with 90% (score) or 84% (release type), the RB had the highest probability to detect a stable slope (specificity, Eq. (1)) but also the highest rate of undesired false-stable predictions (1-sensitivity, Eq. (2)). Combining the score with the release type did not improve the unweighted average accuracy. If both score and release type had critical values, the specificity increased to 99% and the sensitivity

Table 4
Classification results for the different stability indicators; if the indicator was in the critical range, the slope was rated unstable.

Test or indicator	Specificity (correct stables)	Sensitivity (correct unstables)	Unweighted average accuracy
RB score	0.90	0.78	0.84
RB release type	0.84	0.78	0.81
RB score and RB release type	0.99	0.61	0.80
RB score or RB release type	0.75	0.94	0.84
2 of RB score, RB release type, threshold sum	0.80	0.86	0.83
ECT*	0.79 *(0.78/0.79)*	0.83 *(0.81/0.86)*	0.81 *(0.79/0.83)*
ECT (score ≤21)*	0.82 *(0.81/0.83)*	0.83 *(0.81/0.86)*	0.83 *(0.81/0.84)*
CT score*	0.45 *(0.47/0.44)*	0.90 *(0.89/0.91)*	0.68 *(0.68/0.67)*
CT fracture character*	0.22 *(0.21/0.24)*	0.99 *(0.97/1.00)*	0.60 *(0.59/0.62)*
CT score and CT fracture character* (any layer)	0.50 *(0.50/0.50)*	0.89 *(0.86/0.91)*	0.69 *(0.68/0.71)*
CT score and CT fracture character* (same layer)	0.56 *(0.56/0.56)*	0.93 *(0.91/0.94)*	0.75 *(0.74/0.75)*
CT score or CT fracture character*	0.17 *(0.18/0.17)*	1.00 *(1.00/1.00)*	0.59 *(0.59/0.58)*
CT score and CT fracture character and threshold sum* (any layer)	0.62 *(0.61/0.63)*	0.74 *(0.71/0.77)*	0.68 *(0.66/0.70)*
CT score and CT fracture character and threshold sum* (same layer)	0.86 *(0.88/0.85)*	0.13 *(0.16/0.11)*	0.50 *(0.52/0.48)*
2 of 3: CT score, CT fracture character or threshold sum* (same layer)	0.37 *(0.37/0.37)*	0.89 *(0.87/0.92)*	0.63 *(0.62/0.64)*
Threshold sum	0.38	0.86	0.62

Asterisk (*) indicates results for averaged values where two tests were available; numbers in italic denote values using either the first or the second test of each pair.

Table 5
Pair-wise reproducibility of the ECT, the CT score, the CT fracture character and the RB test.

Result of the test pairs	ECT			CT score			CT fracture			RB		
	Stability of the slope			Stability of the slope			Stability of the slope			Stability of the slope		
	Unstable	Stable		Unstable	Stable		Unstable	Stable		Unstable	Stable	
		Fair	Good		Fair	Good		Fair	Good		Fair	Good
Twice unstable	16	4	2	28	12	23	28	16	38	22	1	0*
Once unstable, once stable	2	5	3	1	6	16	1	5	10	12	24	3
Twice stable	2	9	36	0	3	15	0	0	6	2	2	80
Total pairs	79			104			104			146		
Same critical layer	16	14	22	26	13	27	25	12	25	35	27	83
Different critical layers	5	4	19	6	8	28	7	9	30	1	0	0
Total pairs	80			108			108			146		

For the RB test no pairs were available; instead it was compared whether the RB score and the RB release type indicated the same level of stability. The asterisk (*) denotes an artefact resulting from the definition of “fair”.

decreased. This very low false alarm rate was not attained by any other of the instability indicators ($p=0.003$) but came at the price of a significantly reduced sensitivity ($p=0.035$, except for the combination of CT and threshold sum, for which the difference was not significant). With either the RB score or the RB release type in the critical range, the probability to detect an unstable slope was large, and consequently the proportion of false-stable prediction was low (6%), but the specificity decreased and was significantly lower than for the RB score alone ($p=0.003$). With the combination of RB score, RB release type and threshold sum (at least two of the three variables in the critical range) a good, balanced performance was reached with 14% false stables and 20% false alarms.

With 99%, the CT fracture character showed the highest sensitivity (= probability to detect unstable slopes) of all tests. The sensitivity was significantly better than with the RB ($p=0.041$); the difference to the ECT was marginally not significant ($p=0.051$). However, the specificity of the CT was low and this deficiency cannot easily be compensated. The best criterion for an unstable test result was that CT score and CT fracture character both had to indicate instability for one and the same layer. Still, the false alarm ratio was high (42%).

With 86%, the threshold sum reached a high sensitivity and proved to be helpful in combination with the RB test. On its own, the threshold sum showed a poor specificity.

