**Bojan Leskošek [1]**
**Rok Blagus [1,2]***

# THE RELIABILITY OF THE DIFFICULTY GRADING OF SPORT CLIMBING ROUTES

# ZANESLJIVOST OCENJEVANJA TEŽAVNOSTI ŠPORTNOPLEZALNIH SMERI

## ABSTRACT

This is the first study aiming to evaluate the reliability of assigning a difficulty grade to a climb. Each of the 70 climbers, divided by their abilities into group A (intermediate) and group B (advanced), climbed 5 different routes, and independently suggested a grade for the routes. Although the reliability was generally high (Kendall W = .89 and .86, intraclass correlation coefficient ICC = .82 and .89 in groups A and B, respectively), the grades proposed by a single climber have too large a standard error of measurement to confidently claim that two routes with different grades are truly of different difficulties. The most important factor associated with the accuracy and bias of grading, was the gap between a climber's ability and a route's difficulty (easier routes were graded less accurate, advanced climbers underestimated the true difficulty of the routes). The following factors were also found important for grading; their effect however, was lower, or was only found in some routes or in one group: success (failed attempts were graded higher), tiredness, sex and body height (but only unadjusted for sex). Surprisingly, no significant effects were found for climbing experience, ability level, and style of climbing (onsight, flash).

*Keywords:* rock climbing, performance, accuracy, validity

***[1]Faculty of Sports, University of Ljubljana, Slovenia***

***[2]Institute for Biostatistics and Medical Informatics, University of Ljubljana, Slovenia***

*Corresponding author*: assist. prof. Rok Blagus, Ph.D., Faculty of Sports, University of Ljubljana, Gortanova ulica 22, 1000 Ljubljana, tel. +386 1 520 77 00, fax +386 1 520 77 40*
E-mail: rok.blagus@fsp.uni-lj.si

## IZVLEČEK

Gre za prvo raziskavo zanesljivosti ocenjevanja v športnem plezanju. Vsak od 70 plezalcev, po sposobnostih razdeljenih v skupini A (srednja) in B (visoka), je poskusil preplezati 5 različnih smeri in neodvisno od ostalih predlagal oceno težavnosti smeri. Čeprav je bila zanesljivost ocenjevanja na splošno visoka (v skupinah A in B: Kendall W = 0,89 in 0,86, koeficient intraklasne korelacije ICC = 0,82 in 0,89 ), imajo ocene, ki jih predlaga en sam plezalec, preveliko standardno napako merjenja, da bi lahko za dve smeri podobnih težavnosti zanesljivo trdili, da se težavnost teh dve smeri v resnici razlikuje. Najpomembnejši dejavnik, povezan z natančnostjo in pristranskostjo ocenjevanja, je bila razlika med plezalno sposobnostjo plezalca in zahtevnostjo smeri (lahke smeri so bile nasploh ocenjene manj natančno, visoko sposobni plezalci pa so podcenjevali resnično težavnost lahkih smeri). Za ocenjevanje smeri so bili pomembni tudi naslednji dejavniki, ki pa imajo manjši vpliv ali pa se je vpliv kazal samo v nekaterih smereh ali skupinah plezalcev (A ali B): uspeh vzpona (neuspeli poskusi so bili ocenjeni višje), utrujenost, spol in telesna višina (slednja le neprilagojena za spol). Presenetljivo pa plezalne izkušnje, stopnja plezalne sposobnosti (A ali B) in slog plezanja (na pogled, flash) niso pomembno vplivali na zanesljivost ali pristranskost ocenjevanja.

*Ključne besede:* plezanje, zmogljivost, natančnost, veljavnost

## INTRODUCTION

Sport climbing is a type of rock climbing which was initially an outdoor sport performed on natural rock. It has recently gained high popularity largely due to the emergence of indoor climbing facilities making the sport more accessible to a wider public audience (Valenzuela, de la Villa, & Ferragut, 2015; Watts, 2004). Sport climbing is now a well-established competitive sport with athletes competing in three disciplines: speed, boulder and lead. The sport has been approved as part of Tokyo and Paris Olympics program. Outdoor sport climbing on natural rock is mostly limited to lead climbing and bouldering (Schöffl & Kuepper, 2006); in both, the goal is to reach the top of the route by free climbing, i.e. progressing using only the natural features in the rock without weighting the rope or pulling on carabiners clipped to anchors (Draper, Dickson, Fryer, & Blackwell, 2011). A successful attempt of the route counts only if the route was climbed in lead (the climber clips the belay rope into preplaced equipment attached to bolts) and done by following the free climbing ethics (the climber progresses using handholds and footholds without weighting the rope). The route can be climbed successfully in three valid styles: onsight, flash and redpoint (Draper, Dickson, Blackwell, et al., 2011; Draper, Dickson, Fryer, et al., 2011; Draper et al., 2016). The most valued (and difficult) style is onsight (OS), in which a climber has no prior knowledge of the route. Flash (F) ascents are made in the first try, in which any prior information is allowed. A redpoint (RP) ascent is completed by a climber that has previously climbed any part of the route at least once. After the first successful ascent of a route, the first ascensionist gives it a name and suggests a grade. The grade should objectively describe the difficulty of the route, which may depend on several factors, e.g. the technical difficulty, the power required to execute the single moves, or the stamina needed for long passages in a route without good rest points. However, grades of sport climbing routes are subjective (Morrison & Schöffl, 2007) and are based on the comparison of the difficulty of a particular route with other routes of similar style. After the first ascent, other climbers may agree with the grade suggested by the first ascensionist, or they suggest lower and/or higher grades. Although different individual grades are given by different climbers for the same route, usually only one (consensus) grade is published in a guidebook. The consensus grade is usually given by the author of a guidebook, preferably after consulting other climbers that have successfully climbed or at least tried the route. Different grading scales are used, with the French and the YDC (Yosemite Decimal System) scales probably the best known and most commonly applied (Draper et al., 2016). These grading scales are widely utilised to measure performance, to discriminate between ability groups in studies on rock climbing (de Moraes

