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 $\pm 1.96 \times SD_{diff}$) 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.**

(*Br J Sports Med* 1999;**33**:336–339)

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

Department of Physical Education and Sports Science, University College Chester, United Kingdom
K L Lamb

School of Sport, Health and Physical Education Sciences, University of Wales, Bangor, Wales
R G Eston
D Corns

Correspondence to:
Dr K L Lamb, Department of Physical Education and Sports Science, University College Chester, Parkgate Road, Chester CH1 4BJ, United Kingdom.

Accepted for publication
14 June 1999

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 a widely accepted means of estimating exercise intensity in adults, and, to a lesser extent, in children.[1] Its validity has been claimed for different modes of exercise, including cycling,[2 3] walking and running,[4] stepping,[5] swimming,[6] and rowing,[7] and its use has been advocated as a means of providing a safe and effective training intensity for aerobic exercise.[8] In the same way, RPE is also widely used clinically, particularly with cardiac patients[9] and patients receiving β-blocker therapy.[10] Recent research, however, has begun to question the efficacy of RPE in both healthy and cardiac populations,[11] 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[12] and/or so-called "anchoring" techniques,[13] then their understanding of its function has been established.

On the basis of empirical evidence, the early studies by Skinner *et al*[2] and Stamford[14] 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*[15] 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[16–19] 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 1 Test-retest rating of perceived exertion values at each exercise intensity level and across trials

Stage	*T1 Mean (SD)*	*T2 Mean (SD)*	*t*	*p*
1	10.5 (1.75)	9.6 (1.54)	3.42	0.004
2	12.1 (1.86)	11.9 (1.20)	0.77	0.451
3	13.6 (1.90)	13.7 (1.35)	−0.34	0.736
4	15.4 (1.86)	15.5 (1.37)	−0.33	0.743

technique[20] is the more appropriate statistical approach, as it allows reliability judgments to be based on the size of the within-subjects (trial to trial) variability, and not the relative position of scores across the two trials (whether the subject with the highest score in trial 1 also has the highest in trial 2, or whether the same subject has the lowest score in both trials, and so on). Accordingly, the purpose of this study is to examine the reliability of the RPE scale during standardised and replicated exercise conditions, using the LoA form of statistical analysis.

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 physiological testing guidelines of the British Association of Sport and Exercise Sciences[21] 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

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

ICC, Intraclass coefficient.

In the initial test, subjects were familiarised with the treadmill and introduced to the Borg 6–20 RPE Scale.[12] Before each exercise session, subjects were given standardised RPE instructions[22] to read and seek clarification if necessary. In this way, the RPE scale was being used in its so called estimation or response mode.[23]

The testing sessions took place no more than five days and no less than two days apart. Height and body mass data were collected at the beginning of the initial session using standard laboratory procedures. Subjects' heart rates were measured by telemetry (Polar, Beat; Bodycare Products), at rest (after remaining supine for five minutes) and during exercise, and were subsequently expressed as a percentage of maximal heart rate reserve (%MHRR) for each exercise stage. The ambient temperature in the laboratory over the course of the study was 18–23°C, and for each test cool air was directed on to the subject by a pedestal fan (Pifco 1004) for added comfort. The RPE scale was positioned within sight and reach throughout each exercise bout.

STATISTICAL ANALYSIS

Data were analysed with a two way analysis of variance (trials × levels) with repeated measures to assess the variability of RPE responses across trials and exercise intensities. Post hoc analysis used the 95% limits of agreement procedure of Bland and Altman[20] to examine the test-retest reliability of the RPE values recorded for each of the first four exercise intensities (as all subjects completed at least four stages). This technique requires the calculation of the mean difference (bias) between trial 1 (T1) and trial 2 (T2) and ± 1.96 × SD of these differences (the 95% limits). Assuming that the test-retest differences are (*a*) not significantly greater than zero, (*b*) normally distributed and (*c*) unrelated to the mean of the two trials (homoscedastic), these 95% limits form the reliability statistics. Accordingly, condition (*a*) was examined using paired *t* tests (with a Bonferroni adjustment of α to 0.0125), condition (*b*) with the K-S Lilliefors statistic, which tests whether the sample data are from a normal population, and condition (*c*) with a Pearson correlation coefficient.

Following the recommendations of Atkinson and Nevill,[19] the reliability analysis was extended with the calculation of both the intraclass correlation (ICC) and Pearson correlation coefficients. These are the statistics most often used to assess the reliability of the RPE scale. 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 MSw = $(SS_{Trials} + SS_{Interaction})/(df_{Trials} + df_{Interaction})$). 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.

Table 3 Test-retest maximal heart rate reserve (%) at each exercise stage and across trials

Stage	T1 Mean (SD)	T2 Mean (SD)	t	p
1	63.8 (8.3)	61.5 (6.5)	1.79	0.093
2	70.7 (8.3)	70.3 (6.6)	0.34	0.739
3	77.6 (8.7)	76.5 (6.9)	0.94	0.363
4	84.8 (8.7)	84.8 (7.0)	−0.06	0.951