In our dataset, all the correct unstable predictions from the ECT occurred at ≤21 taps. All unstable predictions with taps >21 were false-alarms. Thus we can increase the specificity of the ECT without reducing the sensitivity by considering only the fractures occurring up to the 21st tap. However, this threshold might be specific to our dataset. The ECT had neither the best specificity (82%) nor the best sensitivity (83%), but showed balanced values for both measures. The specificity was significantly higher than for the CT, the threshold sum and all the combinations of these two indicators ($p=0.007$).

4.2. Reproducibility: test result

When stability tests are repeated side by side as in our set-up similar test results should be expected — at least on rather uniform slopes (Jamieson and Johnston, 1993). In fact, for the ECT, in 87% of the cases the same stability class was found, i.e. twice stable or twice unstable (Table 5). The reproducibility increased to 92% if only the test pairs were considered that were done on a slope that was classified as either “unstable” or as “good”. On the slopes classified as “fair” the reproducibility was with 72% significantly lower ($p=0.048$). When two ECTs were done close together and both tests indicated rather unstable conditions, the slope was rarely stable, as indicated by the high specificity of 90% in Table 6. Similarly, there was a high probability (90%) that the slope was not unstable if both ECTs showed stable test results. The remaining 13% included 2 cases from rather unstable slopes, and 8 cases from rather stables slopes (5 “fair”; 3 “good”).

Compared to the ECT, the CT fracture character (85%) and the CT score had both a lower reproducibility (78%), but the differences were not significant. Both indicators showed a very high reproducibility (97%) on slopes rated as rather unstable. This follows from the high sensitivity and poor specificity of the CT.

For the CT, combining the pair-wise test results did not improve classification results satisfactorily since the specificity remained low. For comparison with the RB, we considered RB score and RB release type instead of two different tests. On all the slopes where both test results indicated the same stability class, the unweighted average accuracy was very high (97%). However, in the remaining 27% of the cases when the test results did not agree, the combination did not reduce uncertainty.

4.3. Reproducibility: critical failure layer

Of the 80 cases considered, the critical failure layer as found with the first ECT of each pair, coincided in 52 cases (65%) with the critical failure layer found with the second ECT (Table 5). In the 23 cases where both tests indicated rather unstable conditions, the critical failure layer agreed more often (83%) than in other cases 58% ($p=0.036$).

For the CT the agreement was slightly lower: 61% for the CT score, 57% for the fracture character. However, these differences were statistically not significant. As in the case of the ECT, the failure layer agreement was higher on unstable slopes (81% for the CT score; 78% for the CT fracture character) than on stable slopes ($p=0.005$).

The nearly perfect agreement of the critical failure layers found with the RB test cannot be compared with the above values, because

Table 6
Classification results by using stability test pairs (ECT, CT).

Tests	Unstable, if both test unstable		Stable, if both tests unstable		Pairs gives the same result				1 test stable, 1 test unstable	
	Specificity	Sensitivity	Specificity	Sensitivity	Ratio	Specificity	Sensitivity	Unweighted average accuracy	Ratio	Classification
2 ECT	0.90	0.80	0.76	0.90	0.87	0.90	0.90	0.90	0.13	Suspect
2 CT, score	0.53	0.97	0.24	1.00	0.78	0.53	1.00	0.77	0.22	Suspect
2 CT, fracture character	0.28	0.97	0.08	1.00	0.85	0.28	1.00	0.64	0.15	Suspect
1 RB, score and release type	0.99	0.61	0.75	0.94	0.73	0.99	0.94	0.97	0.27	Suspect

For the RB, the RB score and RB release type from the same test were compared.

Table 7
Agreement of the critical failure layer: probability to find a critical failure layer identified by the first test, in the second test as well.

First test	Second test					
	RB score	RB release type	CT score	CT fracture character	ECT	Threshold sum
RB score	147	147/0.99	254/0.43	254/0.46	226/0.51	147/0.32
	(64)	(64/0.98)	(117/*0.59*)	(117/*0.60*)	(102/*0.64*)	(64/*0.41*)
RB release type	147/0.99	147	254/0.42	254/0.45	226/0.51	147/0.33
	(64/0.98)	(64)	(117/*0.57*)	(117/*0.58*)	(102/*0.64*)	(64/*0.42*)
CT score	255/0.42	255/0.42	255	255/0.83	266/0.45	255/0.29
	(125/*0.55*)	(125/*0.54*)	(125)	(125/*0.88*)	(128/0.50)	(125/0.33)
CT fracture character	252/0.46	252/0.45	252/0.84	252	263/0.48	252/0.29
	(122/*0.57*)	(122/*0.56*)	(122/*0.90*)	(122)	(125/*0.54*)	(122/0.32)
ECT	229/0.51	229/0.51	275/0.43	275/0.45	228	229/0.26
	(105/*0.62*)	(105/*0.62*)	(127/*0.50*)	(127/*0.54*)	(105)	(105/*0.32*)
Threshold sum	297/0.17	297/0.18	502/0.15	502/0.15	457/0.13	295
	(116/0.22)	(116/*0.23*)	(209/*0.20*)	(209/*0.19*)	(181/*0.19*)	(181)

Values in the upper line give the number of pairs and the probability. Values in brackets in the lower line are for tests performed on either rather unstable slopes or stable slopes that were rated as "fair", again the number of pairs and the probability are given. Numbers are given in italic if the difference to stable slopes rated as "good" was statistically significant ($p<0.05$).

they do not result from two different tests, but from analysing two different test results from a single RB test. They do not represent the reproducibility.

