Bayesian inference of the climbing grade scale

Alexei Drummond and Alex Popinga

October 25, 2021

Abstract

Climbing grades are used to classify a climbing route based on its perceived difficulty, and have come to play a central role in the sport of rock climbing. Recently, the first statistically rigorous method for estimating climbing grades from whole-history ascent data was described [1], based on the dynamic Bradley-Terry model for games between players of time-varying ability. In this paper, we implement inference under the whole-history rating model using Markov chain Monte Carlo and apply the method to a curated data set made up of climbers who climb regularly. We use these data to get an estimate of the model's fundamental scale parameter m, which defines the proportional increase in difficulty associated with an increment of grade. We show that the data conform to assumptions that the climbing grade scale is a logarithmic scale of difficulty, like decibels or stellar magnitude.

We estimate that an increment in Ewbank, French and UIAA climbing grade systems corresponds to 2.1, 2.09 and 2.13 times increase in difficulty respectively, assuming a logistic model of probability of success as a function of grade. Whereas we find that the Vermin scale for bouldering (V-grade scale) corresponds to a 3.17 increase in difficulty per grade increment. In addition, we highlight potential connections between the logarithmic properties of climbing grade scales and the psychophysical laws of Weber and Fechner. Finally, we note a natural corollary of this model that describes how the grade of a "cruxy" route can be decomposed into the grades of its constituent sections, or moves, and illustrate how this approach can be used, along with the slope parameter m, to quantify the contribution of fatigue to overall difficulty.

Author summary

In this paper, we ask the question: "What does an increment of the climbing grade really mean?" Climbing grades originated as a way of classifying the difficulty of ascending particular routes based on subjective determination by the climbers. However, the grades exhibit strongly quantitative behaviour, which we analyse by forming a model that explains the grade of a route as a function of the number of sessions that a climber of a certain ability will take to successfully climb the route. We propose giving climbers a grade as well and then expressing the expected number of

failures before success as a logarithmic function of the difference between the climber's grade and the grade of the route. We find that by using this approach, we can estimate both the grade of the climber and the slope of the climbing grade scale itself from a logbook of attempts that includes all of the climber's successes and failures to ascend the route. Using data from a popular public climbing logbook, we find that an increment of the climbing grade corresponds to slightly more than a doubling of the expected number of failures before successfully climbing the route.

Introduction

Climbing grades play a central role in the sport of rock climbing [2, 3]. There are a number of commonly employed grading systems used by sport climbers globally to classify climbing routes into grades that increase with difficulty. In North America, the three-part Yosemite decimal system (YDS) is employed. In many parts of Europe, the French sport grading system is used. In Australia and NZ, climbers use the simple numerical Ewbank grading scale. The Ewbank scale is open-ended and begins at 1. Most amateur climbers will be able to climb routes up to about grade 18, while the most difficult routes currently graded under the Ewbank system are grade 35.

For lead climbing at the advanced end of the spectrum (i. e., from about Ewbank 23 onwards) there is an almost one-to-one correspondence between these three major grading systems [3], (see Table 1). However, this simple relationship is not universal, because a fourth major grade system (UIAA) has wider intervals of difficulty in this range [3].

For bouldering, the two common grading systems are the Vermin scale and the Fontainbleau scale. These two scales also have a correspondence in the higher grades, so that V11 $\equiv$ 8A, V12 $\equiv$ 8A+, ..., V17 $\equiv$ 9A.

Climbing grades were developed subjectively by different climbing communities around the world to classify climbing routes by difficulty. Yet, for difficult routes, these grading systems appear to have converged. They are also quite predictive of the chance of a climber of known ability to successfully climb a new route. This has led to the idea that there may be a quantitative law that underlies climbing grades [2], perhaps similar to the psychophysical Weber-Fechner law [4, 5]. The Weber-Fechner law states that the human perception of sensation intensity is proportional to the logarithm of the stimulus. In climbing, the grade represents the perceived difficulty (usually as determined by the first ascensionist). In controlled conditions it has been shown that climbing grades are proportional to the logarithm of some objective measures of difficulty [2].

