Barnett, Lisa; Reynolds, John; Faigenbaum, Avery D.; Smith, Jordan J.; Harries, Simon; Lubans, David R.; 'Rater agreement of a test battery designed to assess adolescents' resistance training skill competency'. Published in Journal of Science and Medicine in Sport Vol. 18, Issue 1, p. 72-76 (2015)

Available from: http://dx.doi.org/10.1016/j.jsams.2013.11.012

Blind Title Page

Rater agreement of a test battery designed to assess adolescents’ resistance training skill competency

Abstract

Objectives: The study aim was to assess rater agreement of the Resistance Training Skills Battery (RTSB) for adolescents. The RTSB provides an assessment of resistance training skill competency and includes six exercises. The RTSB can be used to assess performance and progress in adolescent resistance training programs and provide associated feedback to participants. Individual skill scores are based on the number of performance criteria successfully demonstrated and an overall resistance training skill quotient (RTSQ) is created by summing the six skill scores. **Design/Method:** The eight raters had varying experience in movement skill assessment and resistance training and completed a 2-3 hour training session in how to assess resistance training performance using the RTSB. The raters then completed an assessment on six skills for 12 adolescents (mean age=15.1 years, SD =1.0, six male and six female) in a randomised order. **Results:** Agreement between seven of the eight raters was high (20 of the 21 pairwise correlations were greater than 0.7 and 13 of the 21 were greater than 0.8). Correlations between the eighth rater and each of the other seven raters were generally lower (0.45 to 0.78). Most variation in the assigned RTSB scores (67%) was between cases, a relatively small amount of the variation (10%) was between raters and the remainder (23%) was between periods within raters. The between-raters coefficient of variation was approximately 5%. **Conclusion:** The RTSB can be used reliably by those with experience in movement skill assessment and resistance training to assess the resistance skill of adolescents.

Word Count: 239
2995 for the article

Key Words: Physical fitness; exercise; strength training; motor skills

Introduction

Youth physical activity guidelines have identified strength as an important health related factor [1], and current public health objectives now aim to increase the number of school-age youth who participate in muscle strengthening activities [2]. Regular participation in an age-appropriate related resistance training program can enhance muscular fitness, power and motor skill performance [3-5]. Furthermore, resistance training interventions in youth can have a positive influence on metabolic health, body composition, cardiorespiratory fitness, blood lipids, bone mineral density and insulin sensitivity [6, 7]. There is clear evidence that resistance training can be a safe, effective and worthwhile activity for children and adolescents provided that appropriate training guidelines are followed and qualified instruction is available [8-10].

Resistance training programs are usually evaluated using ‘product’ type fitness tests that assess muscular strength and local muscular endurance [11, 12] (i.e. ‘how heavy’ or ‘how many repetitions), rather than providing meaningful feedback on movement skill technique. Movement skill technique is important when assessing the fundamental movement skill competency (i.e. the ability to throw and kick) of children and adolescents as this type of ‘process’ assessment involves specific feedback regarding which particular components of the skill need to be improved for satisfactory movement skill performance. A process oriented skill assessment involves assessing the ‘presence’ or ‘absence’ of a number of components/criteria per skill that are considered essential for mastery of that particular skill. For example, a component of a successful kick is the ability to place the non-kicking foot even with or slightly behind the ball [13].

At present, a process oriented assessment is not commonly used in youth resistance training programs. Therefore, the Resistance Training Skills Battery (RTSB) was developed to assess adolescents’ skill competency in resistance training [14]. Potentially, the RTSB could be used to assess each participant's individual performance and, when appropriate, provide general information regarding group level performance and progress in adolescent resistance training programs, while providing constructive feedback to participants. The RTSB includes six skills with each skill involving movements which are considered to provide the basis for strength development. These six skills are summed to provide a resistance training competency total quotient (RTSQ). Initial research was conducted to determine the one week test-retest reliability of the RTSB with 63 adolescents (mean age of 14 years). It was found the RTSB could be used to reliably rank both male and female adolescents on overall resistance training competency and that the RTSB had the necessary sensitivity to detect small changes in resistance skill competency. The RTSB also showed evidence of construct validity, with the RTSQ predicting 39% of variance in muscular fitness (assessed using handgrip strength, timed push-up and standing long jump tests) [14]. However, the skills in this study were all assessed by the same research assistant, so rater agreement for the RTSB has not been established.

