Snowpack tests for assessing snow-slope instability

Jürg SCHWEIZER,[1] J. Bruce JAMIESON[2,3]

[1]*WSL Institute for Snow and Avalanche Research SLF, Flüelastrasse 11, CH-7260 Davos Dorf, Switzerland*
E-mail: schweizer@slf.ch
[2]*Department of Civil Engineering, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*
[3]*Department of Geoscience, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada*

ABSTRACT. Information on snowpack instability is crucial for assessing avalanche risk in backcountry operations as well as for operational forecasting of the regional avalanche danger. Since slab avalanche release requires both fracture initiation and fracture propagation in a weak snowpack layer, field observations should ideally provide reliable information on the probability or propensity of both fracture processes. Even simple field observations that do not require digging a snow pit can provide useful information. Traditional snowpack tests include the shovel shear test, the shear frame test, the compression test (CT) and the rutschblock test (RB). Interpretation of the test results for the CT and RB has been improved by considering the appearance or type of the fracture in addition to the score. More recently, two tests have been developed that focus on fracture propagation rather than initiation: the extended column test (ECT) and the propagation saw test (PST). We compare the sensitivity, specificity and unweighted average accuracy of various stability tests. Comparative studies indicate that the RB, ECT and PST have comparable accuracy. For most test methods the unweighted average accuracy of a single test was 70–90% depending on the dataset. Test methods such as the RB, ECT and PST, which fracture an area large enough to include fracture propagation, are generally more accurate than test methods that fracture smaller areas (e.g. the CT). The threshold-sum method was also less accurate. Even with very experienced observers for the RB, ECT and PST an error rate of at least about 5–10% has to be expected. Performing a second, adjacent test on the same slope improves test reliability.

INTRODUCTION

Snow stability data are, as avalanche occurrence data, most closely related to avalanche release probability, which is the key parameter to be forecasted by any avalanche-forecasting service or backcountry operation. In the context of avalanche forecasting, observations that provide information on snow stability have been termed low-entropy data (LaChapelle, 1980) or Class I data (McClung and Schaerer, 2006).

Snow avalanches are, unlike most other mass movements, to a certain degree predictable (Schweizer, 2008). Even more important, in the case of snow avalanches direct observations can be made that provide information on the probability of a mass movement within the next few days. No simple in situ tests exist to assess, for example, the landslide risk. In that respect, snow stability tests are unique. In this paper, we call any test that is made in situ and provides information on snowpack instability simply a 'snow stability test'.

These tests are primarily indicators of whether triggering by localized dynamic loading (e.g. people or explosives), is likely. This is appropriate because most fatalities are caused by human triggering (e.g. Schweizer and Lütschg, 2001).

Besides simple field observations on avalanche occurrence, shooting cracks and whumpfs (i.e. sound of collapsing weak layers), which all indicate instability (e.g. Jamieson and others, 2009), a variety of in situ tests has been developed over the last five decades. All these tests aim to determine whether the sloping snowpack is stable. None of them provides the definitive answer. The reasons for the insufficiency of the tests are mainly the inherent limitations of the test (and its loading method) in replicating the avalanche release process, the spatially variable nature of the mountain snowpack (e.g. Schweizer and others, 2008a) and the complexity of the avalanche release process. Repeatedly, the question arose as to whether the tests are useful at all.

In this paper, we describe test requirements, review the most common existing tests and assess their performance. Test performance is assessed primarily based on a number of studies that compared various snow stability tests (e.g. Gauthier and Jamieson, 2008b; Moner and others, 2008; Simenhois and Birkeland, 2009; Winkler and Schweizer, 2009).

REQUIREMENTS

It is generally perceived that snow stability tests should determine whether a particular slope is stable or unstable. This seems a rather unrealistic aim. First of all, when a test is made on an avalanche slope, that slope is assumed stable, otherwise the field team doing the test might have been caught in an avalanche and should not have entered the slope. Consequently, tests are often made on small slopes that do not avalanche or on slopes with a steepness $<30^\circ$. By doing so, one has to assume that the conditions found on this relatively safe slope are representative of larger and/or steeper slopes of similar aspect in the surroundings. However, due to the spatially variable nature of the snowpack, extrapolation is not straightforward and requires experience. Therefore we suggest that a snow stability test cannot provide the ideal information, that is whether a slope is stable.

For the release of a dry-snow slab avalanche a number of requirements have to be fulfilled. These are: (1) the slope has to have a minimal slope angle ($\sim 30^\circ$); (2) the snow layering has to be such that a cohesive slab layer overlies a weak layer; (3) this slab/weak-layer stratigraphy has to exist over a minimal extent of several tens of m^2; (4) the snowpack has to

be in a metastable condition, i.e. the strength of the weak layer is, at least at some locations, of similar magnitude to the applied stress; (5) there is an external (or internal) trigger present; (6) an initial failure tends to propagate; and (7) the slab breaks up and slides downslope, i.e. friction is overcome. The requirements that can be checked are (1), (2), (4) and (6). A trigger (5) inherently exists when applying a stability test. Information on requirement (3) is usually not available, but occasionally can be estimated if the slab/weak-layer combination and its formation are known, and (7) is given in most cases on large and steep slopes.

In the context of fractures within weak snowpack layers, we distinguish between fracture initiation and fracture propagation as in Schweizer and others (2003). A fracture or crack initiates if a localized dynamic perturbation (e.g. a trigger such as a skier or tapping on the top of a snow column) causes a crack in a weak layer. A crack or fracture that does not advance (propagate) beyond the influence of the localized perturbation is subcritical. A crack or fracture that advances beyond the influence of the localized perturbation has started to propagate. Fractures that were initiated but did not propagate across the slope were documented, for example, by Van Herwijnen and Jamieson (2005).