Many physical and physiological determinants of climbing performance have been investigated [6, 7, 8]. If it exists, objective climbing difficulty must be a very complex quantity, but in certain situations it can perhaps be approximated by some relatively simple measurable proxy. One example of a measurable proxy exists in electromyographic stimulus of the *flexor digitorum profundus* (FDP) [2]. A climbing route that is primarily difficult because of the "crimpy" nature of its holds could be well characterised by such a proxy, since the FDP is the primary muscle involved in

Ewbank	French sport	YDS
23	7a	5.11d
24	7a+	5.12a
25	7b	5.12b
26	7b+	5.12c
27	7c	5.12d
28	7c+	5.13a
29	8a	5.13b
30	8a+	5.13c
31	8b	5.13d
32	8b+	5.14a
33	8c	5.14b
34	8c+	5.14c
35	9a	5.14d
36	9a+	5.15a
37	9b	5.15b
38	9b+	5.15c
39	9c	5.15d

Table 1: The correspondence between three of the major sport climbing grade systems in the range of grades relevant to advanced and elite climbers. For a more complete table see [3].

closing the fingers into the "crimp" grip position used by climbers to hold small ledges. Another example of a measureable proxy for assessing the complex difficulty of a climb was investigated in a recent study that found that shoulder strength was the most significant factor when considering performance on indoor lead-climbing routes [8], which are often set to be more "athletic" and less "fingery" (mostly relating to the crimps) than outdoor routes.

Recently, a new statistical approach to objectively estimate climbing route difficulty using whole-history ascent data of the climbing community was described [1]. This approach adapts an earlier method for Bayesian inference of player ratings in two-player games [9], by recasting the sport of climbing as a game between the climber and the route. Such models have a long history dating back to at least Zermelo [10], whose model was rediscovered decades later by Bradley and Terry [11]. These models have been developed, extended, and generalised to estimate the time-varying skill of players of board games like Chess [12, 13] and Go [9], computer games like Starcraft [14], and sports such as tennis [14] and basketball [14].

By applying this two-player game framework to sport climbing, climbers can be graded in the same way that routes are graded. An attempt to climb a route is a game between the climber and the route. If the climber and the route have the same grade, then the game will be fair, and there will be even odds that the outcome of the game is success on the part of the climber. This model introduces a scale parameter m to describe how

the probability of success changes as a function of the difference between the grade of the climber and that of the route.

Below, we investigate the suitability of climber ascent data to be described by the Bradley-Terry model. We then apply Bayesian MCMC inference to estimate the time-varying skill of a set of sport climbers along with the fundamental parameter m, which, by establishing a probability of success on the part of the climber, can also be used to describe the increase in difficulty associated with an increment in the climbing grade. We compare our estimates with earlier work. We also examine the assumptions of the whole-history method, leading us to suggest some improvements that take into consideration common self-reporting behaviours that exist in popular public log books of climbing ascents and can significantly affect the accuracy of the model.

Model

Ascent Data

Let $\mathcal{D} = \{(c_i, r_i, t_i, y_i) : i \in [N]\}$ be a dataset of N outcomes of sport climbing ascents. Each ascent i, occurs on date t_i and involves climber $c_i \in \{1, \ldots, N_c\}$ attempting an ascent of route $r_i \in \{1, \ldots, N_r\}$ with outcome $y_i \in \{0, 1\}$. An outcome of $y = 1$ represents a successful 'clean' ascent of the climb, while $y = 0$ represents a failure to make a clean ascent (including "hang dog", failed attempt, or retreat).

The Bradley-Terry model

Following previous work [1] we apply the dynamic Bradley-Terry model [10, 11] to ascent data. We define the grade, $C(t)$, of a climber at time t to be the grade for which they would have an even chance of climbing a route cleanly of that grade on their first (flash) attempt. A 7a climber, thus defined, would have a 50% chance of successfully "flashing" a 7a route.

Following previous work [1] we will posit the following equation for the probability of a successful attempt:

$$\Pr(y = 1) = p_{send} = \frac{e^{mC(t)}}{e^{mC(t)} + e^{mR}} \tag{1}$$

where $C(t)$ is the grade of the climber and R is the grade of the route, and m is a slope parameter that defines the proportional increase in difficulty associated with an increment of the grade of the route. This equation can be re-arranged to reveal that it represents a logistic function of p versus the grade difference $C(t) - R$:

$$p_{send} = \frac{1}{e^{m(R-C(t))} + 1} = \text{logit}^{-1}(mC(t) - mR), \tag{2}$$

where logit^{-1} is the cumulative distribution of the logistic distribution,

also known as the inverse logistic function:

$$\text{logit}^{-1}(x) = \frac{1}{1+e^{-x}} \tag{3}$$

Therefore estimating a climber's grade based on the outcomes of their ascent attempts is equivalent to a logistic regression with one independent variable. The success or failure of the attempt is the outcome, and the difference in the grade of the climber and the route is the independent variable. Assuming a logistic function is the same as stating that the log-odds of success is a linear function of the difference in the grades.