Considering different tests, the critical failure layer identified by the different tests did less frequently agree than in the case of two of the same tests (Table 7). In slightly more than half of the snow pits (51%), the RB and the ECT identified the same layer as the critical failure layer. Comparing CT score or CT fracture character with the RB or the ECT revealed a slightly lower agreement (42–48%). With agreement scores of 13% to 33%, the threshold sum only relatively rarely identified the same layer as the critical failure layer that was found with the other tests. As often several critical failure layers were identified with the threshold sum the agreement depends on the direction of comparison.

Analysing two different test results from the same stability test showed good agreement of the critical failure layers: 99% for the RB test, 84% for the CT. However, these values are not independent of each other. Except for the comparison of the RB score with the RB release type, the agreement between the critical failure layers was always higher between tests performed on either unstable slopes or stable slopes that were rated "fair", than on stable slopes rated as "good"; the differences were in most cases significant.

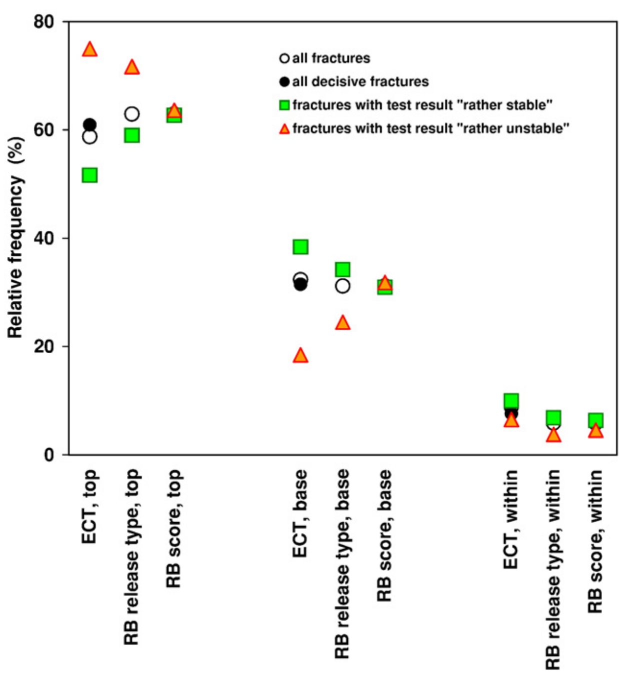

Fig. 4. Location and its frequency of the fracture in the failure layer for the ECT, the RB release type and the RB score.

4.4. Weak layer and slab properties

For the ECT, in half of the stable (52%) and three quarters of the unstable (75%) test results the fracture was found at the top of the failure layer (Fig. 4). Fractures at the top were significantly more often ($p=0.001$) associated with unstable test results than with stable test results and vice versa for fractures at the base ($p=0.006$). For the RB release type, the same behaviour was observed, but the trend was

Table 8
Comparison of weak layer and slab properties from stable and unstable test results of the ECT.

Variable	Stable median	Unstable median	p	Threshold
Location of fracture in failure	55,15,65	12,4,46	**0.003**	–
layer* (base, within, top)	*56,16,78*	*10,3,42*	***0.007***	–
Hardness failure layer	1.5	1	**<0.001**	<1.5
	1	*1*	***0.001***	*<1.5*
Hardness adjacent layer	3	3	0.254	–
	3	*3*	*0.290*	–
Difference in hardness	1	1.5	0.447	–
	1	*1.5*	*0.604*	–
Hardness below failure layer	3	2	**0.005**	<2.5
	3	*2*	***0.002***	*<2.5*
Hardness above failure layer	2.5	2.75	0.488	–
	2.5	*2.5*	*0.550*	–
Slab thickness (cm)	30.5	36	**0.005**	≥44
	30.1	*39*	***0.001***	*≥44.5*
Slab average hardness	1.69	1.81	0.261	–
	1.75	*1.84*	*0.309*	–
Maximum slab hardness	3	3.5	0.137	–
	3	*3.5*	*0.144*	–
Slab thickness × average hardness	51.5	68.9	**0.007**	≥149
	56.6	*72.7*	***0.007***	*≥149*
Slab strength index	64.3	85.4	**0.006**	≥91
	68.4	*92.4*	***0.005***	*≥199*
Grain size failure layer (mm)	0.88	1.50	**<0.001**	≥1.13
	0.88	*1.50*	***<0.001***	*≥1.20*
Grain size adjacent layer (mm)	0.75	0.69	0.767	–
	0.75	*0.75*	*0.988*	–
Difference in grain size (mm)	0.38	0.75	**0.005**	≥0.63
	0.38	*0.75*	***0.004***	*≥0.75*
Grain type*	61/74	49/13	**<0.001**	Pers.
(persistent: FC, DH, SH)/non-persistent)	*63/87*	*44/11*	***<0.001***	*Pers.*
Threshold sum	3	4	**<0.001**	≥4
	2.5	*4*	***<0.001***	*≥3*