In summary, the snow stratigraphy (slab over weak layer), failure initiation and fracture propagation are the most important requirements (apart from steepness) for a dry-snow slab avalanche (McCammon and Sharaf, 2005). A snow stability test needs to provide information on whether (or better, to what extent) these requirements are fulfilled. In addition, the test should not be difficult to perform, not require special equipment, be completed in less than ~30 min and provide robust repeatable results.

Not much is known about the application of snow stability tests in wet snow, and all results described below refer to dry snow conditions.

SNOW STABILITY TESTS

We consider two types of observation that provide information on snowpack instability: (1) observations that do not require digging (Jamieson and others, 2009); and (2) observations that require digging a snow pit.

The first category includes whumpfs (sudden failure of the weak layer due to rapid localized loading manifesting itself by collapse), shooting cracks and recent avalanching. These three observations are all unambiguous signs of instability. Further simple observations are, for example, the ski-pole test (Tremper, 2008) and cracking at skis.

The second category includes the snow profile and all tests that in one way or another apply an additional load to the snowpack to induce a failure. The latter include the shear frame test (Roch, 1966; Jamieson and Johnston, 2001), the shovel shear test (McClung and Schaerer, 2006), the hand shear test (Tremper, 2008), the rutschblock test (RB; Föhn, 1987a), the compression test (CT; Jamieson, 1999), the extended column test (ECT; Simenhois and Birkeland, 2006) and the propagation saw test (PST; Gauthier and Jamieson, 2006). The detailed procedures for all these tests are described in Greene and others (2009).

The investigation of snow stratigraphy can be combined/quantified with a structural instability index such as the threshold sum (Schweizer and Jamieson, 2007). In addition, we include the snow micro-penetrometer (SMP; Schneebeli and Johnson, 1998) in our comparison. It records a penetration resistance–depth profile and potentially can be used to assess snowpack instability (Bellaire and others, 2009).

The shovel (or hand) shear test attempts to shear off the slab above the weak layer and hence provides an index of shear strength. It is primarily used to find weak layers rather than to assess weak-layer strength. However, to load the column properly, the weak-layer depth has to be known. The application of the load is delicate and the rating of the load at failure is highly subjective. The shovel shear test has not been validated, whereas the hand shear test has been correlated with local avalanche danger (Jamieson and others, 2009) but not with slope scale instability.

The shear frame test is the only in situ measurement method that measures shear strength. The weak layer to be tested has to be identified by other means such as a snow profile. The slab is removed apart from a few centimetres just above the weak layer. The shear frame test and the stability indices derived from it (Föhn, 1987b) were shown to be related to snowpack instability (Jamieson and Johnston, 1998; Jamieson and others, 2007). Shear frame measurements from study plots are particularly useful to monitor the temporal evolution of snow stability.

The RB, which can be considered the grandfather of all snowpack stability tests, involves isolating a snow column of 2.0 m (cross-slope) × 1.5 m (upslope). The block is then loaded in stages by a skier until slab failure. The loading step or score (1–7) is noted as well as the release type, i.e. the proportion of the block that released (whole block, most of block, edge only). It has been shown that the RB score is related to the probability of skier-triggered avalanches (Föhn, 1987a; Jamieson, 1995) on the adjacent slope. The RB release type is assumed to be related to fracture propagation propensity, in particular since the RB area (3 m^2) is – except for deep weak layers – larger than the area for which the skier's load is significant (~1 m^2) (Schweizer and Camponovo, 2001). The fact that Schweizer and others (2008b) found a substantially higher sensitivity for the RB release type (81%) than for the RB score (61%) likely supports this assumption.

With the CT a much smaller area (30 cm × 30 cm) is loaded by tapping onto the isolated column. The CT score was related to the probability of skier-triggered avalanches on adjacent slopes (Jamieson, 1999) and the CT score can be related to the RB score. By introducing the fracture character (Van Herwijnen and Jamieson, 2007), the interpretation of the CT was improved and weak-layer/slab properties associated with sudden fractures (equivalent to Q1 shear quality as introduced by Johnson and Birkeland, 1998) suggest that the fracture character is related to fracture propagation propensity (Van Herwijnen and Jamieson, 2007b).

A number of other small column tests have been developed that all aim at replacing the subjective tapping onto the isolated column by a more quantitative loading procedure. Among those are the rammrutsch (or drop hammer) test (Schweizer and others, 1995; Stewart and Jamieson, 2002), the stuffblock test (Birkeland and Johnson, 1999) and the quantified loaded column test (Landry and others, 2001). Since these are variations of the compression test and most have limited validation data, we do not include them in our analysis.