Bertuzzi, Franchini, Kokubun, & Kiss, 2007; Grant et al., 2001; Saul, Steinmetz, Lehmann, & Schilling, 2019; Wall, Starek, Fleck, & Byrnes, 2004), and also as a basis for comparison, e.g. they may be used to describe the progress in a single climber's career or the overall progress of climbing over time (Draper, Jones, Fryer, Hodgson, & Blackwell, 2008; Morrison & Schöffl, 2007; Sherk, Sherk, Kim, Young, & Bemben, 2011; Valenzuela et al., 2015).

Especially important is the comparison function of a grade in outdoor climbing, where no (official) competitions exist and the grades serve as the basis for ranking the routes and the climbers. The subjective evaluation of athletes' performances is common in sports, e.g. in gymnastics, figure skating, wrestling and dance (Bučar, Čuk, Pajek, Karacsony, & Leskošek, 2012). Contrary to other sports, however, there are no written rules and no officially recognised judges that evaluate an athlete's performance when outdoor climbing. Additionally, in climbing, especially in the most difficult routes that are not repeated for several years or even decades, there is only one "judge" that evaluates his/her own performance.

Several questions arise from the fact that the individual grades for the same route are different, variable. First, how variable are they, and what are the factors that may influence this variability? Second, should a procedure be constructed that may "extract" a "valid" consensus grade from these individual grades; and if it does, how many individual grades are needed to achieve this–is every rater eligible to propose a grade, etc.? Third, how sensitive are the grading scales, i.e. how different should be the grades of two routes, so that we may be "certain" one is more difficult than the other? To answer these questions, we conducted a study enrolling a sample of 70 climbers. Each climber attempted to climb 5 indoor routes and suggested, completely independently from other climbers, his or her opinion about the difficulty grade of the route. We studied the reliability of the proposed difficulty grades and factors related to variability and bias of the grading of sport climbing routes.

## METHODS

### Study protocol

The study was conducted on a lead climbing wall in the Climbing Centre Ljubljana (CCL), Slovenia. On three consecutive working days in October 2015, visitors of the climbing centre were invited to participate in the study. If they agreed to participate and met the inclusion criteria (able to lead climb routes with difficulties of 6a and above according to the French scale), they signed an informed written consent about the conditions of their participation. After their usual warmup they climbed 5 new routes (assigned A1 to A5 for group A and B1 to B5 for group B) in an order they chose. All routes were positioned on the main 850 m2 indoor climbing wall inside the CCL (http://www.plezalnicenter.si/english-page/); they were 15.0 to 16.4 m high and set by two experienced route-setters, who regularly set the routes on this wall. After lowering down from the top of each route, participants immediately had to record, on a prepared form, their grade proposition, ascent style (OS, F) and whether they were able to successfully complete the route in a selected style. No communication regarding the difficulty of the route was allowed between the participants. Only one attempt per route was allowed. As is a common practice in CCL, the letter, "+" and slash grades were allowed also for easier routes, e.g. 4a, 5a+, 5b+/c. The study was conducted with the institutional ethical approval.

### Participants

The accidental sample consisted of 70 climbers. Each participant took part in a short survey, from which some basic demographic and physical characteristics and information about self-reported climbing level were obtained. According to their self-reported climbing level and their most difficult redpoint (RP) ascent in the last 12 months, they were included in either group A (intermediate group: RP level 5c–6c+), or group B (advanced group: RP level 7a upwards) (Draper et al., 2016); some exceptions were made by the authors after interviewing the participants: if a climber was currently not in good enough shape to reach the top (successfully or not) of all 5 routes, he/she may be transferred from group B to group A). The basic characteristics of the intermediate and the advanced groups are provided in Table 1. No significant differences between the groups were observed for age, sex, and self-reported body height and weight ($p > .05$), while the advanced climbers were significantly more experienced and had higher OS and RP levels.

Table 1. Basic characteristics of the study sample. The reported p-value is either from a two-sample equal variances t-test, Mann-Whitney test or Pearson's chi-squared test, where appropriate.