Assuming the route grade is known, a climber's grade, $C(t)$, can be seen as a scaled version of the intercept of the logistic regression. So defined, it does not have to be a whole number. The Ewbank grading scale is attractive for considering fractional climbing grades as it is purely numerical. So we could consider a climber to have grade 29.4 on the Ewbank scale, meaning that they have a better than even chance of flashing a grade 29 route and a worse than even chance of flashing a grade 30 route.

The grade of a climber is not static. The ability of a climber tends to increase rapidly in the first few months and years in the sport. Besides that, form waxes and wanes with injury, seasons, training intensity, et cetera. So any mechanism for estimating a climber's grade over a period of time must take account of the fact that the climber's "effective" grade will, in general, increase or decrease over time.

Following previous authors [9, 1] we consider a climber's grade to change through time according to a Wiener process, so that the prior distribution on $C(t+1)$ is:

$$C(t+1) \sim \mathcal{N}(C(t), w^2) \tag{4}$$

If we know the grade of a climber, we can predict their expected performance on a route of some other grade, including grades and routes they haven't tried.

If a climber has a probability p_{send} of success on any particular attempt of a route, then it follows that the expected number of failed attempts before the redpoint (first "clean" success) will equal to the odds-ratio of failure:

$$E[a] = \frac{1-p_{send}}{p_{send}} \tag{5}$$

So if the climber is climbing a route for which $p_{send} = 0.5$ (i. e., a fair game) then they would expect to climb that route in an average of 2 attempts. They would flash it with a 50% probability, take two goes with a 25% probability (i. e., 1 fail, 1 success), 3 goes with a 12.5% probability, et cetera.

Realising this relationship between p_{send} and $E[a]$ means that we can convert between one and the other. So if a climber fails 9 times for every send when climbing routes of a given grade, then we can compute the probability $p_{send} = \frac{1}{E(a)+1} = 0.1$.

Session grades

In the above model, we have considered a game to be defined by a single attempt by a climber to climb a route cleanly, (otherwise known as a "send" or "to send the route", wherein a climber moves from the ground to the top using only the rock or plastic holds assigned to the route, in the case of indoor climbing; the gear and rope are only there for safety and do not assist in the climbing on a "send"). An attempt is also known within the climbing community as a "tie-in" or a "go". A tie-in is the act of a climber tying the rope to their harness, then attempting to send the route, and either succeeding or returning to the ground and untying in order to rest before another go. During a tie-in, a climber may choose to continue to ascend the rock even after a failure to send has occurred (by falling or intentionally resting on the rope or gear); they may wish to practice moves or positions on the rock, place more gear for the next attempt, etc.

A single climber may have multiple tie-ins for a given route in a single day or "session" (a period of time, usually lasting from a couple of hours up to a whole day, during which a climber is trying to succeed in sending a route). Unfortunately, few climbers log the result of every tie-in during a session, and in these cases it is more realistic to define the climbing game by the outcome of each session's attempts.

If the game is defined on a session-basis, then the outcome of the game is the best result that the climber achieved during the session. The climber wins if they eventually achieved a clean ascent during the session, no matter how many attempts they made to achieve the clean ascent. Although this form of ascent data is more granular (at most one ascent outcome per route per day, assuming a day is equivalent to a session), it may be preferred because a larger fraction of the climbing community logs data that conforms to this definition.

We would anticipate that a climber's session grade, thus defined, is greater than their flash grade (in climbing lingo, their redpoint grade is higher than their flash grade).

Methods

We analysed 20 datasets across three countries (New Zealand, Australia, Germany) with the largest sets of data available on The Crag (`http://thecrag.com`), a popular online climbing logbook. Our datasets feature four different grading systems (Ewbanks, French sport, UIAA, V-scale) and three styles of climbing (two types of route climbing: sport and trad; as well as bouldering). Table 2 summarises the datasets.

We describe two analyses in more detail. The first highlighted analysis uses self-reported data on whole-history of ascents for 100 Australian climbers of differing abilities to estimate the fundamental slope parameter for the Ewbank climbing grade scale for sport climbing. We repeated this analysis using the same criteria for selection (see Supplementary Material) for climbers in New Zealand, resulting in an additional dataset of 89 climbers.

