Psychology of Sport and Exercise 15 (2014) 569–579

Contents lists available at ScienceDirect

Psychology of Sport and Exercise

journal homepage: www.elsevier.com/locate/psychsport

Development and psychometric evaluation of a post exercise exhaustion scale utilising the Rasch measurement model

Mark D. Hecimovich [a, *], Jeremiah J. Peiffer [a, 1], Allen G. Harbough [b, 2]

[a] School of Psychology and Exercise Science, Murdoch University, South Street, Western Australia 6150, Australia
[b] Boston University, Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 2 November 2013
Received in revised form
2 June 2014
Accepted 11 June 2014
Available online 24 June 2014

Keywords:
Exercise
Exhaustion scale
Rasch model
Development

ABSTRACT

Objectives: The objective of this study to report on the development and psychometric analysis of a scale to measure post exercise exhaustion.
Design: This study utilised the Rasch measurement model for the psychometric analysis of a new scale aimed at measuring acute onset exhaustion in athletes.
Method: An extensive literature review, feedback from athletes and an expert panel from educators in psychology, sports science and exercise physiology provided feedback on the scale, providing evidence of content validity. A final survey, consisting of the 25 items and completed by three hundred and seventy-nine athletes (Sport: 187 tri-athletes and 192 cyclists; gender: 211 males, and 168 females; age: 18–25 [31], 26–35 [114], 36–45 [120], and 46+ [114]), was submitted to Rasch analysis.
Results: After amendments a final 14 item scale provided internally consistent and reliable measures of exhaustion for participants. The items of the final scale have good fit, and the scale has high PSI providing statistical evidence of reliability. The scale could benefit from items dealing with mid-range levels of exhaustion. The correlational association between the new scale and a similar scale was positive and significant correlation adding to the evidence of the validity of the new scale.
Conclusions: The scale appears to be a valuable tool for the assessment of exercise-induced acute onset exhaustion and may be an attractive option for researchers, clinicians, and coaches seeking to measure the levels of exhaustion in individuals. In addition to its valid theoretical structure and sound psychometric properties, the scale has advantages over other exhaustion or fatigue scales as it is not disease-specific.
© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Within sport, the ability to quantify an athlete's level of fatigue and/or exhaustion is essential to provide the best environment for training and competition success. Exhaustion and fatigue are two related constructs which have been examined in sport and exercise science with fatigue described as a poorly defined feeling, referring to a subjective symptom of malaise and aversion to activity, or to objectively impaired performance (Sharp & Wilks, 2002), and physical exhaustion referring to the state where a person can no longer effectively continue doing exercises and a total loss of strength or vitality (Heywood, Sabado, & De Leon, 2012). Although often used interchangeably, within the literature the term fatigue tends to be used in the medical field (Smets, Garssen, Bonke, & De Haes, 1995), while exhaustion is more commonly associated with psychological literature.

Currently, among the most common instruments employed in sports and kinesiology research, no two fatigue/exhaustion scales measure exactly the same thing with some measuring subjective experiences, others measuring exhaustion severity or impact, whilst many assess a mixture of these. Currently one and two dimensional instruments exist which measure exhaustion or fatigue; these include the Occupational Fatigue Exhaustion Recovery Scale (Winwood, Winefield, Dawson, & Lushington, 2005), the Fatigue Severity Scale (Krupp, LaRocca, Muir-Nash, & Steinberg, 1989), the Dutch Exertion Fatigue Scale (Tiesinga, Dassen, & Halfens, 1998), and the Situational Fatigue Scale (Yang & Wu, 2005). Although valid and reliable, these scales are intended to measure aspects of exhaustion, fatigue, or other constructs that may be chronic or near chronic in nature and often related to disease or pathology. Consequently, these scales may not be sensitive enough to assess exhaustion that is more transient and does not cause significant functional impairments (i.e., exercise).

Within the literature, exhaustion or fatigue is viewed as a continuum from mild, frequent complaints to severe and disabling

* Corresponding author. Tel.: +61 8 9360 2988.
E-mail addresses: m.hecimovich@murdoch.edu.au (M.D. Hecimovich), j.peiffer@murdoch.edu.au (J.J. Peiffer), agh@bu.edu (A.G. Harbough).
[1] Tel.: +61 8 9360 7603.
[2] Tel.: +1 617 353 2000.

http://dx.doi.org/10.1016/j.psychsport.2014.06.003
1469-0292/© 2014 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/).