The recently developed ECT was introduced as a test that should provide information on fracture initiation and propagation (Simenhois and Birkeland, 2006). The ECT involves isolating a column of 30 cm × 90 cm (with the longer side

Table 1. Information on dry-snow slab avalanche formation provided by various observations or tests that do not require digging a snow pit

Observation/test	Stratigraphy (weak layer below slab)	Failure initiation	Fracture propagation
Whumpfs	Yes	Yes	Yes
Shooting cracks	Yes	Yes	Yes
Recent avalanching	Yes	Yes	Yes
Cracking at skis	Partly	No	No
Ski-pole test	Partly	No	No

Table 2. Information on dry-snow slab avalanche formation provided by various observations or tests that do require digging a snow pit

Observation/test	Stratigraphy (weak layer below slab)	Failure initiation	Fracture propagation
Snow profile	Yes	No	No
Snow profile + threshold sum	Yes	Partly	Partly
Shear frame test	No	Yes	No
Shovel (hand) shear test	Partly	Yes	No
RB: score + release type	Yes	Yes	Yes
CT: score + fracture character	Yes	Yes	Partly
ECT	Yes	Yes	Yes
PST	No	Partly	Yes

cross-slope) and loading it in one corner, as with the CT. It is noted whether a fracture crosses the entire column. Simenhois and Birkeland (2009) showed that the ECT is highly indicative of snowpack instability on nearby slopes.

The PST was inspired by traditional beam-type tests. It is a fracture mechanical test in which the resistance of a material to fracture in the presence of a crack is tested. In a fracture mechanical test, either the sample is loaded until failure for a given crack length, or the crack length is continuously increased (under constant load) until failure occurs. Gauthier and Jamieson (2006) and Sigrist and Schweizer (2007) were the first to report on a suitable design for a field test. A snow column (30 cm × ~100 cm) is isolated with the longer side upslope. The length should be at least 100 cm or the slab thickness, whichever is longer. After the weak layer is identified by a separate test (e.g. the CT or profile), a cut is made with a snow saw along the weak layer until the crack length becomes critical and self-propagation of the crack starts. The critical crack length is noted and whether the crack propagates to the end of the column. As the free surface influences fracture propagation, D. McClung (personal communication, 2009) suggested not isolating the column at its upslope end. Gauthier and Jamieson (2008a) validated the PST and showed that at sites where weak-layer fracture initiation was confirmed on adjacent slopes, PST results were clearly related to observations of fracture propagation.

TEST INDICATION

Below we rate the above-described observations of snowpack instability with regard to the three principal requirements outlined above: (1) slab/weak-layer stratigraphy; (2) failure initiation; and (3) fracture propagation.

Table 1 compiles the ratings for the simple observations that do not involve digging. Whereas whumpfs, shooting cracks and recent avalanching all indicate that the three requirements are fulfilled, the ski-pole test might at best show the existence of a (thick) weak layer (e.g. when a cohesionless layer (often consisting of depth hoar) is found below a slab). Cracking at skis only indicates that the surface layer is cohesive, but does not provide information on a possible weak layer.

A snow profile provides snow stratigraphy and shows whether a weak layer below a slab exists (Table 2). Whereas the shear frame test only indicates the strength of the weak layer, the shovel shear test is in addition partly suited for identifying weak layers. For the three tests that involve loading an isolated snow column, the slab/weak-layer stratigraphy is shown implicitly when the column fails. In addition, the RB (score and release type) and the ECT both provide information on initiation and propagation. For the CT the information on fracture propagation is less well related to fracture propagation than for the RB and ECT. Finally, the PST is clearly an index of fracture propagation propensity, but gives no indication on stratigraphy and limited indication on failure initiation. However, digging the pit and sawing might provide some indirect information on stratigraphy.

TEST LIMITATIONS

The use of the tests is influenced by their reliability (see below) and their practicality. Table 3 summarizes some key practical limitations of the tests including the time requirement, required slope angle, effective depth and required technical skill level.

Other than the snow profile, which accompanies most tests, the RB is the slowest of the tests and the only one to require a sufficiently steep slope. The hand shear test is fast but is limited to weak layers within about 45 cm of the snow surface. The RB, CT, ECT, PST and SMP are all indicative in the 30–70 cm range, which is important for skier-triggered dry-snow slab avalanches. Ross and Jamieson (2008) report the ECT to be reliable up to about 70 cm in the typically soft snow of the Columbia Mountains of western Canada, while Simenhois and Birkeland (2009) report indicative results up to about 100 cm in snow climates with wind-stiffened slab layers. The hand shear test, RB, CT and ECT have the advantage of requiring the least skill, whereas the snow profile, shear frame test and SMP require the most skill. The SMP is the only test mentioned in this study that requires expensive electromechanical equipment.

TEST ACCURACY

Early results relating the RB score to the probability of skier-triggered avalanches on nearby slopes have clearly shown that even for high RB scores of 6 or 7 occasionally skier-triggered avalanches were observed (e.g. Jamieson, 1995). These false-stable predictions have shown that snow stability tests are not foolproof and hence that decisions on where and when to travel in avalanche terrain should never rely solely on stability test results. The false predictions have been attributed to spatial variations of snowpack stability but may also be related to differences between the slab-release process and the loading or support in the stability test.

Table 3. Key practical limitations of tests including time requirement, required slope angle, effective depth and required technical skill level

Test	Time	Slope	Depth	Technical skill
	min	°	cm	
Snow profile	>30	Any	Unlimited	High
Shear frame*	>15‡	Any	Unlimited	High
Shovel shear*	10‡	Any	Unlimited	Moderate
Hand shear	5	Any	<45	Low
RB	25‡	>25	30–90	Low
CT	10‡	Any§	<100	Low
ECT	15‡	Any§	30–70¶	Low
PST*	15‡	Any§	>30	Moderate
SMP†	15	Any	<150	High

*Difficult for layers that are hard to find in a snow profile.
†Expensive and not used by practitioners.
‡Best done near a snow profile. The time for the snow profile is not included.
§Results are easier to see on steeper slopes.
¶Ross and Jamieson (2008). Simenhois and Birkeland (2009) report reliable results down to about 100 cm.