| | Group A | Group B | *p*-value |
|---|---|---|---|
| Total count | 41 | 29 | |
| Sex | 23 male, 18 female | 23 male, 6 female | .078 |
| Age (years) | 13–49, *M* = 29.9, *s* = 9.9 | 12–50, *M* = 31.4, *s* = 10.6 | .551 |
| Body height (cm) | 157–185, *M* = 172.2, *s* = 8.2 | 150–189, *M* = 173.7, *s* = 9.7 | .514 |
| Body weight (kg) | 47–85, *M* = 65.5, *s* = 10.9 | 40–94, *M* = 68.6, *s* = 10.6 | .241 |
| Climbing experience (years) | 1–25, *M* = 5.2, *s* = 5.2 | 2.5–30, *M* = 13.3, *s* = 8.6 | < .001† |
| Hardest climb OS (ever) | 5c–7a+/b, *Mdn* = 6b+ | 6c/c+–8a+, *Mdn* = 7a+/b | < .001† |
| Hardest climb RP (ever) | 6a+–8a+, *Mdn* = 6c | 7a+–8c/c+, *Mdn* = 7c/c+ | .016† |
| Hardest climb OS (last year) | 5a–7a, *Mdn* = 6a+/b | 6c/c+–7c/c+, *Mdn* = 7a+ | .006† |
| Hardest climb RP (last year) | 5a–7b+, *Mdn* = 6b+ | 7a–8a+/b, *Mdn* = 7c | < .001† |
| Hardest climb OS (last month) | 5a–6c/c+, *Mdn* = 6a+ | 6c–7c/c+, *Mdn* = 7a | .245† |
| Hardest climb RP (last month) | 5a–7a+/b, *Mdn* = 6a+/b | 6c–8a+, *Mdn* = 7a+/b | < .001† |

Note. OS = onsight; RP = redpoint; M = mean; s = standard deviation; Mdn = median.
† The analysis considered numeric grades.
P-values reported are from Mann-Whitney test.

## Data analysis

Grades on the French scale were transformed to a numeric scale (i.e. decimal climbing grade): each number grade was worth 1 "point", each letter grade above "a" adds 1/3 point, and each "+" adds 1/6 of a point (e.g. grade 6b+ was transformed into 6 + 1×1/3 + 1/6 = 6.5). "Slash" grades add 1/12 to an adjacent lower grade, e.g. 6b+/6c was translated into 6.5 + 1/12 = 6.583. Converting grades on the French scale to numeric scale is common in sport climbing research (Draper et al., 2016; Watts, 2004). Our numeric grades can be converted to IRCRA reporting scale (Draper et al., 2016) by using a simple linear transformation (Table 2). Numeric grades were summarized with mean, median, standard deviation and interquartile range.

Table 2: A comparison of different grading scales.

| French/sport | UIAA | YDC | Watts | IRCRA | Numeric |
|---|---|---|---|---|---|
| 4b | V- | 5.6 | 0.00 | 6 | 4.33 |
| 4c | V | 5.7 | 0.25 | 7 | 4.67 |
| 5a | V+ | 5.8 | 0.50 | 8 | 5.00 |
| 5b | VI- | 5.9 | 0.75 | 9 | 5.33 |
| 5c | VI | 5.10a | 1.00 | 10 | 5.67 |
| 6a | VI+ | 5.10b | 1.25 | 11 | 6.00 |
| 6a+ | VII- | 5.10c | 1.50 | 12 | 6.17 |
| 6b | VII | 5.10d | 1.75 | 13 | 6.33 |
| 6b+ | | 5.11a | 2.00 | 14 | 6.50 |
| 6c | VII+ | 5.11b | 2.25 | 15 | 6.67 |
| 6c+ | VIII- | 5.11c | 2.50 | 16 | 6.83 |
| 7a | VIII | 5.11d | 2.75 | 17 | 7.00 |
| 7a+ | VIII+ | 5.12a | 3.00 | 18 | 7.17 |
| 7b | IX- | 5.12b | 3.25 | 19 | 7.33 |
| 7b+ | | 5.12c | 3.50 | 20 | 7.50 |
| 7c | IX | 5.12d | 3.75 | 21 | 7.67 |
| 7c+ | IX+ | 5.13a | 4.00 | 22 | 7.83 |
| 8a | X- | 5.13b | 4.25 | 23 | 8.00 |
| 8a+ | | 5.13c | 4.50 | 24 | 8.17 |
| 8b | X | 5.13d | 4.75 | 25 | 8.33 |
| 8b+ | X+ | 5.14a | 5.00 | 26 | 8.50 |
| 8c | XI- | 5.14b | 5.25 | 27 | 8.67 |
| 8c+ | | 5.14c | 5.50 | 28 | 8.83 |
| 9a | XI | 5.14d | 5.75 | 29 | 9.00 |

Note: UIAA – Union Internationale des Associations d'Alpinisme, YDS – Yosemite Decimal System, IRCRA – International Rock Climbing Research Association. Sources: (Draper et al., 2016; Watts, 2004).

Agreement between the raters in each group was computed by the Kendall coefficient of concordance W (corrected for ties); 95% confidence interval for W was obtained with bootstrap (Davison & Hinkley, 1997). Intra-rater reliability was evaluated under the two-way random model with intraclass correlation coefficients (ICC), under the consistency (ICCC) and agreement (ICCA) models. Under the agreement model, standard error of measurement (SEM) was computed as $SD \times (1 - ICCA)^{1/2}$ and minimal differences needed to be considered real (MD) as $MD = SEM \times 1.96 \times 2^{1/2}$ (Weir, 2005).

To evaluate the factors of the grading bias, the association between the individual grades, and style of climbing (OS, F), type of ascent (successful, failed), sex, age, height and other variables from the questionnaire, was estimated with linear mixed effects models, where the subject was included in the model as a random effect, and the route number as a fixed effect. Separate analyses were performed for groups A and B. Interaction between the route number and the covariate was considered. The significance of the interaction term was verified with the likelihood ratio test; in case of a non-significant interaction effect the model without the interaction was considered. For the continuous covariates, a possible non-linear association was modelled by using cubic splines. In case of a non-significant non-linear term the linear association was considered.