Asterisk (*) denotes categorical variables. All fractures with a slab of at least 13 cm were analysed. The values in the upper lines include the cases where only one test was made and the first test of each pair for the cases with two tests (135 stable and 62 unstable); the values in the lower lines (*italic*) include the cases where only one test was made and the second test of each pair (150 stable and 55 unstable). The level of significance p (Mann–Whitney U-Test, or Pearson χ^2 statistics for categorical variables) and a threshold determined by tree statistics are given. For significant variables p is shown in bold.

statistically not significant. In the case of the RB score, the location of the fracture in the failure layer had no influence on the test result.

For the ECT, failure layer hardness, failure layer grain size, failure layer grain type and the difference in hardness across the failure interface were highly significant variables, i.e. for each variable the distributions for stable and unstable test results were statistically significantly different (*U*-test). Unstable test results were associated with low failure layer hardness, large and persistent grains in the failure layer, and a large grain size difference across the failure interface (Table 8).

Thus the threshold sum, which includes these four variables (e.g. Schweizer and Jamieson, 2007) seems to be a good indicator of fracture propagation propensity of a known weak layer. A threshold sum ≥ 4 indicated rather unstable conditions, i.e. fracture propagation is favored, whereas values <3 indicated rather stable conditions. The difference in hardness across the failure interface — a variable included in the threshold sum — did not show up as significant variable. Other variables that were also not significant were the hardness of the adjacent layer and the hardness of the layer above the failure layer, whereas a soft layer immediately below the failure layer was associated with unstable results.

For the RB release type, the same four variables as described above for the ECT, were found to be significant variables with partly different thresholds (Table 9). For the RB score, partly other variables were significant than for the ECT and the RB release type. In contrast to the ECT, the location of the fracture in the failure layer, the grain type and size in the failure layer and the threshold sum were not significant

Table 9
Comparison of weak layer and slab properties from stable and unstable test results of the RB.

Variable	Stable median	Unstable median	p	Threshold
Location of fracture in failure	39,8,79	14,2,28	0.908	–
layer* (base, within, top)	*40,8,69*	*13,2,38*	*0.271*	–
Hardness failure layer	1	1	**0.014**	<1.5
	1	*1*	**<0.001**	*<1.5*
Hardness adjacent layer	3	2.5	**0.007**	<3.5
	3	*2.5*	**<0.001**	*<3.5*
Difference in hardness	1.5	1	0.119	–
	1.5	*1*	*0.066*	–
Hardness below failure layer	2.5	2	**0.035**	<2.5
	2.5	*2*	**<0.001**	*<2.5*
Hardness above failure layer	3	2	**0.002**	–
	2.5	*2.5*	*0.099*	–
Slab thickness (cm)	39.5	29.3	**0.001**	<24
	37	*38*	*0.588*	–
Slab average hardness	1.81	1.47	**<0.001**	<1.43
	1.75	*1.73*	*0.633*	–
Maximum slab hardness	3.5	3	**<0.001**	<3.0
	3.5	*3*	***0.003***	*<4.5*
Slab thickness × average hardness	74.2	47.8	**<0.001**	<61
	69	*68*	*0.830*	–
Slab strength index	91.8	53.0	**<0.001**	<72.0
	79.6	*82.2*	*0.876*	–
Grain size failure layer (mm)	1.5	1.56	0.475	–
	1.25	*2*	**<0.001**	*≥1.5*
Grain size adjacent layer (mm)	0.88	0.81	0.574	–
	0.88	*0.88*	*0.387*	–
Difference in grain size (mm)	0.5	0.75	**0.035**	≥1.5
	0.5	*0.88*	**<0.001**	*≥1.5*
Grain type*	80/46	26/18	0.604	–
(persistent: FC, DH, SH)/non-persistent)	*62/55*	*44/9*	**<0.001**	*Pers.*
Threshold sum	3.5	4	0.552	–
	3	*5*	**<0.001**	*≥4*

Asterisk (*) denotes categorical variables. All fractures with a slab of at least 13 cm were analysed. The values in the upper lines are for the score (126 stable and 44 unstable); the values in the lower lines are for the release type (italic, 117 stable and 53 unstable). The level of significance *p* (Mann–Whitney *U*-Test, or Pearson χ^2 statistics for categorical variables) and a threshold determined by tree statistics are given. For significant variables *p* is shown in bold.