Rater agreement is the measurement of the consistency or agreement in scores obtained from two or more raters [15, 16], and is important to consider when assessing movement skill proficiency. It is imperative to demonstrate that if a group of raters receive the same training in instrument administration, that they are then able to reliably assess participants' skill competency, otherwise the instrument has limited applicability in the field. Studies of rater agreement in the health literature are often underreported, and when they are reported, they tend to be incomplete and inadequate; therefore, there is a need for such studies to be performed in the future [17].

When assessing rater agreement it is possible to test the effect of the participant, the rater and also the order of assessment. Analysing for a potential order effect enables an understanding of whether there is a systematic difference occurring during assessment independent from rater differences. For example, if a rater first assesses two adolescents who are poor performers of a skill, the rater as a consequence may then inflate the score of the next adolescent simply because the performance is so much improved from the previous skill performance. Agreement studies that don't test for an order effect are therefore not assessing a potential source of systematic variation. Therefore, the aim of the current study was to assess inter-rater agreement and reliability of the RTSB using the RTSQ. Ordering effects were also assessed.

Methods

Approval for the study was gained from the University Research Ethics Committee and the school principal from one secondary school in New South Wales (NSW), Australia. Parental permission and child assent were obtained. The protocol is described elsewhere [14], but briefly, students completed assessments at school as part of 'all male' or 'all female' groups of three or four. Students observed demonstrations by a research assistant and only questions relevant to the particular exercise (e.g., number of repetitions) were allowed. Encouragement was provided but not skill specific feedback. Students completed two trials of four repetitions for each skill in the following order: (i) body weight squat (ii) push-up (iii) lunge (iv) suspended row (v) standing overhead press and (vi) front support with chest touches. Trunk stability is assessed via *front support* and *chest touches*, upper body pushing strength is assed via a *push-up*, upper body pulling strength is assessed via a *suspended row*, lower body bilateral strength is assessed via a *squat*, and lastly, lower body unilateral strength is assessed via the *lunge*. The exercises therefore target the major muscle groups: lower body (*squat/lunge*), chest, back and arms (*push-up* and *suspended row*), shoulders (*standing overhead press*) and core (*front support with chest touches*). The exercises were all done with only body weight – no additional weight was added. A digital video camera recorded skill attempts. Each skill has four (push-up and suspended row) or five (body weight squat, lunge, standing overhead press and front support with chest touches) performance criteria. Please see Table 1. Scoring was based on the best performance of the skill during the four repetitions for each of the two trials. Participants were awarded a '1' for each criteria correctly demonstrated and '0' if it was not correct. The score for each trial were summed and then totaled for each skill and then the skill scores were all summed for the resistance training skill quotient (RTSQ) (possible range 0 to 56) [14].

TABLE 1 – see supplementary file

For this current study, video assessments of the six skills were selected by taking a stratified random sample of 12 students from the pool of 63 students in the original study (44 males, 19 females, Mean Age 15.1, SD = 1.0). Assessments used for analysis in this manuscript were the first assessments of two trials (assessments were conducted on two occasions seven days apart to determine test retest reliability; this has already been reported [14]). Firstly, all video assessments were grouped by sex and then tertiles were assigned based on the scores assigned previously by the research assistant. Girls and boys performed differently in this original assessment. For girls, the first tertile was a score less than 43 out of the possible 56, the second tertile was from 43 to <47 and the third tertile was ≥ 47. For boys, the first tertile was < 40, the second tertile 40 to < 47 and the third tertile was ≥ 47. Then two students were randomly selected from each of the six strata.

Eight raters independently assessed the six videotaped skills for all 12 students (a total of 72 skill assessments per rater). Raters had a range of backgrounds with varying combinations of relevant qualifications, movement skill assessment coding and resistance training experience. Please see Box 1.

Box 1

Rater	Relevant Degree/Qualification	Movement skill assessment experience	Resistance training experience
r1	Physical Education	Extensive experience	25 years recreational Strength/Conditioning Coach
r2	Physical Education	Limited experience	10 years recreational
r3	No	Extensive experience	<5 years recreational
r4	Physical Education	Limited experience	<5 years recreational
r5	Physical Education Strength/Conditioning Coach	Limited experience	10 years recreational
r6	Physical Education	Moderate experience	8 years recreational
r7	Physical Education	Extensive experience	8 years recreational
r8	Exercise Science	Little experience	<5 years recreational

Note. Extensive experience = coding >500 performances, Moderate experience = coding >300 performances, Limited experience = undergraduate unit, Little experience = a lecture or two.