When analysing the performance of snow stability tests, observed stability is compared with predicted stability. In most cases, only two categories of instability were considered: stable and unstable. For example, RBs were performed on skier-triggered (unstable) as well as skier-tested (stable) slopes and RB scores <4 were considered as unstable and ≥4 as stable. This type of analysis simplifies the comparison, but oversimplifies the problem. Nevertheless, we follow this approach and report the test performance by providing the probability of detection (POD, also called sensitivity), the probability of null events (PON, also called specificity) (Doswell and others, 1990) and their mean, i.e. the unweighted average accuracy (RPC):

sensitivity:

$$\mathrm{POD} = \frac{\text{predicted unstable slopes}}{\text{all observed unstable slopes}} \quad (1)$$

specificity:

$$\mathrm{PON} = \frac{\text{predicted stable slopes}}{\text{all observed stable slopes}} \quad (2)$$

unweighted average accuracy:

$$\mathrm{RPC} = \frac{\text{sensitivity} + \text{specificity}}{2} \quad (3)$$

A test should have both a high sensitivity and a high specificity. A high sensitivity means that most unstable situations are detected. If the specificity is high as well, then there are only few false alarms. A high sensitivity combined with a low specificity means that the test is oversensitive and produces many false alarms. Whereas false alarms have less severe consequences than misses, a low specificity is not desired since this will promote overcautious decisions which for regional forecasting will lead to a credibility problem in the long run (Williams, 1980). On the other hand, as snow stability tests are commonly used to seek instability rather than stability (McClung, 2002), an unbalanced performance with sensitivity larger than specificity is better than with specificity larger than sensitivity.

To assess the performance of snowpack tests, we consider various datasets (Table 4), primarily recent comparative studies. All datasets include a stability test score and an observed stability. The definition of observed stability may vary. For example, in Schweizer and others (2008b) unstable refers to slopes that were skier-triggered. So the tests were made near the perimeter of a skier-triggered avalanche. In most other datasets slopes were rated as unstable if recent avalanches were observed on adjacent slopes, a whumpf was triggered on the test slope or any other sign of instability was observed. Occasionally, as in the study by Winkler and Schweizer (2009), the slope was also rated as unstable based on an additional criterion, that is whether the profile was rated as poor (Schweizer and Wiesinger, 2001). This rating may favour the RB over the other test methods. For most datasets, at least two different tests were performed on the same test slope, thus allowing comparison of the relative performance of the tests (e.g. Gauthier and Jamieson, 2008b; Moner and other, 2008; Simenhois and Birkeland, 2009; Winkler and Schweizer, 2009). For the dataset presented by Schweizer and others (2008b), only RB results are available. Dataset A contains test results from slopes where several CTs were performed adjacent to an RB by University of Calgary avalanche researchers. These so far unpublished data were collected in the Columbia Mountains of western Canada between December 1996 and March 2008. For all datasets, the predicted stability for a given test method is based on a threshold value. Depending on the test score, the test result is rated either as stable or unstable (Table 5).

Table 4. Datasets used to assess test performance. Source, type of snowpack tests performed and the number of test slopes (*N*) are given. The last column indicates how test slopes were classified as unstable: (1) avalanche on test slope; (2) avalanche on adjacent slope; (3) signs of instability on test slope; (4) profile rated as 'very poor' or 'poor' according to Schweizer and Wiesinger (2001); (5) stability rating 'very poor' or 'poor' according to Simenhois and Birkeland (2009)

Dataset	Source	Tests	*N*	Observation of instability
A	Unpublished data (see text)	RB, CT	139	1, 3
B	Schweizer and others (2008b)	RB	512	1
C	Winkler and Schweizer (2009)	RB, CT, ECT	146	2, 3, 4
D	Gauthier and Jamieson (2008b)	RB, CT, PST	See Table 6	1
E	Simenhois and Birkeland (2009)	ECT	311	5
E1	Simenhois and Birkeland (2009)	ECT, PST	78	1, 3
F	Moner and others (2008)	RB, ECT	63	1
G	Bellaire and others (2009)	SMP	60	2, 3, 4

Table 5. Critical ranges or thresholds to rate stability test results with regard to stability (stable/unstable)

Test	Unstable	Stable
RB score	≤3	≥4
RB release type	Whole block release	Partial release
CT score	<14	≥14
CT fracture character	Sudden planar (SP), sudden collapse (SC)	Others
ECT	Fracture crosses entire column within one tap of initiation	Others
PST	<50% cut length and first propagation reaches end of column	≥50% cut length or first propagation stops before end of column
Threshold sum	≥5	≤4

Table 6 compiles the performance measures for the datasets presented in Table 4. As the datasets have distinct properties (e.g. in terms of design and circumstances) and not all test methods were included in each dataset, it is not meaningful to calculate an average performance for a given test method. Instead, one way of assessing the different test methods is to compare their unweighted average accuracy with that of the RB within the same dataset (where the sample size is large or similar), which provides the relative performance (Table 7). To check for differences in the performance of the various tests, the two-proportion *Z*-test (Spiegel and Stephens, 1999) was used.