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, $p>0.4$). 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[17], 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 al*[2] 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 Robertson[13] challenged this on the grounds of the 36% of unexplained variance in the relation. The claim of Stamford[14] 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 stool stepping) 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.[19] 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 Altman[20] 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,

perceptions changing (in either direction) from, for example, "extremely light" to harder than "very light", "light" to harder than "somewhat hard", or "hard" to harder than "very hard". Such inconsistency may have particular relevance for situations in which RPE is used as a dependent variable in some form of intervention study, or where it is used as a surrogate measure of heart rate to reflect a person's state of metabolic stress and/or exercise tolerance, or as an adjunct indicator (or precursor) of physical work capacity or maximal oxygen uptake. For example, Noble[24] cites a "rule of thumb" that coronary heart disease patients who reach an RPE rating of 15 will not complete more than one more stage of the Bruce treadmill protocol. If the reliability of the scale for such patients is no better than that of the current sample, the above marker for test termination may be equivalent to a rating as low as 12 for some or as high as 18 for others. Likewise, RPE unreliability would undermine the efficacy of perceptually based submaximal exercise protocols, such as the Sjostrand cycle test and the perceptually based run test, described by Noble and Robertson.[13] With these protocols, improvements in physical work capacity or running speed after aerobic training are estimated on the basis of a criterion RPE of 15, the physical work capacity/running speed at RPE 15 before training being compared with that at RPE 15 after training.

From a methodological perspective, this study did not allow any "improvements" in reliability to occur through repeated exposure to the RPE scale. It is not known whether a third, or even fourth trial, would have yielded narrower (better) limits of agreement as a consequence of the subjects becoming more fully habituated to the RPE concept. In addition, no attempt was made to employ an "anchoring" technique analogous to that described recently for cycling exercise by Noble and Robertson.[13] Although no empirical evidence has been published to support the effectiveness of such a preparatory technique, it does seem to have face validity and deserves to be investigated further.

In conclusion, the present findings cast doubt on the test-retest reliability of the established 6–20 Borg RPE scale for estimating exercise effort during progressive exercise. In adopting a more appropriate form of statistical analysis than has previously been used with RPE data (the 95% LoA), trained male athletes were found to differ in their responses to repeated exercise trials by as much as three RPE units. The implication of this for other trained and untrained people is the prospect of a tool that is invalid for use in exercise testing. 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.

1 Lamb KL, Eston RG. Effort perception in children. *Sports Med* 1997;**23**:139–48.
2 Skinner JS, Hutsler R, Bergsteinova V, *et al*. The validity and reliability of a rating scale of perceived exertion. *Med Sci Sports* 1973;**5**:94–6.
3 Morgan WP, Borg GAV. Perception of effort in the prescription of physical activity. In: Craig TT, ed. *Humanistic and mental aspects of sports, exercise and recreation*. Chicago: American Medical Association, 1976:126–9.
4 Robertson RJ. Central signals of perceived exertion during dynamic exercise. *Med Sci Sports Exerc* 1982;**14**:390–6.
5 Walker CAH, Lamb KL, Marriott HE. The validity of using ratings of perceived exertion to estimate and regulate exercise intensity during stepping ergometry. *J Sports Sci* 1996; **14**:102–3.
6 Ueda T, Kurokawa T. Relationships between perceived exertion and physiological variables during swimming. *Int J Sports Med* 1995;**16**:385–9.
7 Marriott HE, Lamb KL. The use of ratings of perceived exertion for regulating exercise levels in rowing ergometry. *Eur J Appl Physiol* 1996;**72**:267–71.
8 American College of Sports Medicine. *Guidelines for exercise testing and prescription* (3rd ed.). Philadelphia: Lea & Febiger, 1991.
9 Pandolf KB. Advances in the study and application of perceived exertion. *Exerc Sport Sci Rev* 1983;**11**:118–58.
10 Eston RG, Connolly D. The use of ratings of perceived exertion for exercise prescription in patients receiving β-blocker therapy. *Sports Med* 1996;**21**:176–90.
11 Whaley MH, Brubaker PH, Kaminsky LA, *et al*. Validity of rating of perceived exertion during graded exercise in apparently healthy adults and cardiac patients. *J Cardiopulm Rehabil* 1997;**17**:261–7.
12 Borg GAV. *An introduction to Borg's RPE-scale*. New York: Mouvement Publications, 1985.
13 Noble BJ, Robertson RJ. *Perceived exertion*. Champaign: Human Kinetics, 1996:78–9.
14 Stamford BA. Validity and reliability of subjective ratings of perceived exertion during work. *Ergonomics* 1976;**19**:53–60.
15 Wenos DL, Wallace JP, Surburg PR, *et al*. Reliability and comparison of RPE during variable and constant exercise protocols performed by older women. *Int J Sports Med* 1996;**17**:193–8.
16 Atkinson G. A comparison of statistical methods for assessing measurement repeatability in ergonomics research. In: Atkinson G, Reilly T, eds. *Sport, leisure and ergonomics*. London: E & FN Spon, 1995:218–22.
17 Nevill AM, Atkinson G. Assessing agreement between measurements recorded on a ratio scale in sports medicine and sports science. *Br J Sports Med* 1997;**31**:314–18.
18 Lamb KL. Test-retest reliability in quantitative physical education research: a commentary. *European Physical Education Reviews* 1998;**4**:145–52.
19 Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;**26**:217–38.
20 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.
21 British Association of Sport and Exercise Sciences. *Position statement on the physiological assessment of the elite competitor.* Leeds: White Line Press, 1988.
22 American College of Sports Medicine. *Guidelines for exercise testing and prescription*. Philadelphia: Lea & Febiger, 1991:71.
23 Myles WS, Maclean D. A comparison of response and production protocols for assessing perceived exertion. *Eur J Appl Physiol* 1986;**55**:585–7.
24 Noble BJ. Clinical applications of perceived exertion. *Med Sci Sports Exerc* 1982;**14**:406–11.

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.