(Kant et al., 2003), with acute onset exhaustion characterised as protective, identifiably linked to a single case, generally occurring in healthy individuals, perceived as normal, and with a rapid onset and short duration (Piper, 1989). Participation in sporting events or strenuous exercise can lead to acute onset exhaustion, which left unchecked can have accumulating effects leading to prolonged exhaustion, decrease in athletic performance and musculoskeletal injuries (Cresswell & Eklund, 2006; Galambos, Terry, Moyle, & Locke, 2005). Indeed, extreme exhaustion and a lack of energy (e.g. fatigue) are the most central aspects of the burnout experience (Goodger, Wolfenden, & Lavallee, 2007; Gustafsson, Hassmen, Kentta, & Johansson, 2008). Furthermore, mental and physical exhaustion are signs of overtraining and burnout leading to decreased athletic performance (Cox, 2007; Weinberg & Gould, 2007; Williams, 2006). Attempting to avoid the accumulative influences of acute onset exhaustion is a challenge for sport scientists, medical professionals, and coaches who need to structure stress and recovery cycles to yield optimum performance increments for the athlete without producing negative outcomes such as overtraining, burnout or overuse injuries (Kellman, 2002; Kenttä & Hassmen, 1998). A psychometrically robust instrument capable of measuring acute onset exhaustion would therefore aid sport scientists, medical professional and coaches accomplish this task. However, a scale which can be utilised rapidly after an event or exercise session which provides an objective marker for a level of exhaustion consistent with sport is not currently available.

The lack of available scales to assess fatigue and exhaustion in sport warrant the need for research in this field. For this reason, the purpose of this study was to develop a brief scale for the measurement of post exercise exhaustion. To accomplish this, in this paper we describe how a Rasch measurement model (Rasch, 1960/1980) was used to develop and psychometrically analyse a scale to measure post exercise and sport exhaustion, referred to here as acute onset exhaustion. Whilst the Rasch measurement model is being increasingly used in the development and evaluation of clinical tools in health and medical sciences, including rehabilitation science, psychology, nursing and podiatry (Hargquist, Bruce, & Gustavsson, 2009; Ramp, Khan, Misajon, & Pallant, 2009), and has been present in the exercise science literature (Linacre, 2000; Strauss, Büsch, & Tenenbaum, 2012; Wood & Zhu, 2006), it has not been elevated to mainstream consideration (as can been seen by a majority of research articles still reporting Classical Test Theory statistics and analyses) and therefore is relatively novel. Furthermore, no conceptual separation between exhaustion and fatigue has been made in this paper and the choice of the term 'exhaustion' as the principal terminology was motivated by the understanding that fatigue is often viewed and measured in sport and exercise science in a variety of laboratory-based measures (e.g. hyperthermia, accumulation of metabolic bi-products, dehydration) (Noakes, 2000). However, the ineffective adaptation to fatigue may be the antecedent for exhaustion (Olson, 2007), and therefore exhaustion may be viewed as more extreme by others. Nonetheless, and due to the definition of the constructs of exhaustion and fatigue remaining elusive, we make no attempt at elaborating on the conceptual boundaries of either. Instead, we try to make a clear distinction that the intent of the scale is to measure acute onset exhaustion across a range of mild (or none) to extreme (or severe).

Method

Survey development

Developing items that best represent acute onset exhaustion following exercise or sport as opposed to items which pertain to chronic or long-term exhaustion required consideration of a participant's internal resources and the context of the situation (Yang & Wu, 2005); in other words, the participant's subjective feeling about the state of their internal resources and also the demands of the activity (or the physical and mental work) being performed.

Although no generally accepted definition for exhaustion or fatigue exists, Smets et al. (1995) have proposed five ways it can be expressed: 1) by general remarks of a person concerning his or her functioning (for example, I do not feel rested); 2) by referring to physical sensations related to the feeling of tiredness; 3) by referring to cognitive symptoms, such as having difficulty concentrating; 4) a description of a lack of motivation to start any activity resulting in a dimension labelled Reduced Motivation, and 5) reference to a reduction in activity—a dimension labelled Reduced Activity. Points one through three are labelled General Physical and Mental Fatigue and correspond with scales developed on the basis of factor analyses as reported by others (Fawzy et al., 1990; Greenberg, Sawicka, Eisenthal, & Ross, 1992; Kobashi-Schoot, Hanewald, Van Dam, & Bruning, 1985; Vertommen & Leyssen, 1988; Wessely & Powell, 1989).