Each rater was sent a RTSB training package that included videos for each skill that had been classified in terms of the previous scoring as ‘poor’ (i.e. few criteria performed correctly), ‘medium’ (most criteria performed correctly) or ‘high’ performance (all criteria performed correctly). For example there were three videos of three different students performing the squat to a ‘poor’, ‘medium’ or ‘high’ level. Raters were asked to firstly view these videos and the accompanying scoring sheets which showed how the student had been previously coded. When raters considered they understood the scoring protocol they were asked to code the six skills for each of the 12 students in a specific pre-determined order that was assigned to them.

Raters spent on average 90 minutes developing an understanding of the scoring protocol and 120 minutes scoring the trials.

The order of student assessment (i.e. 1-12 positions) was randomised for each rater. A rater (1 ... 8) was allocated to a presentation order for the assessments by randomly selecting a column from the design matrix for a row-column design. The row-column design (rows = positions and columns = raters) had the following properties: (1) Each student was assessed once by each rater, (2) Each student was evaluated no more than once in a position, (3) Each pair of students appeared in the same assessment position between 4 and 7 times, and (4) Each student was preceded by every other student no more than once. The design was not balanced for the residual effect (if any) of the evaluation of the preceding student on the evaluation of the current student as this would have required a larger design (such as a Williams' Square) and recruitment of more raters. Nevertheless the chosen design allowed these residual, or carryover, effects to be estimated.

As a check on the overall discrimination of the eight raters, a nested analysis of variance (ANOVA) with raters regarded as a random effect and students regarded as a fixed effect, explored whether there was significant variation between the means of the 12 students. In addition, for each student, the variance between the raters was calculated as a check on the stability of the overall assessments and Bartlett's test was used to assess the homogeneity of these within-student (i.e. between-rater) variances. Similarly, for each rater, the variance between the students was calculated as a check on their discrimination and Bartlett's test was used to assess the homogeneity of these within-rater (i.e. between-student) variances. Diagnostic plots of fitted values and residuals were viewed to assess outliers and to check for variance-mean relationships. Agreement between pairs of raters was assessed by computing Pearson's correlation coefficient. The residual effect of the assessment of the previous student on the assessment of a student was investigated via a mixed model analysis (using REML) in which raters (1 to 8) and positions (1 to 12) were regarded as random effects and students, and the previous student (including no previous student, i.e. assessment occurred in the first position), were regarded as fixed effects. Lastly, in a random effects analysis, variance components for students, raters, and, assessments within raters were estimated to enable intraclass correlations to be reported All analyses were conducted using GenStat Release 14.2 statistical software [18].

Results

Mean scores for the 12 cases ranged from 33.9 to 49.75 (Table 2). Table 2 also shows the original tertile assigned to each case (i.e. High/Medium/Low) and the minimum and maximum score assigned for each case by any of the raters. Diagnostic plots of fitted values and residuals showed only one potential outlier (rater 8's relatively low assessment of case #139). The nested ANOVA indicated significant variation between the cases ($p < 0.001$). Two of the 12 cases appeared to have relatively high between-rater variance (or potential discordance), namely cases #130 and #157 and two of the 12 cases, namely cases #146 and #139, appeared to have relatively low between-rater variance (or reasonable concordance). The variance also appeared to vary with the mean (lower variances at the high end of the scale where the scores have an upper bound of 56, and higher variances in the middle of the scale, namely 27 to 40). Homogeneity of these between-rater (within-case) variances was explored using Bartlett's test and, despite the apparent differences, there was no significant departure from homogeneity of variance ($\chi^2_{11} = 8.18$; $p = 0.697$).

TABLE 2

Agreement between seven of the eight raters was high (20 of the 21 pairwise correlations were greater than 0.7, 13 of the 21 were greater than 0.8 and the range was 0.67 to 0.94). Correlations between the eighth rater (r8) and each of the other seven raters were generally lower (0.45 to 0.78) and this eighth rater also had the highest mean score (Table 3). Mean scores for the eight raters ranged from 37.50 to 43.67. Table 3 also shows the maximum and minimum score given by each particular rater. (Table 3). ANOVA indicated significant variation between the raters ($p < 0.001$). Two of the raters (r1 and r7) appeared to have relatively high between-case variance, indicating either high discrimination or instability, or, both. One rater (r2) appeared to have relatively low between-student variance, indicating either low discrimination or, moderate to high, stability, or, both. Homogeneity of these between-student (within-rater) variances was explored using Bartlett's test and, despite the apparent differences, there was no significant departure from homogeneity ($\chi^2_7 = 5.79$; $p = 0.565$).