To evaluate factors that affect the accuracy (discrepancy) of grading, the outcome variable was defined as the absolute difference between the individual grade of the route and the mean grade for that route. This outcome was then associated with the covariates by using the same approach as presented for the bias of grading. A multivariate linear mixed effects model was used to compare the reliability of grading in groups A and B, adjusting for sex, years of climbing and the highest RP level, which were included in the model because of potential confounding. To take into account the difference in the difficulty of the routes in both groups, the difference between the OS level and the mean grade of each route was also included in the multivariate model.

A p-value of less than .05 was considered as statistically significant. The analysis was performed with the R language for statistical computing, R version 3.0.3.

## RESULTS

The distribution of individual grades is given in Figure 1. For each route, there are between 6 and 10 different grades. However, in most cases, only 2 or 3 grades prevail. Slash grades (e.g. 4c+/5a) are uncommon. In group A, "+ grades" (e.g. 5b+) are rare, while in group B, they are more common, except for the route B3, where they are much less frequent. As the distribution of grades is in most cases approximately symmetrical, average and median grades are similar (Table 3).

Figure 1. Frequency distributions of grades for routes 1–5 of A and B groups of climbers.

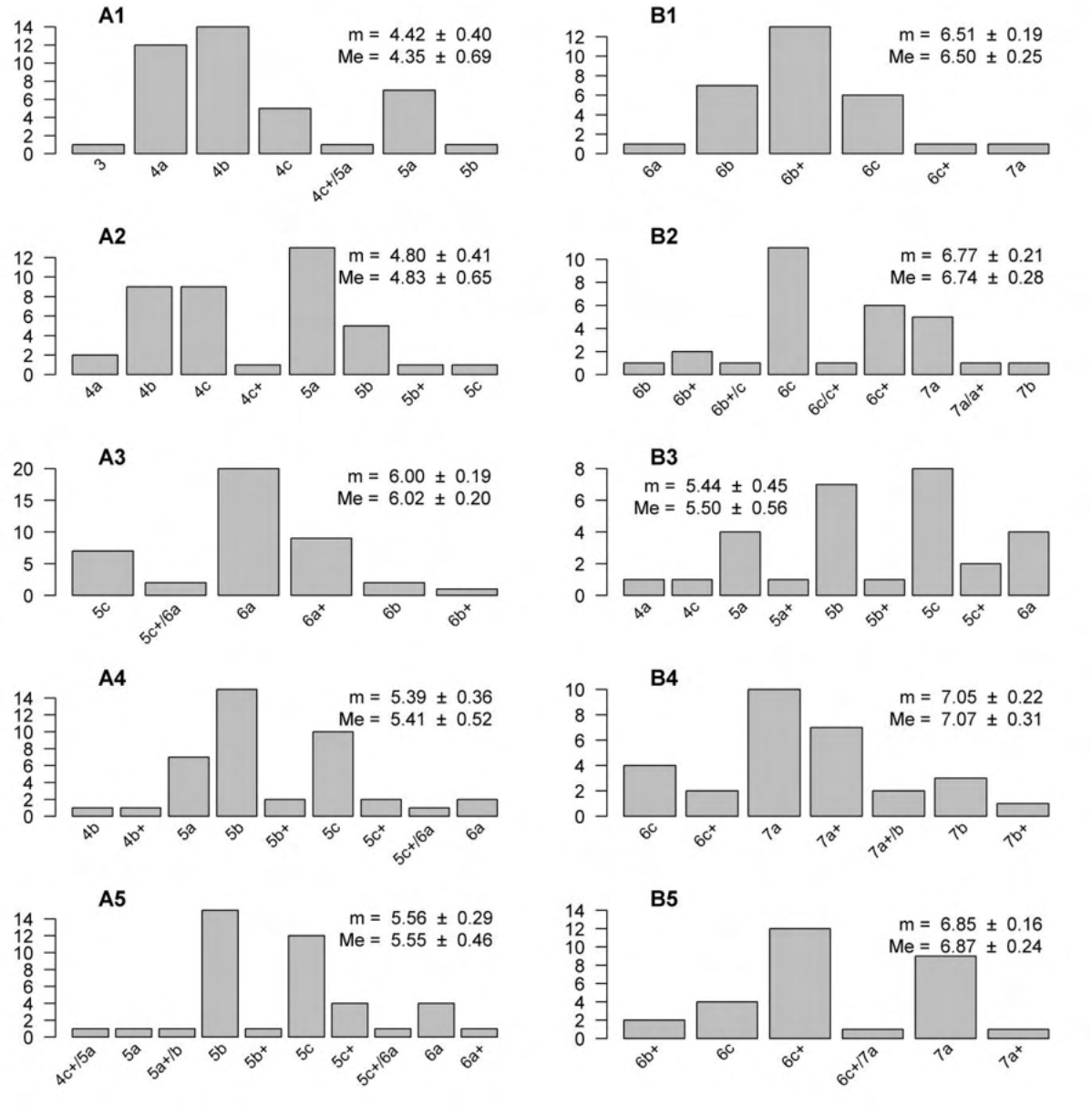



Table 3. Distributional statistics of routes' numeric grades and their average (m1–5).

| Route | Group A | | | | | Group B | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *Mdn* | *s* | *IQR* | % success | *M* | *Mdn* | *s* | *IQR* | % success |
| 1 | 4.42 | 4.35 | .40 | .69 | 100% | 6.51 | 6.50 | .19 | .25 | 86% |
| 2 | 4.80 | 4.83 | .41 | .65 | 98% | 6.77 | 6.74 | .21 | .28 | 83% |
| 3 | 6.00 | 6.02 | .19 | .20 | 56% | 5.44 | 5.50 | .45 | .56 | 100% |
| 4 | 5.39 | 5.41 | .36 | .52 | 90% | 7.05 | 7.07 | .22 | .31 | 55% |
| 5 | 5.56 | 5.55 | .29 | .46 | 93% | 6.85 | 6.87 | .16 | .24 | 66% |
| $m_{1–5}$ | 5.23 | 5.27 | .21 | .31 | 87% | 6.52 | 6.50 | .18 | .19 | 78% |

Note. M = mean; Mdn = median; s = standard deviation; IQR = interquartile range; % success = percentage of climbers who successfully climbed (i.e. onsighted or flashed) the route.