As the development of the current scales focused on measuring acute onset as opposed to chronic or long standing exhaustion, it was important that items be constructed to discriminate between these two. Smets et al.'s (1995) dimensions are more descriptive of chronic, long standing, exhaustion where as Winwood et al. (2005), in their validation of a fatigue exhaustion recovery scale, referred to Bartlett's (1953) suggestion that fatigue represents an incapacitation after an activity, suggesting a philosophical base for selecting items representing acute fatigue. For instance, the depletion of available energy by an activity (exercise or sport), initiating acute fatigue, produces consistent changes in an individual's ability to undertake self-chosen non-essential tasks in non-activity time (non-exercise). Therefore the self-report of an incapacity or unwillingness to engage in self-chosen pleasurable activities in non-activity time is related to the level of acute (post activity) fatigue. Conversely, chronic fatigue is indicated by self-reported doubt and despair in the capacity to maintain current exercise patterns; declining interest, involvement, and commitment; reduced concentration and motivation; and negative emotions, combined with physical manifestations of persistent tiredness (Meijman & Schaufeli, 1996).

In this study we carefully considered Smets et al.'s (1995) five ways via which fatigue, or exhaustion, can be expressed. It was concluded that the development of items based on these suggestions may reflect physical and emotional feelings associated with acute onset exhaustion. We also incorporated Winwood et al. (2005) suggestion of a philosophical base as well by creating items pertaining to levels of energy and engagement in activities. To ascertain how athletes viewed the construct, individuals from various sports and activities (endurance cyclists, tri-athletes, and Australian rules football) were asked to complete a six-item questionnaire following the completion of a high-intensity training session to determine the participant's current level of exhaustion and provide descriptions of how they currently felt physically and mentally. Descriptions such as "legs feeling tired", "heavy and weak" were consistently observed. Descriptors such as "sharp", "frustrated", and "satisfied" were used to describe mental exhaustion. Utilising Smets et al. and Winwood et al.'s formulations (2005), feedback and consultation with professionals and academics in exercise science and physiology, psychology and related fields and athlete feedback resulted in a preliminary 36 item scale with strong face validity.

Three versions of a 36 item scale were initially developed using similar wording for each item but differing numbers of response categories, namely, 1) a dichotomous scale, 2) a 10-point scale comparable to the Piper Fatigue Scale, and 3) a 5-point scale similar to the Multidimensional Fatigue Symptom Inventory-Short Form

(MFSI-SF). Subsequently, a panel of six experts drawn from educators in psychology, sports science and exercise physiology provided feedback on the proposed survey instrument, thus providing consensus evidence of content validity. Panel members were asked to review the scales and comment on each item and the overall format. From their suggestions a final version of the scale was developed consisting of 25 items each with a 10-point Likert-type frequency/intensity scale of response categories anchored at extreme ends by 'not at all' to 'extremely difficult'. The final set of items addressed a series of questions pertaining to affect on functioning (walking, running, activities of daily living) and associated symptoms (muscle cramping representing). The intent was to examine these items' performance using statistical and substantive criteria (i.e., items that not only show statistical dependence with another item, but also from the wording, which assesses a similar aspect of the variable, will be considered for deletion).

Though the expert panel's feedback was focused on the item content, there was a collective preference for the 10-point scale. It was determined that a 10-point Likert-type scale was reasonably justified based on the following considerations: (1) There was precedence for this level of granularity, and the choice of a 10-point scale mirrors other closely related scales such as the Lee Fatigue Scale (LFS). (2) While the focus in this study is to develop a brief instrument that can be validated via a Rasch model, the use of a 10-point scale makes analysis via alternative statistical modelling techniques (e.g., CFA) available to future researchers. (3) As the items are assessed by participants on a non-valenced frequency/intensity (non-negative quantity) scale, the lack of an exact middle (or neutral) category was not a pressing issue. Thus, the choice between an even or odd number of categories was arbitrary. (4) A 10-point scale allows for flexibility of modelling individual response styles (e.g., extreme responders or extreme avoidance) and accounting for this via collapsing categories. The benefit of this is that a set of items can be found that has an adjusted response scale with (a) all options along the scale being modal responses for some (latent) level of exhaustion, and (b) a comparable interpretation from item to item. (To accommodate the final point, it was determined that only items that adhered to a global restructuring of the response categories would be included in the final scale.)

