Reliability of ratings of perceived exertion during progressive treadmill exercise

Kevin L Lamb, Roger G Eston, David Corns

Abstract
Objective—To assess the test-retest reliability (repeatability) of Borg’s 6–20 rating of perceived exertion (RPE) scale using a more appropriate statistical technique than has been employed in previous investigations. The RPE scale is used widely in exercise science and sports medicine to monitor and/or prescribe levels of exercise intensity. The “95% limits of agreement” technique has recently been advocated as a better means of assessing within-subject (trial to trial) agreement than traditional indicators such as Pearson and intraclass correlation coefficients.

Methods—Sixteen male athletes (mean (SD) age 23.6 (5.1) years) completed two identical multistage (incremental) treadmill running protocols over a period of two to five days. RPEs were requested and recorded during the final 15 seconds of each three minute stage. All subjects successfully completed at least four stages in each trial, allowing the reliability of RPE responses to be examined at each stage.

Results—The 95% limits of agreement (bias ± 1.96 × SDp) were found to widen as exercise intensity increased: 0.88 (2.02) RPE units (stage 1), 0.25 (2.53) RPE units (stage 2), −0.13 (2.86) RPE units (stage 3), and −0.13 (2.94) RPE units (stage 4). Pearson correlations (0.81, 0.72, 0.65, and 0.60) and intraclass correlations (0.82, 0.80, 0.77, and 0.75) decreased as exercise intensity increased.

Conclusions—These findings question the test-retest reliability of the RPE scale when used to monitor subjective estimates of exercise intensity in progressive (or graded) exercise tests.

Keywords: rating of perceived exertion (RPE); limits of agreement analysis; exercise testing; Pearson correlation coefficients; intraclass correlation coefficients

On account of its strong positive associations with physiological variables, such as oxygen uptake, heart rate, and blood lactate concentrations (typically established during continuous, incremental exercise) the rating of perceived exertion (RPE) concept is widely accepted as a means of estimating exercise intensity in progressive (or graded) exercise tests. The 95% limits of agreement (LoA) for aerobic exercise. In the same way, RPE is also widely used clinically, particularly with cardiac patients and patients receiving β-blocker therapy.

Recent research, however, has begun to question the efficacy of RPE in both healthy and cardiac populations, the indications being that ratings recorded during graded exercise testing do not match the levels of relative physiological intensity that they are assumed to.

Fundamental to this concern over the validity of the RPE scale is the issue of its reliability. As a measurement tool cannot be deemed valid without also being reliable, it is surprising that little attention has been paid to establishing the reliability (or repeatability) of RPE under repeated (identical) exercise testing conditions. Instead, it has often been assumed that once subjects have been “introduced” to the Borg 6–20 RPE scale through standardised instructions and/or so-called “anchoring” techniques, then their understanding of its function has been established.

On the basis of empirical evidence, the early studies by Skinner et al and Stamford are often referred to in support of the reliability of the RPE scale. These articles reported test-retest correlation coefficients ranging from 0.71 to 0.90, depending on the mode of exercise and whether the protocol was incremental or otherwise, which were deemed sufficiently high to indicate “consistency of results”. More recently, Wenos et al reported reliability correlations of 0.96, 0.97, and 0.72 at intensities of 30, 50, and 70% of peak oxygen uptake respectively during a discontinuous walking protocol. However, when the same three exercise intensities were applied in separate constant load protocols, the reliability correlations were less impressive (0.53, 0.94, and 0.67 respectively).

A feature common to the limited research on RPE reliability is the lack of regard given to the appropriateness of the statistical techniques used to quantify reliability. A recent movement led by British exercise scientists has highlighted the misuse of certain statistics, especially the bivariate correlation, as indicators of reliability. This concern is applicable to the RPE scale, as it has almost always been considered to provide interval level data which have subsequently been analysed with parametric statistics. As correlation coefficients do not actually assess the level of agreement between repeated measures (they quantify the degree of association), it is not yet known whether the RPE scale yields repeatable values when applied in a typical test-retest investigation. The 95% limits of agreement (LoA)
Table 2  Test-retest analysis of rating of perceived exertion values at each exercise intensity level and across trials

<table>
<thead>
<tr>
<th>Stage</th>
<th>T1 Mean (SD)</th>
<th>T2 Mean (SD)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.5 (1.75)</td>
<td>9.6 (1.54)</td>
<td>3.42</td>
<td>0.004</td>
</tr>
<tr>
<td>2</td>
<td>12.1 (1.86)</td>
<td>11.9 (1.20)</td>
<td>0.77</td>
<td>0.451</td>
</tr>
<tr>
<td>3</td>
<td>13.6 (1.90)</td>
<td>13.7 (1.35)</td>
<td>-0.34</td>
<td>0.736</td>
</tr>
<tr>
<td>4</td>
<td>15.4 (1.86)</td>
<td>15.5 (1.37)</td>
<td>-0.33</td>
<td>0.743</td>
</tr>
</tbody>
</table>

Reliability of ratings of perceived exertion

The reliability of ratings of perceived exertion (RPE) was assessed using the intraclass correlation coefficient (ICC). The ICC was calculated from repeated measures analysis of variance and was of the type that accounted for trial to trial variability (ICC = (MSs − MSw)/MSs, where MSs = (SS_{trials} + SS_{interaction})/(df_{trials} + df_{interaction})). As a secondary marker of the consistency of the exercise protocol over the two trials (and therefore as a check on whether there was a systematic bias between trials), the heart rate responses were also analysed with repeated measures analysis of variance and, as with RPE responses, paired t tests for each exercise stage.