In dataset A, the unweighted average accuracy of the RB *or* release type, the CT, as well as the CT *and* fracture character are within 0.01 of the value for the RB; for the RB *and* release type the accuracy is 0.09 lower; however, the difference is not statistically significant ($p = 0.36$). In dataset B, the unweighted average accuracy for the threshold sum is comparable to that for the RB; however, the value for the RB *and* release type is significantly higher ($p = 0.01$), while the value for the RB *or* release type is 0.04 (not significantly) higher ($p = 0.34$). In dataset C, the accuracy for the RB *and* release type is 0.04 lower, for the RB *or* release type and the ECT it is comparable, while the value for the CT is 0.16 lower and for the threshold sum is even lower – only the latter two differences are statistically significant. In dataset D, the unweighted average accuracy for the RB *and* release type is 0.06 lower, the value for the RB *or* release type

Table 6. Performance of various snow stability tests for the datasets described in Table 4. For the threshold sum, the threshold values ≥5 and ≥4 are considered (the results for the latter are given in parentheses). The base rate gives the proportion of unstable observations

Test	Dataset	*N*	Base rate	Sensitivity	Specificity	Unweighted average accuracy
RB: score	A	139	0.540	0.48	0.84	0.66
	B	457	0.444	0.61	0.73	0.67
	C	146	0.250	0.78	0.90	0.84
	D	23	0.700	0.63	0.86	0.74
	F	62	0.440	0.74	0.77	0.76
RB: score *and* release type	A	33	0.480	0.44	0.71	0.57
	B	185	0.330	0.69	0.85	0.77
	C	146	0.250	0.61	0.99	0.80
	D	23	0.700	0.50	0.86	0.68
	F	29	0.450	0.69	1.00	0.85
RB: score *or* release type	A	33	0.480	0.88	0.47	0.67
	B	185	0.330	0.89	0.53	0.71
	C	146	0.250	0.94	0.75	0.84
	D	23	0.700	0.75	0.71	0.73
	F	29	0.450	1.00	0.75	0.88
CT: score	A	139	0.540	0.52	0.81	0.67
	C	146	0.250	0.90	0.45	0.68
	D	58	0.710	0.63	0.47	0.55
CT: score *and* fracture character	A	33	0.480	0.56	0.76	0.66
	C	146	0.250	0.93	0.56	0.75
	D	58	0.710	0.63	0.65	0.64
ECT	C	146	0.250	0.83	0.79	0.81
	E	311	0.402	0.94	0.82	0.88
	E1	78	0.580	1.00	0.91	0.95
	F	47	0.380	0.89	0.97	0.93
PST	D	187	0.604	0.70	0.88	0.79
	E1	78	0.580	0.56	1.00	0.78
Threshold sum ≥5 (≥4)	B	426	0.502	0.50 (0.74)	0.81 (0.58)	0.66 (0.66)
	C	146	0.250	0.86	0.38	0.62
	D	27	0.630	0.88	0.50	0.69
SMP	G	60	0.380	0.78	0.76	0.77

Table 7. Unweighted average accuracy for the various datasets and test methods. Same data as in Table 6 but compiled with the dataset as main entry to facilitate within dataset comparison. Only datasets that include several test methods are shown. n/a: not applicable

Dataset	RB: score	RB: score *and* release type	RB: score *or* release type	CT: score	CT: score *and* fracture character	ECT	PST	Threshold sum ≥5
A	0.66	0.57	0.67	0.67	0.66	n/a	n/a	n/a
B	0.67	0.77	0.71	n/a	n/a	n/a	n/a	0.66
C	0.84	0.80	0.84	0.68	0.75	0.81	n/a	0.62
D	0.74	0.68	0.73	0.55	0.64	n/a	0.79	0.69
E1	n/a	n/a	n/a	n/a	n/a	0.95	0.78	n/a
F	0.76	0.85	0.88	n/a	n/a	0.93	n/a	n/a

is comparable to that of the RB, the value for the CT is 0.19 lower and for the threshold sum is 0.05 lower than for the RB (while the PST is based on a different sample size). None of the observed differences in unweighted average accuracy in dataset D are statistically significant (due to the partly low sample size). Within dataset E1, the accuracy for the ECT is 0.17 (significantly) higher than for the PST ($p = 0.002$). Finally, in dataset F, the unweighted average accuracy for the ECT is 0.17 (significantly) higher than for the RB ($p = 0.01$).

In summary, only for dataset C is the sample size sufficiently large, and several tests were included to allow a broader comparison. Based on dataset C, the RB and ECT have similar accuracy. On the other hand, dataset E1 suggests that the ECT performs better than the PST, and based on dataset F it seems that the ECT performs better than the RB. Finally, dataset D suggests that the RB and the PST have similar accuracy. Though datasets D, E1 and F are fairly small, the problem of the obviously conflicting conclusions from the various studies cannot be resolved, in particular since none of the studies included all tests and the datasets have partly distinct properties. Based on the available studies and recognizing the differences between datasets, we therefore conclude that the RB, the ECT and the PST have similar accuracy and that the CT and threshold sum are less accurate. Below we report on some of the specific properties of the tests.

If the RB score alone is considered, the RB shows a low false alarm ratio (high specificity) but misses quite a number of unstable situations. The prediction can be improved largely by considering the release type as well (as has been shown by Winkler and Schweizer, 2009). With an RB score ≥4 and only a partial release of the block, the conditions are very likely (99%) rather stable, whereas unstable conditions can be expected (94%) if either the RB score is low (<4) or the whole block is released. These findings have been confirmed by Moner and others (2008). In the case of the CT, considering the fracture character only moderately improves the performance of the CT, which shows a high false alarm ratio, that is the CT is oversensitive.