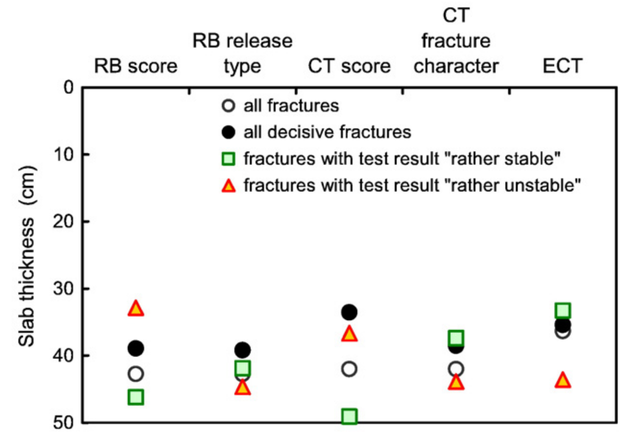

Fig. 5. Average slab thickness vs. stability test results for the RB, the CT and the ECT.

variables that discriminated between rather stable and rather unstable test results. Slab thickness and hardness (and their products) were significant variables for both the ECT and the RB score. However, for the ECT, thin and soft slabs were rather associated with stable results, whereas for the RB score, this type of slab was more often found with unstable results.

Finally, we consider in some more detail the slab thickness (Fig. 5). The average slab thickness for all ECT tests was with 36 cm significantly smaller than the slab thickness found with the RB (43 cm) or the CT (42 cm) ($p \leq 0.002$). If only the decisive (or most critical) fractures were considered, the slab thickness was only slightly thinner for the ECT and the RB, but significantly thinner for the CT ($p \leq 0.001$). The decrease in slab thickness is due to the fact that the test score for different fractures in a CT cannot decrease with increasing slab thickness. As in the case of two identical fracture character results, the decisive fracture was determined with the CT score, a slight (not significant) decrease resulted. The same behaviour can be expected for the RB, but as multiple fractures in a RB test were not frequent, the effect was minor as well.

Considering the slab thickness separately for the tests with stable and unstable results, showed no significant difference in slab thickness between stable and unstable test results for the RB release type and the CT fracture character (Tables 8 and 9, Fig. 5). For the CT and RB score, and the ECT, the slab thickness was a significant variable but with opposite trend. Whereas the slabs were thinner (than the average slabs) with unstable test results for the scores of RB and CT ($p = 0.007$), they were thicker for unstable ECT results ($p = 0.007$). The reduced slab thickness for the unstable RB and CT results seems to be due to the fact that the minimal score usually was found for the thinnest slab.

4.5. Ease of use

For operational use, a stability test has to be easy to perform, and it has to be easy and unambiguous to observe the results. We have not systematically assessed the ease of use, for example, by a questionnaire, but simply report our subjective assessment. In regard to ease of use, the ECT lies in between the RB release type (which is the most simple observation) and the CT fracture character, which some observers had difficulties with. In regard to the time required to do the test, the sequence is the same, but in reverse order: CT, ECT then RB which requires most time.

5. Discussion

In our data set, the ECT showed a well balanced performance with an unweighted average accuracy of slightly more than 80% —

comparable to the performance of the RB test or the combination of rutschblock and threshold sum. However, the performance was significantly lower than reported by Simenhois and Birkeland (2006, 2009-this issue). They found values of the unweighted average accuracy of 98% for their original dataset and 88% for a larger, more diverse dataset. Moner et al. (2008) reported a similarly high unweighted average accuracy of 93% based on a relatively small dataset of 47 test pits.

For the rutschblock and the threshold sum the performance measures can be compared to the results of Schweizer et al. (2008). They reported values of the unweighted average accuracy for RB score of 68%, for the RB release type of 74% and for the threshold sum of 67%. These values are all some lower than the ones in our present study (Table 4) reflecting the large, diverse dataset analyzed by Schweizer et al. (2008).

Due to its low specificity, the performance of the CT was clearly inferior to the ECT and the RB. With an unweighted average accuracy of 60% the performance was lower than the one reported (69%) by van Herwijnen and Jamieson (2007).

We suppose that the reduced performance of the CT and ECT might be due to the fact that in some of the other studies often the most critical weak layer was known a priori. As on our test slopes mostly several potential week layers existed, each of them with an unknown stability, we used for each individual test the result for the most critical weak layer found with the specific test. This approach allowed comparing our estimate of slope stability with the result that every single test would have given at this test location, but it also increased the number of false alarms.

Even if we were unable to reproduce the high performance values reported by Simenhois and Birkeland (2006) and Birkeland and Simenhois (2008), the ECT still separated stable from unstable slopes more reliably than the CT and the threshold sum — and as good as the RB that was slightly favored in our study due to our choice of classifying slopes in regard to stability.

In contrast to Ross and Jamieson (2008) who indicated that a minimal slab thickness of about 30 cm was required for a reliable test, we observed several propagating fractures for slabs that were only a few centimetres thick, in particular for thin wind slabs. Except for soft slabs, the ECT seems appropriate to detect shallow instabilities.

As has been shown previously for the case of the RB test (e.g. Jamieson and Johnston, 1993), the stability assessment became more reliable when results from two adjacent ECTs were combined. For our dataset, it was possible to classify 87% of the slopes with an accuracy of 90%. The remaining 13% of the slopes could not reliably be classified with the ECT. As the frequency of these slopes might depend on the stability distribution of the dataset analysed, a larger and more balanced dataset is needed for a final performance assessment of two adjacent ECTs.