Methods

SUBJECTS

Sixteen healthy male athletes from the University of Wales volunteered to take part in this study (mean (SD) age 23.6 (5.1) years, height 1.80 (0.11) m and body mass 73.5 (9.4) kg). Subjects were habitually engaged in middle or long distance training and club level competition, either as runners or rowers. All subjects abstained from caffeine and strenuous physical activity on the day of each test, and completed an informed consent form and a health questionnaire just before being tested. Approval for the study was granted by the ethics committee of the School of Sport, Health and Physical Education Sciences at the University of Wales.

PROCEDURES

Subjects attended the laboratory on two occasions, each time being subjected to a graded exercise test. The graded exercise tests comprised two identical running protocols on an electronically driven Powerjog (GM200) treadmill. The protocol was extracted from the physiologic testing guidelines of the British Association of Sport and Exercise Sciences and incorporated a five minute warm up at 3.13 m/s (7 mph) at 0% gradient, followed by three minutes at 3.58 m/s (8 mph). Thereafter, the velocity remained constant while the gradient was increased in increments of 2.5% every three minutes. For each session, heart rate and RPE were recorded in the last 15 seconds of each three minute increment until either an RPE of 17 or volitional exhaustion was reached.

Table 2  Test-retest analysis of rating of perceived exertion values at each exercise stage

<table>
<thead>
<tr>
<th>Homoscedasticity</th>
<th>r</th>
<th>p</th>
<th>Bias (SD)</th>
<th>95% Limits of agreement</th>
<th>ICC</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.147</td>
<td>0.88 (1.03)</td>
<td>0.88 ± 2.02</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.833</td>
<td>0.25 (1.29)</td>
<td>0.25 ± 2.53</td>
<td>0.80</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>-0.03</td>
<td>0.914</td>
<td>-0.13 (1.46)</td>
<td>-0.13 ± 2.86</td>
<td>0.77</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>-0.04</td>
<td>0.870</td>
<td>-0.13 (1.50)</td>
<td>-0.13 ± 2.94</td>
<td>0.75</td>
<td>0.60</td>
</tr>
</tbody>
</table>

ICC, Intraclass coefficient.
Table 3 Test-retest maximal heart rate reserve (%) at each exercise stage and across trials

<table>
<thead>
<tr>
<th>Stage</th>
<th>T1 Mean (SD)</th>
<th>T2 Mean (SD)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>63.8 (8.3)</td>
<td>61.5 (6.5)</td>
<td>1.79</td>
<td>0.093</td>
</tr>
<tr>
<td>2</td>
<td>70.7 (8.3)</td>
<td>70.3 (6.6)</td>
<td>0.34</td>
<td>0.739</td>
</tr>
<tr>
<td>3</td>
<td>77.6 (8.7)</td>
<td>76.5 (6.9)</td>
<td>0.94</td>
<td>0.363</td>
</tr>
<tr>
<td>4</td>
<td>84.8 (8.7)</td>
<td>84.8 (7.0)</td>
<td>−0.06</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Maximal heart rate reserve (%) was calculated as (Exercise HR − Resting HR)/(HRmax − Resting HR), where HRmax is estimated from the equation HRmax = 220 − age, and HR is heart rate.

All data analyses were performed using SPSS 8.0 for Windows.

Results

Table 1 presents the mean RPE values recorded for each exercise stage in T1 and T2. Analysis of variance disclosed significant main effects for levels ($F = 358.3$, $p<0.001$), and non-significant effects for trials ($F = 0.59$, $p>0.10$). The levels × trials interaction, however, was significant ($F = 5.8$, $p<0.01$), due solely to significant ($p<0.005$) bias being present at the lowest exercise intensity (stage 1), although the difference is less than one unit. For stages 2–4, the differences between means were not significantly greater than zero.

The normality of the test-retest differences in RPE values was confirmed for each exercise intensity (K-S Lilliefors statistics; $p>0.05$). Likewise, these differences were found to be homoscedastic, with correlations between the absolute differences and the mean of the two trials being small and non-significant (table 2). Consequently, table 2 shows the 95% LoA analyses, and, for comparative purposes, the ICC and Pearson correlation coefficients.