The variability of grades, as expressed by the standard deviation and interquartile range, is generally around 2 times higher in group A than in group B. In each group, there is one exception, with route A3 having a much lower variability than other group A routes, and route B3 having a much higher variability than other routes of group B. The success rate, i.e. the percentage of climbers who successfully onsighted or flashed a route, was higher (87%) in group A than in group B (78%).

Agreement on rankings of the 5 routes is high among climbers of both groups (Table 4), although a little higher in group A (Kendall coefficient of concordance W = .89), than in group B (W = .86). Intraclass correlation coefficients (for a single judge) under both the agreement and the consistency model is somewhat higher in group B than in group A. Related to that, standard errors of measurement (SEM) and minimal differences needed to be considered real (MD) are lower in group B.

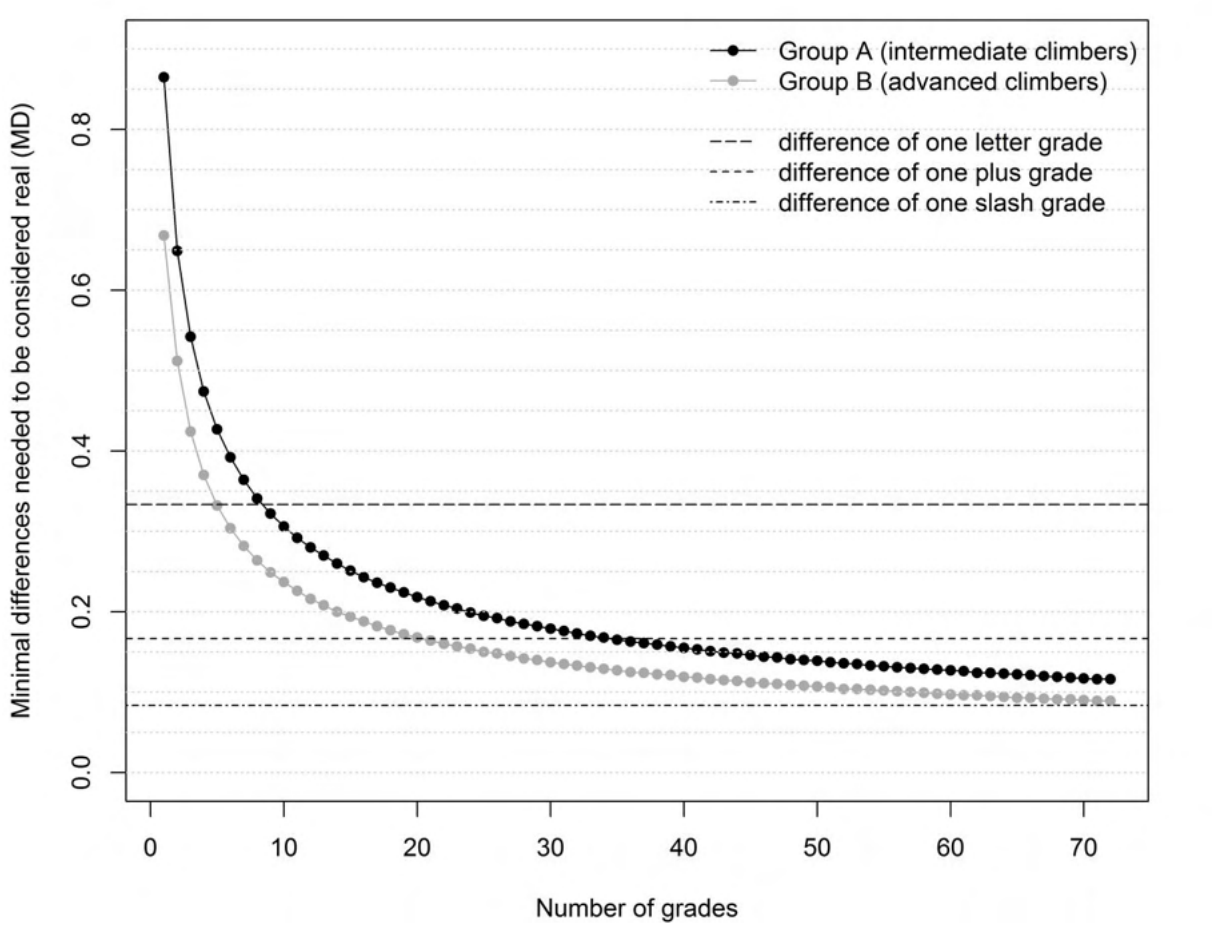
Table 4. Statistics related to the reliability of climbing grades.

| | Group A<br>value, [95% CI] | Group B<br>value, [95% CI] |
|---|---|---|
| Kendall *W* | .89 [.58, 1.00] | .86 [.42, 1.00] |
| Mean correlation between judges | .85 | .95 |
| Standard deviation of all grades | .66 | .63 |
| Intraclass correlation coefficient (agreement) | .78 [.54, .97] | .85 [.67, .98] |
| Intraclass correlation coefficient (consistency) | .82 [.61, .97] | .89 [.74, .99] |
| Standard error of measurement (*SEM*) | .31 | .24 |
| Minimal differences to be considered real (*MD*) | .86 | .67 |

Note. CI = confidence interval.