So far the ECT does only discriminate between rather stable and rather unstable conditions based on whether a fracture propagates across the entire column. For operational use, the introduction of an intermediate stability class would be useful, which calls for further work. Adjacent ECTs with contradicting test results might be candidates for this intermediate class.

When different stability tests are performed close together as in our set-up (Fig. 1) we would expect that in most of the pits the same layer would be found as the critical failure layer by the various test — at least on rather uniform slopes. Interestingly, two different types of stability tests performed adjacent to each other, identified only in about half of the cases the same critical failure layer. The relative low agreement scores represent a challenge for any method aiming at automatically identifying potential failure layers in a snowpack without relying on stability test results.

The CT fracture character is assumed to be less sensitive to spatial variations in snowpack properties than the CT score (van Herwijnen and Jamieson, 2002). However, in our dataset the agreement of the critical layer was not better for the fracture character than for the CT score. A generally higher agreement was obtained for two of the same test than for different types of tests. This suggests that differences in failure layer detection might not only be caused by spatial variations of the layer properties, but also by the type of the test and its realization. On slopes with rather unstable conditions, where prominent weak layers are more frequently expected, the agreement in failure layer detection between the tests was higher than on slopes with rather stable conditions.

For CT fracture character, van Herwijnen and Jamieson (2007) found largely the same layer characteristics associated with unstable test results as our analysis has shown for the ECT and the RB release type (Table 10): a thick slab, large grains in the failure layer, large difference in grain size across the failure interface, persistent grain types and a high threshold sum. Also, the hardness of the layer below the failure layer was a significant variable in all these tests, but with opposite sign for the CT fracture character. In general, the weak layer properties associated with unstable test results as indicated by the CT fracture character were more closely related to those indicated by the ECT and the RB release type than to those indicated by RB score.

Although the average slab thickness found with the ECT was lowest compared to the other tests, the slab thickness for the unstable test results were comparable to those obtained with the CT fracture character and the RB release type and thicker than those obtained with the scores of RB and CT. The increase in slab thickness for the unstable ECT results might be caused by the way the surface load (tapping) is distributed within the column which is specific to the ECT. With increasing depth the area of the ECT should become more uniformly loaded though the load decreases, corresponding to the load under a skier (Schweizer and Camponovo, 2001). Due to decreased, but more uniform load in greater depth, only fractures in prominent weak layers will be initialised. These prominent failure layers usually have a high propagation propensity and thus produce unstable test results.

With 38 cm, the mean slab thickness of the unstable ECT fractures was lower than the slab thickness (46 cm) of skier-triggered avalanches as reported by Schweizer and Jamieson (2001). One of the reasons for the relatively thin slabs found in our tests might be that we tended to dig at thin spots where the snowpack is known to be generally weak. Failure layers with depth >1 m are rarely detected by the RB or the ECT. Others tests such as the propagation saw test are better suited to assess deep instabilities (Ross and Jamieson, 2008).

As similar slab and failure layer properties favoured unstable test results (Table 9), our analysis suggests that beside the RB release type and the CT fracture character also the ECT provides primarily information on the fracture propagation propensity. The RB score,

Table 10
Variables significantly related to unstable test results and their trend: “<” indicates that more likely unstable results were found when the value of the variable decreased; “>”indicates that more likely unstable results were found when the value of the variable increased; crosses “X” indicate significant categorical variables.

Variable	ECT	RB release type	RB score
Location of fracture	X		
Hardness failure layer	<	<	<
Hardness adjacent layer		<	<
Hardness below failure layer	<	<	<
Slab thickness (cm)	>		<
Average slab hardness			<
Max slab hardness		<	<
Slab thickness × average hardness	>		<
Slab strength index	>		<
Grain size failure layer	>	>	
Difference in grain size	>	>	>
Grain type	X	X	
Threshold sum	>	>	

which is more indicative for the ease of fracture initiation (Schweizer et al., 2008) was found to have partly different significant variables.

6. Conclusions

We compared various stability tests and indicators of snow instability in particular to assess the performance of the extended column test (ECT) that has recently been developed by Simenhois and Birkeland (2006, 2007). The data contained 146 sets of various tests performed side by side on potential avalanche slopes. They were collected above tree line in the Swiss Alps during the winter of 2007–2008.

Based on our limited dataset we have shown that the ECT was able to well differentiate stable from unstable slopes. By reducing the number of loading taps to 21, the number of false alarms was slightly reduced, so that the specificity and the sensitivity was 82% and 83%, respectively. This means that the portion of false alarms and false stable prediction was similar. The performance was clearly better than for the CT, which was bothered with a low specificity. The unweighted average accuracy for the ECT was about 80%, comparable to the performance of the RB test. However, the stability classification used for our analysis, was partly based on the RB test.