Heart rate responses did not vary significantly over trials ($F = 0.6$, $p>0.10$), but showed an expected increase across levels ($F = 198.7$, $p<0.001$). The trials × levels interaction was not significant ($F = 2.1$, $p>0.10$), and table 3 shows that the replicated exercise protocol elicited relative heart rates free of significant systematic bias at each intensity level.

Discussion

These data provide a unique perspective on the repeatability of RPE during progressive treadmill exercise. Adopting the “worst case scenario” approach to interpreting LoA analyses of Nevill and Atkinson,1 an athlete in this study reporting an RPE of 12 during stage 2 in trial 1 could possibly have reported a value as high as 15 or as low as 10 during the same stage a few days later (values rounded up). Likewise, a first trial RPE of 16 during stage 4 could have been as high as 19 or as low as 13 in trial 2. As this type of analysis is new to perceived exertion research, there is no scope for comparison with previously published findings. However, given the circumstances of this study, such a degree of “uncertainty” observed in relatively active subjects must raise questions about the reliability of RPE (and therefore its validity) in less active or exercise naïve people.

The more traditional marker of reliability calculated alongside the LoA (the Pearson correlation coefficient) does provide scope for placing the present findings into context. Moreover, three out of the four of this study’s exercise intensities (stages 2–4) lend themselves to an interpretation that is as unfavourable as the LoA. Skinner et al10 reported what can only be an overall Pearson correlation of 0.80 for incremental cycling exercise (the data from all stages being combined), but did not provide statistics on RPE reliability for each intensity across the range used. Interestingly, the same type of analysis of the present data yields a correlation of 0.86. While Skinner et al considered their finding to reflect “sufficiently high reliability”, Noble and Robertson11 challenged this on the grounds of the 36% of unexplained variance in the relation. The claim of Stamford14 to have established the reliability of the RPE scale is questionable not only from a statistical perspective, but also from a design perspective. Although he used different modes of exercise (treadmill walking and jogging, cycling, and stot stopping) and variable intensities in his study, it does not seem that (for each mode) the RPE data were collected in an identical manner over the “repeated” trials.

In this study, the mean %MHHR for each exercise intensity was very similar across the two trials. The difference at the lowest intensity was the largest, reflecting a systematic bias of about 2.3%, although this was not significant. However, in terms of practical significance, such variability is not “large”. Of course, a finding of zero bias between repeated measures does not mean there was no within-subjects variation (random error) in heart rates. Even though the exercise protocol and measurements (potential sources of random error) were controlled, considerable random error (due to biological variation) is to be expected.24 Furthermore, even if a systematic bias was present generally, the relation between RPE and heart rate is not so strong as to be causal, that is, it could not be assumed that a given %MHHR bias (in either direction) would elicit a corresponding RPE bias.

With regard to the RPE correlations in this study, both forms decline in magnitude as the exercise intensity increases, suggesting decreasing reliability. At the same time, the random error can be seen to increase through the 95% LoA becoming wider. Although such concordance (in terms of the trends) is somewhat reassuring, the case of the lowest exercise intensity exemplifies well how inappropriate the two correlation coefficients can be as measures of reliability. Here it is clear that the “high” Pearson correlations and ICCs (0.81 and 0.82 respectively) mask the significant bias (0.88 RPE units), the existence of which Bland and Altman25 would argue (from a medical perspective) is sufficient to render the current data useless for the purpose of assessing reliability. These opposing interpretations reinforce the need for sports and exercise scientists to understand statistical techniques and recognise their importance in the wider process of measurement and evaluation.

The 95% LoA method of analysis indicates a degree of test-retest variability of up to almost three RPE units, or, in qualitative terms,
Reliability of ratings of perceived exertion

339

The implications of this for other repeated exercise trials by as much as three test and the perceptually based run test, exercise protocols, such as the Sjostrand cycle Y

termination may be equivalent to a rating as of the current sample, the above marker for test

the scale for such patients is no better than that

maximal oxygen uptake. For example, Noble24 cites a “rule of thumb” that coronary heart dis-

exercise tolerance, or as an adjunct indicator

of heart rate to reflect a surrogate measure of heart rate during progressive exercise. In a preparatory technique, it does seem to have preparedness for the paper.

Additional research is needed to verify these findings for different exercise tests (with different samples) and to assess the effectiveness of multiple exposures (or habituation) to the scale in enhancing its reliability.

KL led the writing of the paper and analysed all the data for presentation to the Journal. RE generated the idea, supervised the research project at the University of Wales, Bangor, and contributed ideas on revisions and amendments to the paper. In discussion with KL, DC (the supervisee of RE) contributed to the design of the study and collected the data. RE and KL are guarantors for the paper.


Take home message

In adopting more appropriate methods of analysis than previously used, the test-retest reliability of ratings of perceived exertion for estimating exercise effort during incremental (graded) exercise has been found to be suspect. Users of this scale are advised to assess for themselves the reliability of the scale before accepting its validity.