If the routes in group A are graded only by one climber, the MD is .86 of the grade, corresponding to a difference of more than two letter grades on the French grading scale, while in group B, the MD is .67, corresponding to a difference of exactly two letter grades. In group A, 9, 35 and 139 individual grades are needed to be 95% certain that the true differences in the difficulty of two routes is one letter, plus and slash grade, respectively (Figure 2); in group B, 5, 21 and 82 individual grades are needed to achieve the same.

Figure 2. Minimal difference between numeric (decimal) grades of two sport climbing routes needed to be 95% certain that the two routes are really of different difficulty as a function of the number of individual grades for group A and group B.




**Factors of bias in grading**

There was a significant difference in mean grades between the successful and failed attempts in group B. Climbers in group B, failing to successfully complete the route gave, on average, higher grades (mean difference: .15, 95% CI [.04, .26], p = .007). Women in group B graded route number 3 significantly higher than men (mean difference: .36, 95% CI [.11, .61], p = .005). Route B3 was graded lower by taller climbers (mean difference, expressed as a difference in height of 10 cm: .10, 95% CI [.00, .21], p = .058), but this was a consequence of confounding by sex, as no association between the grade and height was observed when adjusting for sex (p-value for height when adjusting for sex p = .325). No association between the grades and the style of the ascent (OS/flash), age, weight of the climbers, BMI, years of, best OS ascent, best

RP ascent, amount of climbing in the past year, amount of climbing in the past month, percentage of outdoor climbing and percentage of indoor climbing within CCL was observed ($p > .05$).

**Factors of accuracy in grading**

There was a significant difference in grading accuracy between routes A3 and A1 (on average, a .20 higher accuracy was observed for route A3 than A1, 95% CI [.11, .29], $p < .0001$), routes A5 and A1 (on average, a .08 higher accuracy was observed for route A5 than A1, 95% CI [.00, .17], $p = .048$) as well as routes B3 and B1 (on average, a .22 less accurate grading was observed for route B3 than for B1, 95% CI [.13, .30], $p < .0001$), while the accuracy of grading between the other pairs of routes was not significantly different (A2 and A1 $p = .830$, A4 and A1 $p = .177$, B2 and B1 $p = .334$, B4 and B1 $p = .239$, B5 and B1 $p = .878$).

Women in group A graded the climbs more accurately than men (.05, 95% CI [.00, .12], $p = .046$), but this was not observed in group B ($p = .693$). Similarly, taller people in group A were less accurate (expressed as a difference in height of 10 cm: .04, 95% CI [.00, .01], $p = .023$). When sex and height were simultaneously included in the model for the accuracy of grading in group A, none of the factors exhibited a significant association ($p = .825$ and $p = .275$, for sex and height, respectively), which was a consequence of strong collinearity between sex and height.

Climbers in group B with a higher OS level graded route B3 significantly less accurately than climbers with a lower OS level (.34, 95% CI [.15, .54], $p = .001$). This was not observed for other routes or in group A ($p > .05$). We identified an influential observation, a climber with a very high OS and RP level, who graded route B3 significantly lower than the other climbers. When removing this observation, climbers in group B with a higher OS level graded the routes significantly less accurately (.09, 95% CI [.01, .18], $p = .033$). Grades of route A1 were significantly more accurate if the climbers spent more time climbing indoors (expressed as a difference of 10 percentage points: .03, 95% CI [.01, .06], $p = .020$). This was not observed for the other routes or in group B ($p > .05$).

No association between the accuracy of grading and type of a climb (successful, failed), style of the ascent (OS/F), age, weight, years of climbing, amount of climbing in the past year, amount of climbing in the past month and the percentage of indoor climbing within CCL was observed ($p > .05$).

In the multivariate analysis (Table 5) climbers in group A graded the climbs significantly less accurately when adjusting for the other covariates (.12, 95% CI [.04, .20], p = .003, Model 1). However, when the difficulty gap (the difference between the best ever OS level and the mean grade of a specific route) was included in the model (Model 2), the difference in the accuracy of grading between the groups was no longer significant, and the only significant independent factor for the accuracy of grading was the difficulty gap. Climbers with a larger difference between their best OS level and the mean grade of the route graded the climbs significantly less accurately when adjusting for the other factors (.11, 95% CI [.08, .15], p < .0001).