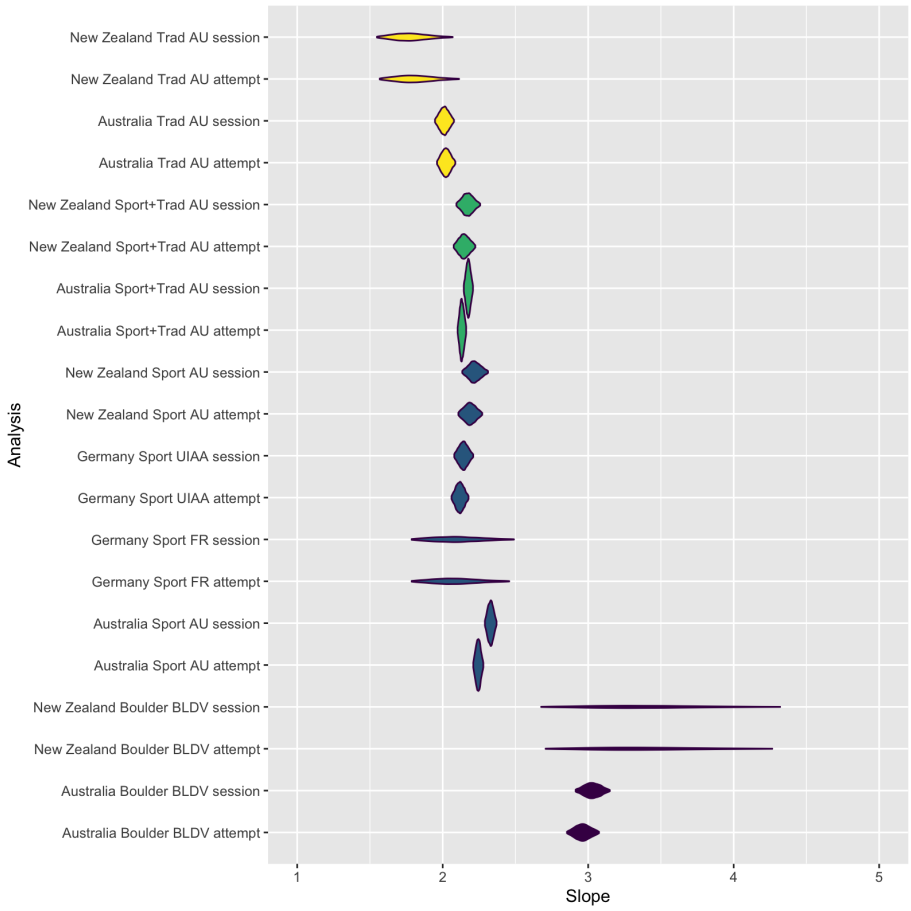

Figure 1: The posterior estimate of the slope parameter ($d = e^m$) for each of the 20 analyses.

country	gear (style)	climbers	ascents	slope	hpd.lower	hpd.upper	min.ascents	min.failures	grade.type	game	time
Australia	Boulder	100	25805	2.96	2.85	3.07	30	1	V-grade	attempt	6468
Australia	Boulder	100	25344	3.03	2.91	3.15	30	1	V-grade	session	5752
Australia	Sport	100	52947	2.24	2.21	2.28	30	1	Ewbanks	attempt	4142
Australia	Sport	100	48679	2.33	2.29	2.37	30	1	Ewbanks	session	4094
Australia	Sport+Trad	100	61931	2.13	2.10	2.16	30	1	Ewbanks	attempt	4905
Australia	Sport+Trad	100	57512	2.18	2.15	2.21	30	1	Ewbanks	session	4729
Australia	Trad	100	14929	2.02	1.96	2.09	30	1	Ewbanks	attempt	4899
Australia	Trad	100	14479	2.01	1.95	2.08	30	1	Ewbanks	session	4324
New Zealand	Boulder	15	1049	3.35	2.71	4.27	30	1	V-grade	attempt	41
New Zealand	Boulder	15	1027	3.35	2.68	4.32	30	1	V-grade	session	40
New Zealand	Sport	89	10027	2.19	2.11	2.27	30	1	Ewbanks	attempt	1561
New Zealand	Sport	89	9679	2.22	2.13	2.31	30	1	Ewbanks	session	2397
New Zealand	Sport+Trad	98	11389	2.15	2.08	2.22	30	1	Ewbanks	attempt	2597
New Zealand	Sport+Trad	98	11024	2.17	2.09	2.26	30	1	Ewbanks	session	2344
New Zealand	Trad	8	524	1.80	1.57	2.11	30	1	Ewbanks	attempt	47
New Zealand	Trad	8	517	1.77	1.55	2.07	30	1	Ewbanks	session	30
Germany	Sport	8	521	2.08	1.79	2.46	30	1	French	attempt	42
Germany	Sport	8	483	2.09	1.79	2.49	30	1	French	session	25
Germany	Sport	100	19497	2.12	2.06	2.18	30	1	UIAA	attempt	5026
Germany	Sport	100	18435	2.14	2.08	2.21	30	1	UIAA	session	4949

Table 2: Summary of Bayesian analyses performed

The estimated slopes of all 20 analyses are presented in Figure 1.