The mixed model analysis showed no significant effect of first position (i.e. no previous assessment) versus the other positions ($p = 0.788$) and no overall residual or carryover effect of the assessment of the previous student on the current assessment of a student ($p = 0.411$). When raters (n=8) and students (n=12) were regarded as random effects, the total variance in the 96 RTSB scores was mostly between students (67%), a relatively small amount of the variation (10%) was between raters and the remainder (23%) was between periods within raters (Table 4). The between-raters coefficient of variation was approximately 5%.

TABLE 3 and 4

Discussion

This study has shown that the RTSB [14] can be used reliably to assess the resistance training skill competency of adolescents. The variation between raters was relatively small, with most of the variation being due to the particular cases that were assessed. Seven of the eight raters commonly had high agreement (pairwise correlations over 0.80). Even the eighth rater (who generally had lower agreement), still had only two pairwise correlations that were below 0.68. Studies which use a process oriented battery to assess the movement skills of children have reported high inter-rater reliability statistics. For example, a recent Brazilian study involving children reported an ICC of 0.88 for the locomotor subtest and 0.89 for the object control subtest in the Test of Gross Motor Development (TGMD-2) [19]. Similarly, a study of Australian preschool children using the TGMD-2 reported similar results for both subtests (locomotor *ICC* = 0.92 and object control *ICC* = 0.90) [20]. Thus, our estimate of the interrater reliability statistic (ICC = 0.67) for our assessment battery of resistance training skills is lower than such statistics reported in studies of children's movement skill ability that use process oriented instruments. This could be for several reasons. Firstly, the Brazilian study described their raters as 'expert' and the Australian study reported raters received 12 hours of training, whereas in the current study only three of the eight raters could be called 'expert' (based on a criteria of extensive experience in movement skill assessment combined with some resistance training experience) and the training period was less. Secondly, our study involved eight raters whereas the Brazilian study involved three raters and the Australian study used four raters. Having a higher number of raters purposively selected to have varying levels of experience will increase the observed between-rater variance component and, all other things being equal, decrease the interrater reliability. Furthermore the inclusion of one relatively inexperienced professional whose agreement with the other seven raters was low may have further inflated the between and within rater variance components [21]. In a post-hoc analysis, we excluded the 8th rater and found that the ICC measure of rater agreement increased from 0.67 to 0.71. Finally it does not appear that either of these studies used a mixed model where potential variance was explained at each potential level (the student/the rater - both between and within) which may also have influenced results. It has been noted in an article which proposes guidelines for reporting reliability and agreement studies that although ICC values are reported in many health research studies it is often not clear what ICC is being reported and how the analysis has been performed [17]. The same article also suggests that values above 0.60, 0.70, or 0.80 are all reported as minimum values for reliability coefficients, and these values should be seen as appropriate for group-level comparisons and/or research purposes; accordingly, the ICC value found for the current study could be regarded as having met a minimum standard[17].

Of note, when considering the mean scores for each rater, the raters with less experience coding movement skills had higher overall means than the three raters with considerable experience, even though all raters had relevant backgrounds. The rater with the highest mean score (r8) was the rater with little previous experience. It might be expected that those with experience in observing and coding movement patterns in adolescents would exhibit higher levels of discernment when assessing movement skills and therefore apply more precise scoring. This information may be useful to researchers recruiting movement skill assessors, as well as physical education teachers who may solicit assistance from others during class testing.

Furthermore 10 of the 12 cases were all rated in the same tertile as those originally assigned, giving further evidence towards the potential of this instrument to be used by a number of raters in a reliable fashion. This study also showed there were no order effects indicating that raters should be able to assess participants in any order and still achieve reliability. However it must be noted that whilst the order of watching and assessment was specified clearly for each rater, the assessment order was not supervised by the researchers. Order effects are not generally reported in literature reporting reliability of movement skill assessment, although one study in preschool children reported that they intentionally ordered the skills for ease of assessment [22]. This study did not however assess any potential order effects in the rater agreement analysis [22].

Conclusion

In conclusion, given the high agreement between seven of the eight raters and the relatively low between-rater coefficient of variation, namely 5%, we believe that the RTSB can be used reliably to assess skill competency in selected resistance training exercises in adolescents.