Table 5. Multivariate models for the accuracy of grading.

| | Coefficient | *SE* | 95% CI | *p*-value |
|---|---|---|---|---|
| Model 1 | | | | |
| Group B:A | -.12 | .04 | [-.20, -.05] | .003 |
| Sex F:M | -.03 | .02 | [-.09, .02] | .146 |
| Climbing experience | .00 | .00 | [-.01, .01] | .817 |
| Best RP level (ever) | .02 | .03 | [-.03, .08] | .396 |
| Model 2 | | | | |
| Group B:A | -.00 | .04 | [-.08, .09] | .995 |
| Sex F:M | -.04 | .02 | [-.09, .01] | .097 |
| Climbing experience | .00 | .00 | [-.01, .01] | .948 |
| Best RP level (ever) | -.05 | .03 | [-.11, .02] | .111 |
| Difficulty gap† | .12 | .02 | [.08, .15] | < .0001 |

Note. SE = standard error of the estimated regression coefficient; CI = confidence interval.
† Difficulty gap is the difference between the best OS level and the mean grade of the route

## DISCUSSION

This is the first study to examine the reliability of assigning a difficulty grade to a climb. As expected, the climbers did not completely agree about the difficulty of any of the routes. However, even though they had only one attempt at the route and were not allowed to communicate about the difficulty of the route with each other, the reliability of grades was relatively high. ICCs for an individual climber were comparable to those observed for the highly trained and experienced judges in the artistic gymnastics of the European championships (Leskošek, Čuk, Pajek, Forbes, & Bučar-Pajek, 2012) or University games (Bučar et al., 2012). A similar trait is observed based also on the Kendall W coefficient, with the only exception that here, the reliability is slightly higher in group A than in group B. Note that the difference is very small, and probably a consequence of the larger difficulty range of the routes climbed in

group A than in group B; it is known that the coefficients of intra-rater reliability increase when the rated objects (in our case routes) are more heterogeneous (Weir, 2005).

A higher reliability of grading in group B is observed when considering the standard error of measurement (SEM), which shows the typical error when a randomly selected climber grades a typical route. MD values, which determine the smallest true difference in the difficulty of two routes, for which one can be 95% confident that one route is harder than the other, are consistent with the SEM values. MD in group A and group B, expressed on the French grading scale, are around two letter grades. As an example, this means that if a typical climber graded a route 6c, then we can be 95% confident that the other route is truly easier if it was graded 6a or less. This example clearly emphasizes that caution is needed when comparing the difficulty of different routes or the accomplishments of different climbers. Claims which route is the hardest, or which climber is the best, are commonly made based on a difference of one plus grade, however, the true difference between two routes, each with only one individual grade, should be at least four times larger, so that we could claim this with 95% confidence. Based on our study, 21 (advanced group) and 35 (intermediate group) individual grades are required to reliably declare the true difference of one plus grade as real. This is also important when evaluating the performance of climbers that is based only on their hardest ascent (either climbed OS or RP), which is common in rock climbing research (Draper et al., 2016).

The only identified source of bias when grading the difficulty of the climbs in our study was the unsuccessful completion of the climb. In our study, the climbers in the advanced group (group B) who did not successfully complete the climb gave, on average, significantly higher grades. The magnitude of the bias was in the range of one plus grade. This was not observed in the intermediate group, which was to a large extent the consequence of a very small number of unsuccessful attempts (less than 10%, except for the hardest route), reducing the power of the statistical analysis. This is important in practice as it suggests that the climbers that successfully complete the climb should be more eligible to propose a grade, otherwise the consensus grade could be biased.

Grades for different routes exhibited significantly different accuracy. The easier routes in both groups of climbers were graded significantly less accurately than the harder routes. Consistent with this result, the climbers with higher climbing level in the advanced group graded the climbs significantly less reliably. The multivariate model showed more accurate grading in the advanced group when controlling for sex, climbing experience and RP level. This effect

disappeared, however, when the difference between the climbing level and the difficulty of the route was included in the model. This is important, as it suggests that the only factor associated with the reliability of grading is the gap between the climbing level and the difficulty of the route. The climbers in our study were mainly regular visitors of the climbing centre and were therefore very familiar with the grading scale used there. This could explain why climbing experience was not a significant factor for the reliability of grades in the multivariate model. Climbing Centre Ljubljana is one of the very few places in Slovenia where it is possible to climb on graded indoor lead routes; therefore, our study sample was in that sense very homogenous. It would be interesting to conduct a study with a heterogeneous sample, where we expect that factors other than the difficulty gap could be related to the reliability of climbing grades.

We used a simple linear transformation of the French grades into numeric grades. There are other linear transformations of the grading scales to a numeric scale (Draper et al., 2016; Watts, 2004). Note however, that using any linear transformation to a numeric scale will lead to identical conclusions as presented herein. What should, however, be investigated in the future, is the appropriateness of assuming linearity throughout the range of the grading scale when converting the grades to a numeric scale for the purpose of a statistical analysis. Namely, it gets progressively more difficult for a climber to make progress on the grading scale as he/she moves from easier to harder grades, i.e., it takes more time (and effort) for a climber to progress by e.g. one letter grade, as the climbers progresses to higher grades. While this holds for any sports discipline (say 100-meter sprint), it would be interesting to see if this, given the subjective nature of grading sport climbing routes (as opposed to objective measurement of time in say the 100-meter sprint), also translates to the scale at which the difficulty of a climb is measured (which is clearly linear for the 100-meter sprint).

The grade given to a route is the main and usually the only indicator of performance in (outdoor) sport climbing. It is used to compare climbers' ability and routes' difficulty. It is also used in scientific research, e.g. as a basis for forming ability groups of climbers or for evaluating the effect of training intervention. Therefore, it is crucial to know the metric characteristics of a climbing grade scale. This is the first study to establish these characteristics. It was found that the grade, given by a single climber, is in general unreliable (inaccurate) and biased. Several factors, which were suspected to influence reliability and bias were examined, but only few were found significant. Between those factors the difficulty gap between climber's ability and route difficulty was found as the most important. In some cases also other factors were found

important, i.e. unsuccessful attempts, tiredness and sex. Surprisingly, some other factors, e.g. body height (adjusted for sex), ability level and climbing experience were not found significant. Based on these findings several recommendations arise. First, it was found that several tens of independent individual grades are needed for the average (or median) grade to be sufficiently accurate (i.e. within one "tick mark" of the difficulty scale). Additionally, to prevent bias of the average grade, only rested climbers, which have actually successfully climbed the whole route in valid style and whose ability is close to the difficulty of the route, should be considered for proposing the grade of the route.