In the second set of analyses we use whole-community successful ascent data from New Zealand and Australia. Finally we analyse a small data set of curated link-up boulder problems to illustrate the limitations of grading climbs in parts without considering fatigue.

Estimating the slope of difficulty for the Ewbank climbing grade scale

For the Australia-Sport analysis we used self-reported whole-history of ascents for $n = 100$ climbers of differing abilities. All of these log books were self-reported on `http://thecrag.com`, a popular online climbing logbook. The total number of ascents analysed was $N = 48,679$. Most climbers do not self-report every attempt. To reduce the bias caused by climbers that don't log every failure we converted all logbooks to a session logging form, in which only the single most successful ascent was chosen for each climber on each route on each day that climbing occurred. In addition we only considered climbers that had at least 30 attempts, with at least 1 explicit failure (i. e., at least one hangdog, attempt, retreat or working). The analysis took 1hr and 8 minutes on an iMac 3.6 GHz 8-Core Intel Core i9.

We implemented full Bayesian MCMC inference of the dynamic Bradley-Terry model in Stan [15]. This code is available to the public domain at `https://github.com/alexeid/climbing-grades`. The Stan model employed is shown in **??** (Supplementary Information).

We used this model to co-estimate the fundamental slope parameter m and the grade of each climber in monthly windows during 60 months between 1st August 2016 and 1st August 2021. For the purpose of this analysis we assumed that the assigned grades for the routes attempted were correct.

Estimating the slope of total successful ascents versus grade in community-wide ascent data

We followed O'Neill [16] in analysing total successful ascent data, as a complementary approach to understanding grade difficulty. We downloaded the total number of successful ascents by grade for New Zealand and Australia from `http://thecrag.com`, for both bouldering and sport climbing on 20th July 2021. For sport climbing we chose to include ascents with tick types: redpoint, flash and onsight. For bouldering we chose to include tick types send, flash and onsight. The purpose was to chose tick types that strongly implied a clean send on the part of the ascensionist (i. e., represented accomplishing success of the full difficulty of the boulder problem or route).

Estimating the fatigue factor for link-up boulder problems

It it common to construct a more difficult boulder problem by linking two existing boulder problems together. Armed with an estimate of the m parameter we can consider whether boulder problems can be described by the sum of the difficulties of the subproblems. We would anticipate that deviations would be caused in the case where fatigue accumulates from the first part of the boulder problem and a complete reset between the subproblems is not possible. So by estimating the expected grade based on the grades of the subproblems, and comparing that to the assigned grade we can get insight into the extent to which unmodeled fatigue contributes to the perceived grade across different boulder problems.

In supplementary information we curated a small table of link-up boulder problems. For each problem there is a grade for the full problem, as well as individual grades for the parts of the problem. Assuming a slope parameter of $m = 0.6$ we compute the expected grade of the full problem from the grades of its parts and define the difference between this expected grade and the actual grade as the fatigue factor F.

Results

Estimating the slope of difficulty for the Ewbank climbing grade scale

Figure 2 shows the estimated grade through time plot between August 2016 and July 2021 inclusive for 100 climbers that fulfilled our selection criteria for the Australia Sport data set. The second panel reports the posterior distribution of the slope parameter which was jointly estimated. The estimate of the m parameter was 0.85 [0.83, 0.86], which is equivalent to 2.33 [2.29, 2.37] times more failed attempts per success per grade increment.

The mean estimate of the m parameter from a simple log-linear regression of each climber individually assuming no change in ability of the analysed period was 0.65 (see Figure **??**; Supplementary Information).

The median number of explicit fails per climber (i. e., hangdog, attempt, working, retreat), was 126 (range: 18-642) or as a fraction 29.9% (interquartile range 4.7%- 72.0%) of ascents.

Figure 3 shows the estimated grade through time plot between August 2016 and July 2021 inclusive for 89 climbers that fulfilled our selection criteria for the New Zealand Sport data set. The second panel reports the posterior distribution of the slope parameter which was jointly estimated.

The estimate of the m parameter was 0.8 [0.76, 0.84], which is equivalent to 2.22 [2.13, 2.31] times more failed attempts per success per grade increment.

The mean estimate of the m parameter from a simple log-linear regression of each climber individually assuming no change in ability of the analysed period was 0.52 (see Figure **??**; Supplementary Information).

