# MINIMIZING "FALSE-STABLE" STABILITY TEST RESULTS: WHY DIGGING MORE SNOWPITS IS A GOOD IDEA

Karl W. Birkeland[1,*] and Doug Chabot[2]

[1]Forest Service National Avalanche Center, Bozeman, Montana, USA
[2]Gallatin National Forest Avalanche Center, Bozeman, Montana, USA

ABSTRACT: The worst nightmare for an avalanche worker is to assess an unstable slope as stable since the consequence of such an assessment is that you, your clients or the public could be caught in an avalanche. Thus, a primary goal in avalanche forecasting is to minimize such "false-stable" errors. In this paper we analyze the first season of data from the SnowPilot database. Starting with nearly 1,000 snowpits and 3,500 stability tests, we use stability test scores, shear quality, and weak layer depth to identify what we term the "critical weak layer" in each pit. We also divide the pits into "stable" and "unstable" categories based on the assessed snow stability and observations of obvious signs of instability (collapsing, cracking and recent avalanche activity). This filtering leaves us with 289 compression, rutschblock and stuffblock stability tests that fractured on the critical weak layer on unstable slopes. Of those 289 tests, 38 of them (13%) presented "false-stable" results, which we define as CT21 or greater, RB5 or greater, or SB drop heights 40 cm or greater. If we include shear quality and consider strong test results with a Q1 shear to be unstable, we decrease our false-stable cases to around 9% of the total. This implies that – if we use only stability test results – around 1 in 10 times we assess unstable slopes we will conclude that it is stable, which is unacceptably high. Recently spatial variability research has led some to argue that digging snowpits is unnecessary or futile, but we believe our data reinforce the idea that the key to analyzing snow stability lies in digging more rather than fewer pits, and using a holistic approach that considers much more than simple stability test results. Though our dataset is limited, it suggests that digging multiple pits might be an effective strategy for minimizing false-stable situations. In fact, having stability tests and associated shear quality from two different, but representative locations on the slope might decrease the chance of a false-stable error from around 10% to closer to 1%.

KEYWORDS: avalanche forecasting, stability assessment, stability tests, false-stable conditions

## 1. INTRODUCTION

Avalanche workers strive to accurately assess snow stability. During that assessment, two possible errors can be made: 1) concluding the snowpack is stable when it is actually unstable (statistically speaking, a Type I error), and 2) concluding the snowpack is unstable when it is actually stable (a Type II error) [*McClung*, 2002]. The consequence of the latter error is not opening a run, a road, or skiing some terrain. Frequently making such errors results in lost credibility. On the other hand, the consequences of the first error are that you, your clients, or the public are caught in an avalanche and possibly injured or killed. The severe consequences of assessing an unstable slope as stable makes minimizing these errors the primary goal of avalanche professionals. This paper focuses on cases when stability tests from snowpits give potentially misleading information (Figures 1 and 2). In particular, we are interested in cases in which stability tests indicate stable conditions on known unstable slopes. Like other past work [e.g., *Jamieson and Johnston,* 1995; *Johnson and Birkeland*, 2002], we call these results "false-stables".

To assess false-stables, we use the SnowPilot database [*Chabot, et al.*, 2006], which currently contains over 3500 stability test results. We use these data to assess the snowpack conditions associated with the false-stable results, and to analyze the frequency of false-stable results. In essence, our results show that false-stable results are relatively rare, and that their frequency can be further reduced by including shear quality or fracture character [*Birkeland and Johnson*, 1999; *Johnson and Birkeland*, 2002; *van Herwijnen and Jamieson*, 2006]. However, even

**Corresponding author address: Karl Birkeland, Forest Service Nat'l Avalanche Center, P.O. Box 130, Bozeman, MT 59771 USA; tel: 406-587-6954; email: kbirkeland@fs.fed.us*


*Figure 1: This picture shows an overview of the Henderson Bench area near Cooke City, Montana. The "X" indicates a snowpit location that gave false-stable results (SB40 Q1).*


*Figure 2: This avalanche, located on Crown Butte near Cooke City, Montana, was triggered by a snowmobiler. A subsequent snowpit (at the "X") demonstrated false-stable stability test results (SB50 Q2).*

when additional information is included, false-stable conditions still constitute an unacceptably high percentage of stability test results. Our limited data suggest that digging additional snowpits might be a good strategy for minimizing the chances of assessing unstable slopes as stable.