In conclusion, if a climbing grade is to be used as a criterion for evaluating the climber's ability, it needs to be based on sufficiently large number of individual grades given by climbers whose climbing ability is close to the difficulty of the route. Importantly, the number of required individual grades depends on the smallest difference of the climber's ability that one would like to measure (evaluate). E.g., if this difference was one plus grade then, based on our study, around 20 independent individual grades for advanced group of climbers and more than 30 for intermediate group are required.

**Acknowledgment**

We would like to thank the participants, Matjaž Jeran and the staff of the Climbing Centre Ljubljana for allowing us to use their facilities and for their help in organising the study, the Faculty of Sports, University of Ljubljana, for funding, Andraž Gregorčič and Blaž Beličič for their technical support and Jernej Maučec for his help with the editing of the manuscript.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## REFERENCES

Bučar, M., Čuk, I., Pajek, J., Karacsony, I., & Leskošek, B. (2012). Reliability and validity of judging in women's artistic gymnastics at University Games 2009. *European Journal of Sport Science*, 12(3), 207-215. doi:10.1080/17461391.2010.551416

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application (Vol. 1):* Cambridge university press.

de Moraes Bertuzzi, R. C., Franchini, E., Kokubun, E., & Kiss, M. A. P. D. M. (2007). Energy system contributions in indoor rock climbing. *European Journal of Applied Physiology*, 101(3), 293-300. doi:10.1007/s00421-007-0501-0

Draper, N., Dickson, T., Blackwell, G., Fryer, S., Priestley, S., Winter, D., & Ellis, G. (2011). Self-reported ability assessment in rock climbing. *Journal of Sports Sciences*, 29(8), 851-858. doi:10.1080/02640414.2011.565362

Draper, N., Dickson, T., Fryer, S., & Blackwell, G. (2011). Performance differences for intermediate rock climbers who successfully and unsuccessfully attempted an indoor sport climbing route. *International Journal of Performance Analysis in Sport*, 11(3), 450-463. doi:10.1080/24748668.2011.11868564

Draper, N., Giles, D., Schöffl, V., Konstantin Fuss, F., Watts, P., Wolf, P., . . . Fryer, S. (2016). Comparative grading scales, statistical analyses, climber descriptors and ability grouping: International Rock Climbing Research Association Position Statement. *Sports Technology*, 1-7. doi:10.1080/19346182.2015.1107081

Draper, N., Jones, G. A., Fryer, S., Hodgson, C. I., & Blackwell, G. (2008). Effect of an on-sight lead on the physiological and psychological responses to rock climbing. *Journal of Sports Science and Medicine*, 7, 492-498.

Grant, S., Hasler, T., Davies, C., Aitchison, T., Wilson, J., & Whittaker, A. (2001). A comparison of the anthropometric, strength, endurance and flexibility characteristics of female elite and recreational climbers and non-climbers. *Journal of Sports Sciences*, 19(7), 499-505. doi:10.1080/026404101750238953

Leskošek, B., Čuk, I., Pajek, J., Forbes, W., & Bučar-Pajek, M. (2012). Bias of judging in men's artistic gymnastics at the european championship 2011. *Biology of Sport*, 29(2), 107. doi:10.5604/20831862.988884

Morrison, A. B., & Schöffl, V. R. (2007). Physiological responses to rock climbing in young climbers. British *Journal of Sports Medicine*, 41(12), 852-861. doi:10.1136/bjsm.2007.034827

Saul, D., Steinmetz, G., Lehmann, W., & Schilling, A. F. (2019). Determinants for success in climbing: A systematic review. *Journal of Exercise Science & Fitness*, 17(3), 91-100.

Schöffl, V. R., & Kuepper, T. (2006). Injuries at the 2005 world championships in rock climbing. *Wilderness and Environmental Medicine*, 17(3), 187-190. doi:10.1580/pr26-05

Sherk, V. D., Sherk, K. A., Kim, S., Young, K. C., & Bemben, D. A. (2011). Hormone responses to a continuous bout of rock climbing in men. *European Journal of Applied Physiology*, 111(4), 687-693. doi:10.1007/s00421-010-1685-2

Valenzuela, P. L., de la Villa, P., & Ferragut, C. (2015). Effect of Two Types of Active Recovery on Fatigue and Climbing Performance. *Journal of Sports Science and Medicine*, 14(4), 769. doi:26664273

Wall, C. B., Starek, J. E., Fleck, S. J., & Byrnes, W. C. (2004). Prediction of indoor climbing performance in women rock climbers. *The Journal of Strength and Conditioning Research*, 18(1), 77-83. doi:10.1519/1533-4287

Watts, P. B. (2004). Physiology of difficult rock climbing. *European Journal of Applied Physiology*, 91(4), 361-372. doi:10.1007/s00421-003-1036-7

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. The *Journal of Strength and Conditioning Research*, 19(1), 231-240. doi:10.1519/15184.1