The ECT shows a very balanced performance and, according to Simenhois and Birkeland (2009), has the best unweighted average accuracy of all tests. However, the results from the study by Hendrikx and others (2009) indicate that under some conditions (which cannot be specified yet) the ECT can also be less accurate (about 40%).

The PST does indicate that propagation is unlikely for quite a number of unstable conditions, but shows an unweighted average accuracy of about 80%, comparable with most other tests.

With an unweighted average accuracy of almost 80%, the performance of the SMP is not much lower than the accuracy of the traditional snow stability tests. However, in contrast to most validation studies of the RB, CT, ECT and PST, Bellaire and others (2009) used the same data to establish the instability criterion as for the accuracy.

Given that most studies involved many observers with varying experience under a variety of different conditions and revealed test accuracies of 70–90%, it is apparent that, even with a very experienced observer, in at least about 5–10% of the cases snow stability will not be predicted correctly by a single stability test.

SOURCES OF ERROR

The relatively high number of false predictions undermines the usefulness of snow stability tests. What causes the false predictions? We propose that there are at least two sources of error. The first is related to the test method, the second to the variable nature of the snowpack. Obviously, all test methods are relatively crude methods that involve many subjective elements such as the way of loading. The support or tested area of some tests (e.g. the CT or shear frame) is too small to capture fracture propagation. Any test result will be specific for the test location since the slab as well as weak-layer properties may vary within the slope and be different on adjacent slopes. Hence an individual test result may well under- or overestimate stability. At present, the contribution of the two sources of error to the overall rate of false predictions is unclear. It has been suspected, for example by Schweizer and others (2008a), that, due to test errors, spatial variations of snow stability cannot be detected easily; this would mean that the two errors have similar magnitude.

Spatial variability studies that used snow stability tests in conjunction with the SMP may shed some light on the source of errors. Kronholm (2004) has provided the quartile coefficient of variation (QCV) for the stability test results as well as the weak-layer strength penetration resistance. The QCV of the stability test scores was in most cases about 30%, whereas it was only about 20% for the weak-layer penetration resistance. As the SMP is considered a high-precision instrument that has produced repeatable results, it can be assumed that at least about one-third of the observed variation in stability test scores was related to test errors and about two-thirds may reflect the real spatial variability of the snowpack. However, it has to be pointed out that the two methods have very different support: 0.09 m^2 (CT) vs

$2 \times 10^{-5}\,m^2$. In fact, one would expect the variation to increase with decreasing support. On the other hand, the variation in stability includes variations of both weak-layer and slab properties so that it is expected to be higher than the variation of weak-layer penetration resistance. The use of stability tests for spatial variability studies seems questionable given the obviously significant test error, but so far no alternative exists.

The uncertainty due to test errors can be decreased substantially if two tests adjacent to each other are conducted. In the case of the RB, Jamieson (1995) has shown that in 97% of cases the test result is within a ± 1 score of the slope median, at least on rather uniform slopes. This implies that the probability of the median of two independent tests being within one-half or one step of the slope median score is 0.91 or 0.99, respectively.

Winkler and Schweizer (2009) found that with two adjacent ECTs which provide the same test result, the unweighted average accuracy increased from about 80% to about 90%. Birkeland and Chabot (2006) proposed that the false-stable error rate could be reduced from about 10% to about 1% by making a second test at a representative site beyond the correlation length from the first test and choosing the less stable of the two test results. As the correlation length is unknown, at least about 10 m has been proposed as the distance between two tests (Jamieson and Johnston, 1993; Schweizer and others, 2008a).

CONCLUSIONS

Snow stability tests represent highly prized Class I data. They can be considered as indices of instability. They represent the only way to obtain information on: (1) layering; (2) failure initiation; and (3) fracture propagation (in the absence of obvious signs of instability). Despite obvious deficiencies, they are useful for assessing avalanche risk in backcountry operations as well as for operational forecasting of the regional avalanche danger, in particular in areas with persistent weak snowpack layers.

A good test method should predict stable and unstable conditions similarly well (sensitivity vs specificity). Combinations of test results (from the same or different methods) are useful, as exemplified by RB scores and release type. Comparisons *across* datasets require cautious interpretation. Nevertheless, with this approach, the ECT has generally higher unweighted average accuracy than other tests. On the other hand, comparisons *within* datasets suggest that the ECT, RB, PST (and the SMP) generally have a comparable accuracy but higher than the CT. This is likely because the areas of the weak layer tested by the ECT, RB and PST are large enough to represent fracture propagation, whereas the CT tests for fracture initiation in about $0.09\,m^2$ of the weak layer, and hence has low specificity. The threshold sum provides no direct information about fracture initiation or propagation and has a lower unweighted average accuracy than any of the tests that fracture weak layers. This is consistent with LaChapelle (1980) who stated that observations of snowpack mechanics were more directly related to avalanching than observations of stratigraphy.

Even with very experienced observers an error rate of at least about 5–10% has to be expected. Site selection and interpretation require experience. Stability tests are not foolproof, and decisions about travelling in avalanche terrain should not be based solely on stability test results. Obviously, test reliability increases when two adjacent tests are carried out. However, a second test on a different slope (or at a second site on the same slope which is more than the autocorrelation length from the first site) should be more useful than the same test repeated in the same snow pit.

While accuracy is relevant when selecting a test for various scales of forecasting or backcountry decisions, considerations such as the effective depth, required time and technical skill are also important.