## 2. METHODS

We use the SnowPilot database to evaluate false-stables. SnowPilot is a free software program that allows users to enter, graph, and database their snowpits at www.snowpilot.org [*Chabot, et al.*, 2006]. In a little more than a year, users entered over 1100 snowpits with more than 3500 stability tests into the database. The disadvantage of such a database is the difficulty with quality control as opposed with studies where an individual or a research group collects carefully controlled data. However, this disadvantage is balanced with the advantage of being able to collect a great deal of data at a low cost. A further strength of the dataset is its diversity, with snowpits from many different areas and snow climates.

Analyzing the data required separating snowpits dug on unstable slopes from those dug on stable slopes. Users typically recorded the snow stability on the slope they were testing using standard definitions [*CAA*, 2002; *Greene, et al.*, 2004]. If they entered the stability as *Poor* or *Very Poor* the pit was classified as unstable, while ratings of *Good* or *Very Good* put the pit into the stable category. If snow stability was not rated or for ratings of *Fair*, evidence of collapsing, cracking or recent avalanches on similar slopes pushed a pit into the unstable category while the rest of the pits were stable.

Although the dataset has over 3500 stability tests, many of those tests are on layers that are not important for the current snow stability. An example would be a case where a stability test showed a break in the new snow, but the real stability problem was a surface hoar layer buried 50 cm down. Thus, our analysis required another step to identify what we termed the "critical interface" in each pit. Only one such "critical interface" could be identified for each pit. *Birkeland* [2001] and *Schweizer and Jamieson* [2003] previously defined the "critical interface" to be the location of the lowest rutschblock or compression test score. For this work we developed an algorithm to identify these critical interfaces using depth to the layer, stability tests, test scores, and shear quality. Our method first looked at interfaces deeper than 15 cm and then looked at the shallower layers only if no stability tests existed below 15 cm. The algorithm assessed interfaces as critical in the following order of priority:

1. Layers with the lowest rutschblock scores having Q1 or Q2 shears, then those having a Q3 shear.
2. Layers with the lowest stuffblock scores having Q1 or Q2 shears, then those having a Q3 shear.
3. Layers with the lowest compression test scores having Q1 or Q2 shears, then those having a Q3 shear.

A visual check of our results showed good agreement between the algorithm and our professional judgment of particular snow profiles.

Once we classified pits as unstable and assigned a critical interface to each one, we could identify false-stable test results. Following *Johnson and Birkeland* [2002] we called rutschblock results of 5 or greater and stuffblock drop heights of 40 cm or greater "stable" results. Others might argue that these numbers should be higher, but we kept them at these levels to be consistent with past work and to ensure a reasonable sized dataset. For the compression test we used hard compression tests (CT21 or greater) [*CAA*, 2002; *Greene, et al.*, 2004]. This is slightly less than the 23 or more taps that *Jamieson* [1999] refers to as "hard" compression test results. In addition to looking only at the stability test score, we also looked at cases where the stability test score was above the threshold levels we set *and* the shear quality was Q1. Finally, we looked at each false-stable profile manually to assure our algorithms had identified the appropriate pits.

For our overall dataset, and for each type of stability test, we calculated a *false-stable ratio*, which is simply the ratio of false-stable results to the total number of tests done on unstable slopes in our dataset. Thus, this *false-stable ratio* roughly quantifies the chances of getting a stable test result on an unstable slope.

## 3. RESULTS AND DISCUSSION

Out of over 3500 tests, SnowPilot users collected 300 tests on unstable slopes (as defined above). We focused only on rutschblock, stuffblock, or compression test results, leaving 289 total tests on unstable slopes. Of those, 38 tests from 29 pits were false-stables. In cases where the weak layer grain type was known, 88% of the cases had a persistent grain type (Table 1).

*Table 1: Characteristics of the critical interface and weak layer (wkl) in the 29 pits exhibiting false-stable stability test results.*

| | Median | Min. | Max. | Low quartile | Upper quartile |
|---|---|---|---|---|---|
| Depth to critical interface (m) | 0.53 | 0.17 | 1.45 | 0.46 | 0.85 |
| Elevation (m) | 2682 | 610 | 3079 | 2484 | 2818 |
| Slope angle (deg) | 30 | 23 | 45 | 27 | 34 |
| Lemon count* at critical interface | 4 | 0 | 5 | 3 | 4 |
| Wkl grain type | 10 mixed facets, 5 facets, 4 surface hoar, 1 cupped crystals, 1 small facets, 2 rounded grains, 1 mixed rounds, 5 unknown | | | | |

*Lemon counts are from parameters in *McCammon and Schweizer* [2002]. Only 13 of our 29 pits had sufficient data for complete lemon counts.