When two different types of stability tests were performed adjacent to each other, in about half of the cases the tests identified the same critical failure layer. On slopes with rather unstable conditions, where prominent weak layers are more frequently expected, the agreement between the tests was higher. Higher agreement was also obtained between the same tests compared to the agreement between different types of tests. The relative low agreement scores represent a challenge for any method aiming at automatically identifying potential failure layers in a snowpack without relying on stability test results.

As has been shown previously for the case of the RB test (e.g. Jamieson and Johnston, 1993), the stability assessment became more reliable when results from two adjacent ECTs were combined. For our dataset, it was possible to classify 87% of the slopes with an accuracy of 90%.

So far the ECT does only discriminate between rather stable and rather unstable conditions based on whether a fracture propagates across the entire column. For operational use, the introduction of an intermediate stability class would be useful, which calls for further work. Adjacent ECTs with contradicting test results might be candidates for this intermediate class.

In most cases, when unstable ECT (or RB release type) test results were observed, the failure occurred at the upper boundary of the failure layer. A soft failure layer consisting of large and persistent grains, and a large difference in grain size across the failure interface favoured unstable test results for the ECT as well as whole block releases with the RB test. Partly different slab and failure properties were associated with low RB scores. Whereas thin and soft slab were associated with lower RB scores, they were more often found with ECT results indicating rather stable conditions. Although the average slab thickness found with the ECT was lowest compared to the other tests, the slab thickness for the unstable test results were comparable to those obtained with the CT fracture character and the RB release type. It seems that the ECT is able to find weak layers with high propagation propensity even if they are somewhat deeper in the snowpack.

As the same slab and failure layer properties favoured unstable test results with the ECT and the RB release type, we suggest that both tests provide information on the fracture propagation propensity, whereas the partly different significant variables for the RB score support the assumption that the RB score is more indicative of the probability of failure initiation.

In terms of ease of use, the ECT did not pose a problem, though requires from the observer some more skills than the RB test. As the ECT is done faster as the RB test, two ECTs can easily be done in the same time. Further work has to show whether this obvious advantage balances the above-mentioned lack of an intermediate stability class.

Finally, snow slope stability evaluation should never rely on the result of single tests whether it is the ECT or any other stability test, but for best results all available information on instability has to be combined.

Acknowledgements

We thank Frank Techel, Martin Oberhammer, Jean-Luc Lugon, Peter Diener, Daniele Degiorgi, Pius Henzen, Martin Hepting, Giovanni Kappenberger, Jörg Kindschi, Luca Silvanti, Giorgo Valenti and numerous colleagues from SLF for snowpack observations. Comments by Christine Pielmeier and two anonymous reviewers helped to improve the manuscript and are gratefully acknowledged.