To assist with the validation process, a subscale from an existing valid and reliable scale, the 13-item *fatigue* subscale from the LFS (Lee, Hicks, & Nino-Murcia, 1991), was administered at the same time. The LFS scale is an 18-item, 2-dimensional scale related to fatigue and energy and has been used with healthy individuals (Gay, Lee, & Lee, 2004; Lee et al., 1991) as well as in patients with cancer and HIV (Lee, Portillo, & Miramontes, 1999; Miaskowski et al., 2008) and has well-established validity and reliability (Lee et al., 1991; Lee, Lentz, Taylor, Mitchell, & Woods, 1994). The 13-item LFS *fatigue* subscale was chosen for the current study because of its measurement of fatigue and it is relatively short and easy to administer. We believe the subscale was useful to include because its construct, fatigue, is an area which is similar to exhaustion. The expectation was that the *fatigue* subscale should correlate positively and moderately with the newly developed exhaustion scale.

The final survey, as part of a four-part questionnaire labelled the Exercise Exhaustion Survey (EES) consisted of the 25 items in the newly developed scale, the Hecimovich–Peiffer–Harbough Exercise Exhaustion Scale (HPHEES); the 13 item *fatigue* subscale of the LFS; and a demographic section (age group, gender, and perceived level of fitness). For a list of the 25 items please refer to Appendix A.

Participants

The target group were adults (18 and older) associated with triathlete and cycling clubs. Numerous clubs throughout the world (Australia, Canada, New Zealand, South Africa, United Kingdom, and United States) were contacted with those located in Australia, Canada, New Zealand, and the United States agreeing to participate. The athletes were eligible for inclusion if they were 18 years of age or older; able to read, write, and understand English; had no ongoing disease which could affect their level of exhaustion; and had participated in a training session or event within 72 h prior to taking the survey. In order to provide sufficient numbers for the psychometric analysis, the researchers set an initial target of 10–15 participants per scale item. With the initial scale (after culling based on review by the expert panel) contained 25 items, this gave a target sample size of 250–375 athletes. As other researchers have indicated that there is no standard protocol for assessing necessary sample sizes (Embretson & Reise, 2000, p. 123), this sample size was deemed appropriate in that it was within the 250–500 target range suggested by Reise and Yu (1990) for models for graded responses. Furthermore, the final sample size was above the suggest range of 108–243 for item calibration suggested by Linacre (1994). As this is the first published evaluation of this survey instrument, this sample size seems to support an exploratory examination of the tool (Linacre, 1994), with the expectation that future studies will explore the instrument's psychometric properties with larger sample sizes.

Procedure

Initial contact was made with a board member of numerous triathlete and cycling clubs informing them of the intent and aim of the study, the procedures involved and time commitment for volunteering participants. If they agreed, the survey was administered via e-mail (on Survey Monkey) by the board member to their club members. All procedures were approved by the institutional ethics review board.

Data analysis

Responses to the $n = 379$ questionnaires were submitted to psychometric analysis using the polytomous Rasch model (PRM) (Andrich, 1978; Rasch, 1960/1980) via the Rasch Unidimensional Measurement Model software RUMM2030 (Andrich, Sheridan, & Luo, 2010) with the partial credit parameterization in which different items have different threshold estimates and is often termed the partial credit model. In the PRM the difficulty for each item is estimated uniquely along with the positioning of the thresholds for each item and a single unidimensional location for each person. Together with the assumption of local independence among responses of all persons to all items, a test of fit is a test of fit of both local independence and undimensionality. Different tests of fit focus on different aspects of these two main properties of the model, considered in the tests of fit in the paper. Because the model is unidimensional, the parameter estimates take account of the first dimension. Then if there is local dependence, which may reflect subdimensions between items or some form of response dependence, this can be studied by analysing the response residuals, which can include factor analysing these residuals. In this paper, where the items were constructed to be unidimensional, correlations of residuals among items were studied and acted upon. Fit to the Rasch model is an indication of the internal consistency of the set of items – one aspect of construct validity. Further evidence of construct validity lays in the basis on which items are developed, that is, the theoretical underpinnings. For more detailed explanations of the Rasch paradigm and procedures, see Bond and Fox (2007), Embretson and Reise (2000), and the online manual for the RUMM2030 software (Andrich et al., 2010). For many researchers, the Rasch model represents an advance on traditional