Following the snowpack parameters identified by *McCammon and Schweizer* [2002] resulted in a median lemon count of four for the false-stable cases (Table 1).

The grouping of stability test results in each pit is interesting for assessing the chances of getting multiple false-positives in one pit. Of the 29 pits with false-stable results, 15 (52%) had a single false-stable result that was near to other weaker stability test results in the same pit, while in two pits (7%) there were two false-stables adjacent to a weaker test result. Four pits (14%) exhibited either two or three false-stable results on the same critical layer, while eight (28%) had a single false-stable result at the critical interface. Looking only at the 21 pits where multiple tests were conducted at the critical layer shows that in 17 cases (81%) the false-stable was accompanied by a weaker test result. However, 19% of the time there were multiple false-positives and no weaker tests on a critical layer. This grouping of false-positives might be related to the spatial autocorrelation of stability test results in some cases (i.e., that closely spaced stability tests are more likely to have similar results than those spaced farther apart). Such autocorrelation at short distances has been suggested in previous work using shear frames [*Logan*, 2005; *Logan, et al.*, In press].

The primary purpose of this work was to examine the false-stable ratio. Of our 289 total tests on unstable slopes, 38 (13%) were false-stables (Figure 3). Breaking our data down by stability test shows 6 of 59 rutschblocks (10%), 11 of 81 stuffblocks (14%) and 21 of 149 compression tests (14%) classified as false-stable. Thus, our overall false-stable rate is on the order of 10 to 15%. This compares favorably to data collected by others on slopes adjacent to recent avalanches; *Schweizer and Jamieson* [2003] found a false-stable ratio of about 9% for compression tests and *Jamieson and Johnston* [1995] had a false-stable ratio of 18% for 22 total rutschblocks with scores of 5 or greater.

*Johnson and Birkeland* [2002] proposed using shear quality to help reduce false-stables. To integrate shear quality into our data, we considered strong stability test scores with a Q1 shear to be unstable. Consistent with previous work, our results demonstrate the importance of incorporating shear quality (or fracture character) into stability assessments. This reduced our overall false-stable ratio to 9%, with the false-stable ratio for rutschblocks dropping to 9%, stuffblocks to 6% and compression tests to 11% (Figure 3).

Though our data are admittedly limited, a false-stable ratio around 10% seems reasonable based on the results of others and on our personal experience of conducting stability tests on unstable slopes. On the one hand, this is good news since about 90% of the time we get a correct