ACKNOWLEDGEMENTS

We thank I. Moner and D. Gauthier for providing additional information on their datasets, and A. van Herwijnen for valuable suggestions on the manuscript. For financial and logistical support, B.J. is grateful to the Natural Sciences and Engineering Council of Canada, HeliCat Canada, the Canadian Avalanche Association, the Canada West Ski Areas Association, Mike Wiegele Helicopter Skiing, Backcountry Lodges of British Columbia, the Association of Canadian Mountain Guides, Parks Canada, the Canadian Ski Guide Association, and Teck Mining Company. J.S. acknowledges support by the European Commission under contract NEST-506 2005-PATH-COM-043386 (Triggering of instabilities in materials and geosystems (TRIGS)). Thorough reviewer comments helped to improve this paper.

REFERENCES

Bellaire, S., C. Pielmeier, M. Schneebeli and J. Schweizer. 2009. Stability algorithm for snow micro-penetrometer measurements. *J. Glaciol.,* **55**(193), 805–813.

Birkeland, K.W. and D. Chabot. 2006. Minimizing 'false-stable' stability test results: why digging more snow pits is a good idea. *In* Gleason, J.A., *ed. Proceedings of the International Snow Science Workshop, 1–6 October 2006, Telluride, Colorado*. Telluride, CO, International Snow Science Workshop, 498–504.

Birkeland, K.W. and R.F. Johnson. 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), 115–123.

Doswell, C., J. Davies and D.L. Keller. 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather Forecast.,* **5**(4), 576–585.

Föhn, P.M.B. 1987a. The 'Rutschblock' as a practical tool for slope stability evaluation. *IAHS Publ.* 162 (Symposium at Davos 1986 – *Avalanche Formation, Movement and Effects*), 223–228.

Föhn, P.M.B. 1987b. The stability index and various triggering mechanisms. *IAHS Publ.* 162 (Symposium at Davos 1986 – *Avalanche Formation, Movement and Effects*), 195–214.

Gauthier, D. and B. Jamieson. 2006. Towards a field test for fracture propagation propensity in weak snowpack layers. *J. Glaciol.,* **52**(176), 164–168.

Gauthier, D. and B. Jamieson. 2008a. Fracture propagation propensity in relation to snow slab avalanche release: validating the propagation saw test. *Geophys. Res. Lett.,* **35**(13), L13501. (10.1029/2008GL034245.)

Gauthier, D. and J.B. Jamieson. 2008b. Predictions of the propagation saw test: comparisons with other instability tests at skier tested slopes. *In* Campbell, C., S. Conger and P. Haegeli, *eds. Proceedings of the International Snow Science Workshop, 21–27 September 2008, Whistler, British Columbia*. Whistler, B.C., International Snow Science Workshop, 408–414.

Greene, E. *and 10 others.* 2009. *Snow, weather, and avalanches: observational guidelines for avalanche programs in the United States. Second edition*. Pagosa Springs, CO, American Avalanche Association.

Hendrikx, J., K. Birkeland and M. Clark. 2009. Assessing changes in the spatial variability of the snowpack fracture propagation propensity over time. *Cold Reg. Sci. Technol.,* **56**(2–3), 152–160.

Jamieson, J.B. 1995. Avalanche prediction for persistent snow slabs. (PhD thesis, University of Calgary.)

Jamieson, J.B. 1999. The compression test – after 25 years. *Avalanche Rev.,* **18**(1), 10–12.

Jamieson, J.B. and C.D. Johnston. 1998. Refinements to the stability index for skier-triggered dry-slab avalanches. *Ann. Glaciol.,* **26**, 296–302.

Jamieson, B. and C. Johnston. 1993. Experience with rutschblocks. *In* Armstrong, R., *ed. Proceedings of the International Snow Science Workshop, 4–8 October 1992, Breckenridge, Colorado*. Denver, CO, Avalanche Information Center, 150–159.

Jamieson, B. and C.D. Johnston. 2001. Evaluation of the shear frame test for weak snowpack layers. *Ann. Glaciol.,* **32**, 59–69.

Jamieson, B., A. Zeidler and C. Brown. 2007. Explanation and limitations of study plot stability indices for forecasting dry snow slab avalanches in surrounding terrain. *Cold Reg. Sci. Technol.,* **50**(1–3), 23–34.

Jamieson, B., P. Haegeli and J. Schweizer. 2009. Field observations for estimating the local avalanche danger in the Columbia Mountains of Canada. *Cold Reg. Sci. Technol.,* **58**(1–2), 84–91.

Johnson, R. and K. Birkeland. 1998. Effectively using and interpreting stability tests. *In Proceedings of the International Snow Science Workshop, 27 September–1 October, 1998, Sunriver, Oregon*. Olympia, WA, Washington State Department of Transportation, 562–565.

Kronholm, K. 2004. Spatial variability of snow mechanical properties with regard to avalanche formation. (PhD thesis, University of Zürich.)

LaChapelle, E.R. 1980. The fundamental processes in conventional avalanche forecasting. *J. Glaciol.,* **26**(94), 75–84.

Landry, C.C., J. Borkowski and R.L. Brown. 2001. Quantified loaded column stability test: mechanics, procedure, sample-size selection, and trials. *Cold Reg. Sci. Technol.,* **33**(2–3), 103–121.

McCammon, I. and D. Sharaf. 2005. Integrating strength, energy and structure into stability decisions: so you dig a pit and then what? *Avalanche Rev.*, **23**(3), 18–19.

McClung, D.M. 2002. The elements of applied avalanche forecasting – Part 1: The human issues. *Natur. Hazards,* **26**(2), 111–129.