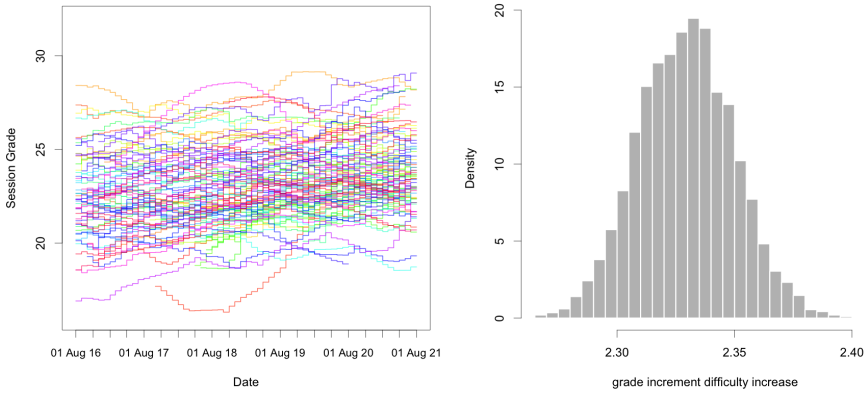

Figure 2: The posterior estimate of each Australian climber's grade ($n = 100$) through time and the posterior distribution of the proportional increase in difficulty per grade increment $d = e^m$.

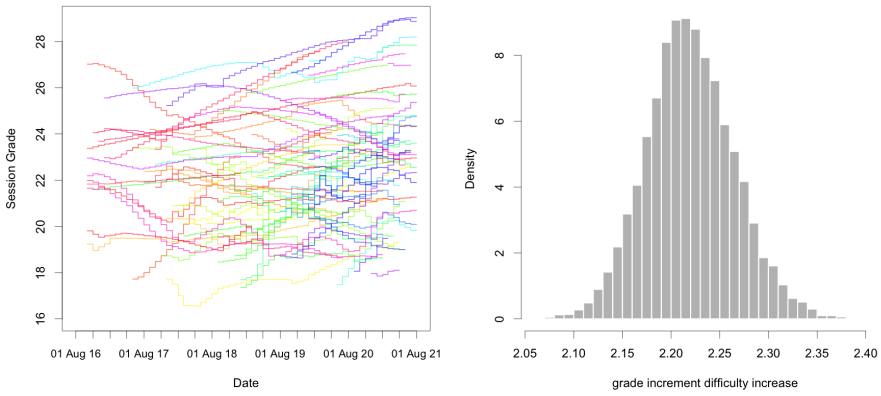

Figure 3: The posterior estimate of each New Zealand climber's grade through time and the posterior distribution of the proportional increase in difficulty per grade increment $d = e^m$.

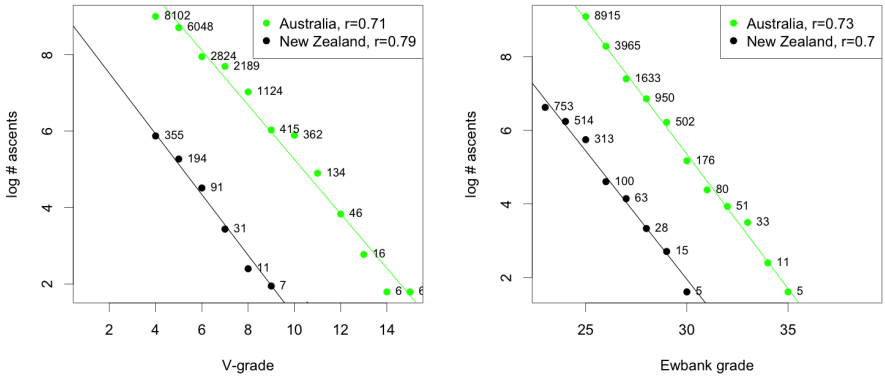

Figure 4: The relationship between grade and the log of the number of successful sends for two styles of climbing (bouldering and sport climbing) in two countries, based on whole climbing community statistics on `http://thecrag.com`

Estimating the slope of total successful ascents versus grade in community-wide ascent data

Figure 4 shows the whole-community counts of successful ascents by grade for Australian and New Zealand for both sport climbs and bouldering. The slope of these curves range between 0.7 and 0.79, conforming remarkably well with the more detailed Bayesian analysis above. This data is not affected by bias in under-reporting of failed attempts, but will be affected by systematic variation in the number of attempts made by the community at each grade.

Discussion and Conclusion

We have endeavoured to bring together various pieces of community knowledge about the mathematics of climbing grades and reframe them in the light of recent work on grade estimation using whole-history ascent data.