Practical Implications

- Raters' with experience in movement skill assessment coupled with at least recreational resistance training experience, are able to reliably assess participants' skill competency after a short training.
- The RTSB can be used reliably in adolescent resistance training interventions when supervised by trained assessors with the appropriate backgrounds.
- Results from the current study coupled with our previous findings highlight the potential usefulness of the RTSB.

Acknowledgements

Thank you to the school and students for their involvement. The first author is supported by a National Health and Medical Research Council early career fellowship. There has been no financial support for this study.

References

1 Strong WB, Malina RM, Blimkie CJR, *et al.* Evidence based physical activity for school-age youth. *J Pediatr* 2005; 146: 732-37.

2 World Health Organization. *Global Recommendations on Physical Activity for Health*. Geneva 2010; 55.

3 Faigenbaum AD, Kraemer WJ, Blimkie CJR, *et al.* Youth Resistance Training: Updated Position Statement Paper from the National Strength and Conditioning Association. *J Strength Cond Res*. 2009; 23: S60-79.

4 Behringer M, vom Heede A, Matthews M, *et al.* Effects of Strength Training on Motor Performance Skills in Children and Adolescents: A Meta-Analysis. *Pediatr Exerc Sci*. 2011; 23: 186-206.

5 Harries S, Lubans DR, Callister R. Resistance training to improve power and sports performance in adolescent athletes: A systematic review and meta-analysis. *J Sci Med Sport*. 2012; 15: 532-40.

6 Faigenbaum AD, Myer GD. Pediatric Resistance Training: Benefits, Concerns, and Program Design Considerations. *Curr Sports Med Rep*. 2010; 9: 161-68.

7 Benson AC, Torode ME, Singh MAF. Effects of resistance training on metabolic fitness in children and adolescents: a systematic review. *Obes Rev*. 2008; 9: 43-66.

8 Behm DG, Faigenbaum AD, Falk B, *et al.* Canadian Society for Exercise Physiology position paper: resistance training in children and adolescents. *Appl Physiol Nutr Metab*. 2008; 33: 547-61.

9 Lloyd R, Faigenbaum AD, Stone M. Position statement on youth resistance training: the 2014 international consensus. British Journal of Sports Medicine 2013(epub ahead of print).

10 Lubans DR, Morgan PJ, Aguiar EJ, *et al.* Randomized controlled trial of the Physical Activity Leaders (PALs) program for adolescent boys from disadvantaged secondary schools. *Prev Med*. 2011; 52: 239-46.

11 Faigenbaum AD, Milliken G, Moulton L, *et al.* Early muscular fitness adaptations in children in response to two different resistance training regimens. *Pediatr Exerc Sci*. 2005; 17: 237-48

12 Lubans DR, Sheaman C, Callister R. Exercise adherence and intervention effects of two school-based resistance training programs for adolescents. *Prev Med*. 2010; 50: 56-62.

13 Ulrich DA. Test of Gross Motor Development (2nd ed). Austin, TX 2000.

14 Lubans DR, Smith J, Harries S, *et al.* Development, Test-Retest Reliability and Construct Validity of the Resistance Training Skills Battery. *J Strength Cond Res*. 9000; Publish Ahead of Print: 10.1519/JSC.0b013e31829b5527.

15 Goodwin LD. Interrater agreement and reliability. *Meas Phys Educ Exerc Sci*. 2001; 5(1): 13-34.

16 Posner KL, Sampson PD, Caplan RA, *et al.* Measuring interrater reliabity among multiple raters: An example of methods for nominal data. *Stat Med*. 1990; 9: 1103-15.

17 Kottner J, Audig´e L, Brorson S, *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011; 64 96-106.

18 Payne RW, Harding SA, Murray DA, *et al.* The Guide to GenStat® Release 14 Part 2: Statistics. VSN International: Hemel Hempstead, UK 2011.

19 Valentini NC. Validity and Reliability of the TGMD-2 for Brazilian Children. *J Mot Behav*. 2012; 44: 275-80.

20 Barnett L, Hinkley T, Hesketh K, *et al.* Use of electronic games by young children and fundamental movement skills. *Percept Mot Skills*. 2012; 114: 1023-34.

21 Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979; 86: 420-28.

22 Williams HG, Pfeiffer KA, Dowda M, *et al.* A field-based testing protocol for assessing gross motor skills in preschool children: The children's activity and movement in preschool study motor skills protocol. *Meas Phys Educ Exerc Sci*. 2009; 13: 151-65.