McClung, D. and P. Schaerer. 2006. *The avalanche handbook. Third edition*. Seattle, WA, The Mountaineers.

Moner, I., J. Gavaldà, M. Bacardit, C. Garcia and G. Marti. 2008. Application of the field stability evaluation methods to the snow conditions of the Eastern Pyrenees. *In* Campbell, C., S. Conger and P. Haegeli, *eds. Proceedings of the International Snow Science Workshop, 21–27 September 2008, Whistler, British Columbia*. Whistler, BC, International Snow Science Workshop, 386–392.

Roch, A. 1966. Les variations de la résistance de la neige. *IASH Publ.* 69 (Symposium at Davos 1965 – *Scientific Aspects of Snow and Ice Avalanches*), 86–99.

Ross, C. and B. Jamieson. 2008. Comparing fracture propagation tests and relating test results to snowpack characteristics. *In* Campbell, C., S. Conger and P. Haegeli, *eds. Proceedings of the International Snow Science Workshop, 21–27 September 2008, Whistler, British Columbia*. Whistler, BC, International Snow Science Workshop, 376–385.

Schneebeli, M. and J.B. Johnson. 1998. A constant-speed penetrometer for high-resolution snow stratigraphy. *Ann. Glaciol.,* **26**, 107–111.

Schweizer, J. 2008. On the predictability of snow avalanches. *In* Campbell, C., S. Conger and P. Haegeli, *eds. Proceedings of the International Snow Science Workshop, 21–27 September 2008, Whistler, British Columbia*. Whistler, BC, International Snow Science Workshop, 688–692.

Schweizer, J. and C. Camponovo. 2001. The skier's zone of influence in triggering slab avalanches. *Ann. Glaciol.,* **32**, 314–320.

Schweizer, J. and J.B. Jamieson. 2007. A threshold sum approach to stability evaluation of manual snow profiles. *Cold Reg. Sci. Technol.,* **47**(1–2), 50–59.

Schweizer, J. and M. Lütschg. 2001. Characteristics of human-triggered avalanches. *Cold Reg. Sci. Technol.,* **33**(2–3), 147–162.

Schweizer, J. and T. Wiesinger. 2001. Snow profile interpretation for stability evaluation. *Cold Reg. Sci. Technol.,* **33**(2–3), 179–188.

Schweizer, J., M. Schneebeli, C. Fierz and P.M.B. Föhn. 1995. Snow mechanics and avalanche formation: field experiments on the dynamic response of the snow cover. *Surv. Geophys.,* **16**(5–6), 621–633.

Schweizerr, J., J.B. Jamieson and M. Schneebeli. 2003. Snow avalanche formation. *Rev. Geophys.*, **41**(4), 1016. (10.1029/2002RG000123.)

Schweizer, J., K. Kronholm, J.B. Jamieson and K.W. Birkeland. 2008a. Review of spatial variability of snowpack properties and its importance for avalanche formation. *Cold Reg. Sci. Technol.,* **51**(2–3), 253–272.

Schweizer, J., I. McCammon and J.B. Jamieson. 2008b. Snowpack observations and fracture concepts for skier-triggering of dry-snow slab avalanches. *Cold Reg. Sci. Technol.,* **51**(2–3), 112–121.

Sigrist, C. and J. Schweizer. 2007. Critical energy release rates of weak snowpack layers determined in field experiments. *Geophys. Res. Lett.,* **34**(3), L03502. (10.1029/2006GL028576.)

Simenhois, R. and K. Birkeland. 2006. The extended column test: a field test for fracture initiation and propagation. *In* Gleason, J.A., *ed. Proceedings of the International Snow Science Workshop, 1–6 October 2006, Telluride, Colorado*. Telluride, CO, International Snow Science Workshop, 79–85.

Simenhois, R. and K.W. Birkeland. 2009. The extended column test: test effectiveness, spatial variability, and comparison with the propagation saw test. *Cold Reg. Sci. Technol.,* **59**(2–3), 210–216.

Spiegel, M.R. and L.J. Stephens. 1999. *Schaum's outline of theory and problems of statistics. Second edition*. New York, McGraw-Hill.

Stewart, K. and J.B. Jamieson. 2002. Spatial variability of slab stability in avalanche start zones. *In* Stevens, J.R., *ed. Proceedings of the International Snow Science Workshop, 29 September–4 October 2002, Penticton, British Columbia*. Victoria, BC, British Columbia Ministry of Transportation. Snow Avalanche Programs, 544–548.

Tremper, B. 2008. *Staying alive in avalanche terrain. Second edition*. Seattle, WA, The Mountaineers Books.

Van Herwijnen, A. and B. Jamieson. 2005. High-speed photography of fractures in weak snowpack layers. *Cold Reg. Sci. Technol.,* **43**(1–2), 71–82.

Van Herwijnen, A. and B. Jamieson. 2007a. Fracture character in compression tests. *Cold Reg. Sci. Technol.,* **47**(1–2), 60–68.

Van Herwijnen, A. and B. Jamieson. 2007b. Snowpack properties associated with fracture initiation and propagation resulting in skier-trigggered dry snow slab avalanches. *Cold Reg. Sci. Technol.*, **50**(1–3), 13–22.

Williams, K. 1980. Credibility of avalanche warnings. *J. Glaciol.,* **26**(94), 93–103.

Winkler, K. and J. Schweizer. 2009. Comparison of snow stability tests: extended column test, rutschblock test and compression test. *Cold Reg. Sci. Technol.,* **59**(2–3), 217–226.