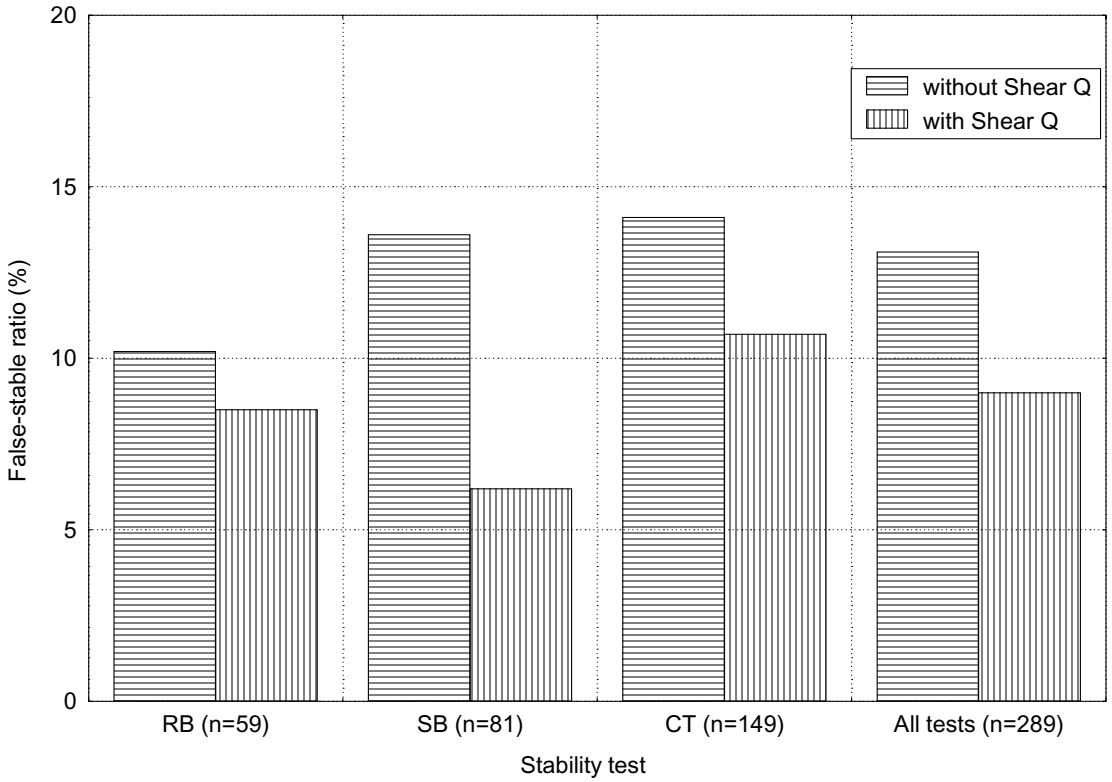


*Figure 3: False-stable ratios for the rutschblock (RB), stuffblock (SB), and compression (CT) tests and for all our tests together. Note the drop in false-stable ratios when shear quality is considered, emphasizing the importance of considering shear quality when doing stability tests. n is the total number of that particular test conducted on unstable slopes.*

result on unstable slopes. However, a false-stable ratio of 10% is still unacceptably high for professional operations and recreationists. Clearly, anyone traveling in avalanche terrain during times of instability that relies solely on single stability tests as their measure of slope stability would soon be caught in an avalanche. This is one reason why a holistic approach is needed for stability assessment, whereby stability tests are combined with knowledge of the snowpack history, weather data, observations of recent avalanche activity or other signs of instability, and other factors.

There are several reasons for false-stable results. First, the spatial variability on the slope might be such that the pit is in an unusually strong location. Second, there is a certain amount of error in all field tests, and some errors might lead to false stable results. Finally, if the targeted weak layer has collapsed at the pit location (due to disturbance or proximity to a nearby avalanche) then the interface will subsequently strengthen [*Birkeland, et al.*, 2006] and may provide false stable results.

Spatial variability research shows that stability test results may vary dramatically across slopes, and the underlying reasons for that variation is sometimes difficult to discern [e.g., *Campbell and Jamieson*, 2006; *Landry, et al.*, 2004; *Logan, et al.*, In press]. Though researchers have cautioned against this conclusion, slope-scale spatial variability has led some to suggest that digging snowpits is either unnecessary or futile. However, this approach is contrary to the goal of avalanche forecasting, which is to search for signs of instability from as many data sources as possible. Ignoring snowpits means that there is less information available in the search for instability. Experienced forecasters who commonly dig snowpits can typically point to a handful of cases when they believed the snowpack was stable, but where a stability test clued them in to a tricky instability. We believe that our results suggest that people should dig

more rather than less. With a false-stable ratio of 10%, there is a 1 in 10 chance that a stability test conducted on an unstable slope will demonstrate stable results. Looking at the problem from a purely probabilistic standpoint, this suggests that if we dig two separate snowpits on the same slope (to avoid any problems with possible spatial autocorrelation, or closely spaced tests being too similar) the chances of getting false-stable results from both would be 1 in 100; these chances would drop to 1 in 1000 by digging three well-placed pits.

Clearly, ours is not a probabilistic textbook problem. It is a practical one with dramatic and sometimes fatal results for incorrect decisions. If systematic human errors exist in the application of the stability test, additional tests may not dramatically reduce the chances of a false-stable result. Further, the spatial pattern on individual slopes is unknown. Some spatial patterns of unstable slopes may have broad areas of stronger snowpack and relatively small weak areas. Therefore, pushing the odds down as low as above might not be feasible, but we can reduce the odds *toward* those numbers. Thus, an effective strategy for minimizing the chances of false-stable results is digging additional snowpits and conducting additional stability tests.