References

Birkeland, K.W., Johnson, R.F., 1999. The stuffblock snow stability test: comparability with the rutschblock, usefulness in different snow climates, and repeatability between observers. Cold Reg. Sci. Technol. 30 (1–3), 115–123.
Birkeland, K.W., Simenhois, R., 2008. The extended column test: test effectiveness, spatial variability, and comparison with the propagation saw test. In: Campbell, C., Conger, S., Haegeli, P. (Eds.), Proceedings ISSW 2008, International Snow Science Workshop, Whistler, Canada, 21–27 September 2008, pp. 401–407.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1998. Classification and Regression Trees. CRC Press, Boca Raton, U.S.A.. 368 pp.
CAA, 2002. Observation Guidelines and Recording Standards for Weather, Snowpack and Avalanches. Canadian Avalanche Association (CAA), Revelstoke BC, Canada. 78 pp.
Doswell, C.A., Davies-Jones, R., Keller, D.L., 1990. On summary measures of skill in rare event forecasting based on contingency-tables. Weather Forecast. 5 (4), 576–585.
Föhn, P.M.B., 1987. The Rutschblock as a practical tool for slope stability evaluation. In: Salm, B., Gubler, H. (Eds.), Symposium at Davos 1986 — Avalanche Formation, Movement and Effects, IAHS Publ., 162. InInternational Association of Hydrological Sciences, Wallingford, Oxfordshire, U.K., pp. 223–228.
Gauthier, D., Jamieson, J.B., 2006. Towards a field test for fracture propagation propensity in weak snowpack layers. J. Glaciol. 52 (176), 164–168.
Gauthier, D., Jamieson, B., 2008a. Evaluation of a prototype field test for fracture and failure propagation propensity in weak snowpack layers. Cold Reg. Sci. Technol. 51 (2–3), 87–97.
Gauthier, D., Jamieson, B., 2008b. Fracture propagation propensity in relation to snow slab avalanche release: validating the Propagation Saw Test. Geophys. Res. Lett. 35 (13), L13501.
Gauthier, D., Jamieson, J.B., 2008c. Predictions of the propagation saw test: comparisons with other instability tests at skier tested slopes. In: Campbell, C., Conger, S., Haegeli, P. (Eds.), Proceedings ISSW 2008, International Snow Science Workshop, Whistler, Canada, 21–27 September 2008, pp. 408–414.
Greene, E. (Ed.), 2004. Snow, Weather and Avalanches: Observational Guidelines for Avalanche Programs in the United States. InAmerican Avalanche Association (AAA), Pagosa Springs CO, U.S.A., 136 pp.
Jamieson, J.B., 1999. The compression test — after 25 years. Avalanche Rev. 18 (1), 10–12.
Jamieson, J.B., Johnston, C.D., 1993. Rutschblock precision, technique variations and limitations. J. Glaciol. 39 (133), 666–674.
Jamieson, J.B., Johnston, C.D., 1995. Interpreting rutschblocks in avalanche start zones. Avalanche News 46, 2–4.
Jamieson, J.B., Schweizer, J., 2005. Using a checklist to assess manual snow profiles. Avalanche News 72, 57–61.
McCammon, I., Schweizer, J., 2002. A field method for identifying structural weaknesses in the snowpack. In: Stevens, J.R. (Ed.), Proceedings ISSW 2002. International Snow Science Workshop, Penticton BC, Canada, 29 September–4 October 2002, pp. 477–481.
Moner, I., Gavaldà, J., Bacardit, M., Garcia, C., Martí, G., 2008. Application of field stability evaluation methods to the snow conditions of eastern Pyrenees. In: Campbell, C., Conger, S., Haegeli, P. (Eds.), Proceedings ISSW 2008, International Snow Science Workshop, Whistler, Canada, 21–27 September 2008, pp. 386–392.
Pielmeier, C., Marshall, H.P., 2009. Rutschblock-scale snowpack stability derived from multiple quality-controlled SnowMicroPen measurements. Cold Reg. Sci. Technol. 59, 178–184 (this issue).
Ross, C., Jamieson, J.B., 2008. Comparing fracture propagation tests and relating test results to snowpack characteristics. In: Campbell, C., Conger, S., Haegeli, P. (Eds.), Proceedings ISSW 2008, International Snow Science Workshop, Whistler, Canada, 21–27 September 2008, pp. 376–385.
Schweizer, J., 2002. The Rutschblock test — procedure and application in Switzerland. Avalanche Rev. 20 (5), 14–15 1.
Schweizer, J., Camponovo, C., 2001. The skier's zone of influence in triggering of slab avalanches. Ann. Glaciol. 18, 193–198.
Schweizer, J., Jamieson, J.B., 2001. Snow cover properties for skier triggering of avalanches. Cold Reg. Sci. Techn. 33 (2–3), 207–221.
Schweizer, J., Lütschg, M., 2001. Characteristics of human-triggered avalanches. Cold Reg. Sci. Technol. 33 (2–3), 147–162.

Schweizer, J., Wiesinger, T., 2001. Snow profile interpretation for stability evaluation. Cold Reg. Sci. Technol. 33, 179–188.
Schweizer, J., Jamieson, J.B., 2003. Snowpack properties for snow profile analysis. Cold Reg. Sci. Technol. 37 (3), 233–241.
Schweizer, J., Jamieson, J.B., 2007. A threshold sum approach to stability evaluation of manual snow profiles. Cold Reg. Sci. Technol. 47 (1–2), 50–59.
Schweizer, J., Jamieson, J.B., Schneebeli, M., 2003. Snow avalanche formation. Rev. Geophys. 41 (4), 1016. doi:10.1029/2002RG000123.
Schweizer, J., McCammon, I., Jamieson, J.B., 2008. Snowpack observations and fracture concepts for skier-triggering of dry-snow slab avalanches. Cold Reg. Sci. Technol. 51 (2–3), 112–121.
Seligman, G., 1936. Snow Structure and Ski Fields. MacMIllan, London, UK. 555 pp.
Sigrist, C., Schweizer, J., 2007. Critical energy release rates of weak snowpack layers determined in field experiments. Geophys. Res. Lett. 34 (3), L03502. doi:10.1029/2006GL028576.
Simenhois, R., Birkeland, K.W., 2006. The extended column test: a field test for fracture initiation and propagation. In: Gleason, J.A. (Ed.), Proceedings ISSW 2006, International Snow Science Workshop, Telluride CO, U.S.A., 1–6 October 2006, pp. 79–85.
Simenhois, R., Birkeland, K.W., 2007. An upgrade on the extended column test: new recording standards and additional data analyses. Avalanche Rev. 26 (2).
Simenhois, R., Birkeland, K.W., 2009. The extended column test: test effectiveness, spatial variability, and comparison with the propagation saw test. Cold Reg. Sci. Technol. 59, 210–216 (this issue).
SYSTAT, 2007. SYSTAT® Statistical Software User Manual, San Jose CA, U.S.A.
van Herwijnen, A., Jamieson, B., 2002. Interpreting fracture character in stability tests. In: Stevens, J.R. (Ed.), Proceedings ISSW 2002. International Snow Science Workshop, Penticton BC, Canada, 29 September–4 October 2002, pp. 514–522.
van Herwijnen, A., Jamieson, B., 2007. Fracture character in compression tests. Cold Reg. Sci. Technol. 47 (1–2), 60–68.