Our main contribution is to focus on quantifying the magnitude of the increase in difficulty representing by an increment in climbing grade. We provide evidence that for the Ewbank, French sport and UIAA grading systems this increment is slightly more than a doubling in difficulty for each grade, as measured by average number of failed sessions required before achieving success. For the V-grade system we find that the increment in difficulty is slightly more than a tripling in difficulty for each grade.

This is quite a surprising quantitative law to arise from what many have assumed to be a subjective process of grading climbs. Whereas climbers do argue whether a climb is too hard or too "soft" for the grade, implying its true grade is off by one, it is much more rare for disagreements about the grade of a route to span more than two consecutive

grades. Typically there is one grade representing the consensus, and a neighbouring grade the alternative. This makes sense if the grade scale is in fact a logarithmic scale of difficulty.

An open question is how these subjective grading systems settled in to such a rigid quantitative law. One hypothesis worth further investigation is that a logarithmic scale is optimal for magnitude estimation and therefore arises naturally in many settings [17]. In many different domains it appears that the perception of a stimulus varied logarithmically with its absolute magnitude. If so, why did these (sport climbing) grading systems settle on a slope so close to 2? Is there an information theoretic reason for this?

One of the key data preparation approaches used here that differs from previous whole-history efforts is to use per-session data. This is a form of aggregation in which only the most successful attempt for each climber on each route on each day is retained and all other attempted ascents are removed from the data set. This caters for a common form of self-reporting that reports the best effort during each day of climbing. We refer to the grade estimated for a climber in this way as the *session grade*. Since it is possible to try a route 2 or more times in a single session, a climber's *session grade* should be higher than the *flash grade*.

The grade of a climb defines its overall difficulty for the average climber in appropriate conditions. However there are many factors besides the grade that can affect the probability of success. Many routes are best climbed in certain conditions, including time of day, season and weather. Direct sun on a route is normally not conducive to optimal performance and many climbers prefer shade, cool rock, a slight breeze and low humidity for the best chances of success. Besides conditions, routes come in many styles; a climber may specialise in slabs (positive angle), vertical or overhanging routes. Depending on the rock type and crag, different hold types may predominate (pockets, slopers, crimps, underclings, jugs), and the climbing may be more "cruxy"/"bouldery" with rests, or of a "pumpy"/"resistance" nature more suited to climbers with endurance. Finally some routes, or their cruxes, are easier to climb for certain body types or sizes. A crux move may be easier for a tall climber because of their extra reach, or easier for a shorter climber because they can fit their body into a small space between holds (called "fitting in the box"). Routes for which the difficulty varies greatly depending on the size of the climber are sometimes termed "morpho". Suffice to say there are many factors that could be taken into account in order to elaborate the simple model described here. Arguably the most obvious would be to include the style of the route (e.g. slab, vertical, steep/overhung, or cruxy versus pumpy) and estimate a correction factor for each climber on each style.

The most questionable detail of the model presented is the idea that the probability of a successful ascent remains the same after each previous attempt of the route, i. e., practising a particle route does not improve the climber's chance of success. This is clearly a bad assumption, so at best the probability of success implied by the Bradley-Terry model should be considered as some sort of "effective" probability, averaged over different levels of practice. The real underlying probability is probably increasing with practice. The problems with developing a model that

admits learning are at least two-fold: (i) we expect some routes are more amenable to practice than others, and (ii) it is unclear what functional form the expected improvement in probability per attempt should take.

Besides details of the model, there are also potential problems related to how climbers choose the routes to climb (selection bias), and also which ascent attempts they decide to log (under-reporting failures).

Selection bias is the propensity of climbers to chose routes that are at the easier end of the grade, or at least routes that suit their climbing style (rather than choosing a random climb at the grade). There are two reasons why we think that selection bias shouldn't have a large impact on the estimate of the climbing grade slope parameter. Firstly, climbers are probably less selective at lower to middle grades in their range, and the slope is derived from the full range of grades that a climber attempts. Secondly, if a climber is selective at all grades, such a bias doesn't change the slope, but instead changes the intercept, because in each grade the result will be to select from the lower end of the grade, which will still result in the same slope. The main result of such selection bias will be an overestimate the grade of the climber by up to 1 grade. If the climbers only select a certain style of climb, then their grade estimate is only relevant to that style of climb. So selection bias will overestimate a climber's ability a bit (either overall, or for styles they don't try), but it won't have a big effect on the fundamental grade scale parameter estimate. The worst case scenario will be if all climbers only climb relatively hard routes in grades low in their range and relatively easy routes from grades high in their range. This would mean the true range in the x-axis of the regression for each climber is actually up to a grade less than what is used to compute the slope. So the estimated slope will be flatter than the true slope. This extreme scenario would flatten the slope by approximately $(k-1)/k$, where k is the difference between the highest grade and the lowest grade. In our data set the interquartile range for k is 8-10. So the worst case scenario if everybody is maximally biased in this way is an underestimate of the slope of about 10%.