## 4. CONCLUSIONS

Assessing snow stability is a difficult and sometimes dangerous task. Of particular concern are situations where a stability test result indicates a stable snowpack on a slope that is clearly unstable. This paper uses a large database to identify 38 such false-stable cases. Integrating shear quality into stability test scores clearly reduces the number of false-stable cases. However, the false-stable ratio (ratio of stable results on unstable slopes to all test results on unstable slopes) still hovers in the neighborhood of an unacceptably high 10%. This work is yet more evidence that more than simple stability test scores and shear quality are needed to adequately assess snow stability. Indeed, a holistic approach utilizing knowledge of the snowpack history, recent weather, avalanche activity, other signs of instability, and many other factors is required for good avalanche forecasts. In particular, we believe that additional well-placed snowpits and stability tests should help to further reduce the chances of obtaining false-stable results.


## Acknowledgements

Mark Kahrl did all the programming for SnowPilot, as well as maintaining and extracting the data for these analyses. Tom Ballard was instrumental in manipulating and cleaning up the database. Ethan Greene and Kalle Kronholm reviewed our paper and provided useful comments and discussions. Finally, we are indebted to all the users of SnowPilot. Thanks for collecting your data and trusting us to analyze it!


## References


Birkeland, K. W. (2001), Spatial patterns of snow stability throughout a small mountain range, *J. Glaciol.*, *47*, 176-186.

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*, 115-123.

Birkeland, K. W., K. Kronholm, S. Logan, and J. Schweizer (2006), Field measurements of sintering after fracture of snowpack weak layers, *Geophys. Res. Lett.*, *33*, doi:10.1029/2005GL025104.

CAA (2002), *Observation guidelines and recording standards for weather, snowpack and avalanches*, 78 pp., Canadian Avalanche Association (CAA), Revelstoke BC, Canada.

Campbell, C., and J. B. Jamieson (2006), Spatial variability of rutschblock resultes in avalanche start zones, paper presented at International Snow Science Workshop, Jackson Hole, Wyoming, 19-24 September 2004.

Chabot, D., M. Kahrl, K. W. Birkeland, and C. Anker (2006), SnowPilot: A "new school" tool for collecting, graphing, and databasing snowpit and avalanche occurrence data with a PDA, paper presented at International Snow Science Workshop, Jackson Hole, Wyoming, 19-24 September 2004.

Greene, E. M., K. W. Birkeland, K. Elder, G. Johnson, C. C. Landry, I. McCammon, M. Moore, D. Sharaf, C. Sterbenz, B. Tremper, and K. Williams (2004), *Snow, Weather and Avalanches: Observational guidelines for avalanche programs in the United States*, 136 pp., American Avalanche Association, Pagosa Springs, Colorado.

Jamieson, J. B. (1999), The compression test - after 25 years, *The Avalanche Review*, *18*, 10-12.

Jamieson, J. B., and C. D. Johnston (1995), Interpreting Rutschblocks in avalanche start zones, *Avalanche News*, 2-4.

Johnson, R. F., and K. W. Birkeland (2002), Integrating shear quality into stability test results, paper presented at International Snow Science Workshop, Penticton, B.C., 29 September-4 October 2002.

Landry, C. C., K. W. Birkeland, K. Hansen, J. J. Borkowski, R. L. Brown, and R. Aspinall (2004), Variations in snow strength and stability on uniform slopes, *Cold Reg. Sci. Technol.*, *39*, 205-218.

Logan, S. (2005), Temporal changes in the spatial patterns of weak layer shear strength and stability on uniform slopes, 169 pp, Montana State University, Bozeman.

Logan, S., K. W. Birkeland, K. Kronholm, and K. Hansen (In press), Temporal changes in the slope-scale spatial variability of shear strength of buried surface hoar layers, *Cold Reg. Sci. Technol.*

McCammon, I., and J. Schweizer (2002), A field method for identifying structural weaknesses in the snowpack, paper presented at Proceedings ISSW 2002. International Snow Science Workshop, Penticton BC, Canada, 29 September-4 October 2002.

McClung, D. M. (2002), The elements of applied forecasting - Part I: The human issues, *Natural Hazards*, *26*, 111-129.

Schweizer, J., and J. B. Jamieson (2003), Snow stability measurements, paper presented at Proceedings International Seminar on Snow and Avalanche Test Sites, Grenoble, France, 22-23 November 2001.

Schweizer, J., and J. B. Jamieson (2003), Snowpack properties for snow profile analysis, *Cold Reg. Sci. Technol.*, *37*, 233-241.

van Herwijnen, A., and J. B. Jamieson (2006), Fracture character in compression tests, paper presented at International Snow Science Workshop, Jackson Hole, Wyoming, 19-24 September 2004.