Selective logging of failures is a more significant concern. It is clear that most climbers don't log all failures. It seems likely that failures on easy routes would be more embarrassing than failures on harder routes. So if this motivation is in action, then harder routes will appear even harder than they actually are, since the failures are underreported for easier routes. This would lead to an overestimation of the slope of the grade scale.

Suffice to say, there is still much work to do. Foremost in our mind is the need to adapt recently developed whole-history inference methods to account for biases in public repository data in a much more rigorous way than pursued here. This will require a better understanding of the differing ways that climbers approach self-reporting of climbing ascents. It seems likely that climber that logs their own ascent and attempts are likely to be susceptible to various biases and to follow differing conventions, depending on their purpose for making a public log book of their ascents. Data-driven approaches to learning about these differing conventions and biases in order to classifying climbers by their logging approach is an obvious next step.

Acknowledgement

The authors would like to thank Simon Dale and Ulf Fuchslueger from `thecrag.com`, for providing access to the theCrag API (`https://www.thecrag.com/en/article/api`) so that the public ascent data could be downloaded programmatically for the analyses produced in this work. In addition we thank Simon Dale, Dr Joseph Heled, Daniel Krippner, Dr Michael Matschiner, John Palmer and Dr Tim Vaughan for helpful discussions on earlier versions of this manuscript.

References

[1] D. Scarff, "Estimation of climbing route difficulty using whole-history rating," *arXiv preprint arXiv:2001.05388*, 2020.

[2] D. Delignières, J.-P. Famose, C. Thépaut-Mathieu, P. Fleurance *et al.*, "A psychophysical study of difficulty rating in rock climbing," *International Journal of Sport Psychology*, vol. 24, pp. 404–404, 1993.

[3] N. Draper, D. Giles, V. Schöffl, F. Konstantin Fuss, P. Watts, P. Wolf, J. Baláš, V. Espana-Romero, G. Blunt Gonzalez, S. Fryer *et al.*, "Comparative grading scales, statistical analyses, climber descriptors and ability grouping: International rock climbing research association position statement," *Sports Technology*, vol. 8, no. 3-4, pp. 88–94, 2015.

[4] *De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*, ser. De pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae, 1834. [Online]. Available: https://books.google.co.nz/books?id=j7CspFMOQTYC

[5] G. T. Fechner, *Elemente der psychophysik*. Leipzig: Breitkopf und Härtel, 1860.

[6] J. Baláš, O. Pecha, A. J. Martin, and D. Cochrane, "Hand–arm strength and endurance as predictors of climbing performance," *European Journal of Sport Science*, vol. 12, no. 1, pp. 16–25, 2012.

[7] J. Baláš, M. Panáčková, B. Strejcová, A. J. Martin, D. J. Cochrane, M. Kaláb, J. Kodejška, and N. Draper, "The relationship between climbing ability and physiological responses to rock climbing," *The Scientific World Journal*, vol. 2014, 2014.

[8] R. MacKenzie, L. Monaghan, R. A. Masson, A. K. Werner, T. S. Caprez, L. Johnston, and O. J. Kemi, "Physical and physiological determinants of rock climbing," *International journal of sports physiology and performance*, vol. 15, no. 2, pp. 168–179, 2020.

[9] R. Coulom, "Whole-history rating: A bayesian rating system for players of time-varying strength," in *International Conference on Computers and Games*. Springer, 2008, pp. 113–124.

[10] E. Zermelo, "Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 29, no. 1, pp. 436–460, 1929.

[11] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[12] A. E. Elo, *The rating of chessplayers, past and present.* New York: Arco Pub., 1978.

[13] M. E. Glickman and A. C. Jones, "Rating the chess rating system," *CHANCE-BERLIN THEN NEW YORK-*, vol. 12, pp. 21–28, 1999.

[14] L. Maystre, V. Kristof, and M. Grossglauser, "Pairwise comparisons with flexible time-dynamics," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1236–1246.

[15] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, no. 1, pp. 1–32, 2017.

[16] T. O'Neill. (2002, June) Grade theory? [Online]. Available: https://web.archive.org/web/20020609071230/http://www.australianbouldering.com:80/table.html

[17] R. Portugal and B. F. Svaiter, "Weber-fechner law and the optimality of the logarithmic scale," *Minds and Machines*, vol. 21, no. 1, pp. 73–81, 2011.