test theory in achieving this measurement goal (Andrich & Styles, 2004; Embretson & Reise, 2000).

Though construct validity cannot be statistically assessed simply by demonstrating a set of items that adequately measures individuals against a unidimensional scale, having such an instrument with said property does allow for measurement of individuals against a comparable scale. Furthermore, if the data fit the model, the relevant statistic to represent a person's level of exhaustion is monotonically related to their total score across items (this aggregated value, often a sum or mean, is that which is traditionally used). However, traditional raw (aggregated) scores may not be linearly incremental and caution should be exercised when attempting to treat them as measurements (Bond & Fox, 2007). That is to say, the resulting scale for the scores may not be an interval level measure (e.g., the difference between a score of 1 and 2 may not be the same "gap" as that between the values of 8 and 9).

The psychometric data analyses addressed three primary aims, the first of which was to establish the internal consistency and reliability of the scale. In other words, do the sets of items each represent a single variable at this level of scale? If they do, then one is justified in adding scores to obtain a total score on each scale and then using those total scores (or their logit-ized equivalents) for other statistical tests, such as comparisons of mean scores amongst groups or over time. The second aim was to determine whether there was evidence of significant Differential Item Functioning (DIF), that is, whether the items have the same psychometric properties across different groups of participants. If items show DIF across groups, they should not be used to compare person performance, unless individuals are from the same group. From an exploratory perspective, in this study, the groups of interest were gender, age, perceived level of fitness, and sport (tri-athlete or cyclist). The third aim was to provide evidence of the convergent validity of the HPHEES by examining its statistical correlation with the *fatigue* subscale from the established LFS (This could be done with the latent estimates from the IRT model (the logit scores) or with the aggregated sum score.)

To address the first aim, which was to establish the internal consistency and reliability of the scale, various aspects of the total scale and individual items were examined. As is common in IRT analyses, numerous aspects (statistics and graphical summaries) are examined simultaneously. A few of these were (in no order of importance): (1) The operation of the response categories is examined. The item thresholds are the cut-points at which one category becomes more likely than its neighbouring category (e.g., between Strongly Agree and Agree in a 5-point Likert scale). If these thresholds are consistently observed in order for each item, this suggests the response scale is being interpreted comparable across the items in the scale. (2) The level of difficulty for the items should target the level of ability of the participant sample. This can be examined via the joint distribution of thresholds and persons on the same continuum. (3) An additional aspect was the possible presence of item dependencies which was examined by inspection of the residual correlations between items. If items show dependency, then one item in each pair is most likely redundant and retaining both would artificially increase the overall reliability for the scale. Such dependencies may also indicate the presence of subscales which can be further examined through the principal component analysis of residuals. (4) Reliability is gauged using the Person Separation Index (PSI), which is the Rasch equivalent of Cronbach's alpha. (5) Finally, individual item fit can be assessed to suggest internal consistency reliability across different items in the scale. Three common tests include the log-residual (a statistical value related to the goodness-of-fit for the categories predicted vs. observed frequencies), the item-trait interaction (a comparable measure to the item-scale correlation in CTT), and the item characteristic curve (ICC, showing the expected item score for varying levels of the unidimensional scale, the curve, in comparison to the observed statistics for groups of participants across the entire range of person locations).

As is common in most models, including the PRM, no single test is sufficient to make a determination of fit, and thus multiple tests of fit were applied. Upon review of the initial analysis of all 25 items in the initial sale, a series of tasks were undertaken in order to achieve a more appropriately balanced scale. This mainly involved removal of items from the scale for lack of fit with these items also being studied in relation to content and other aspects of as well. This protocol is provided in detail in the Results Section below.

To address the second aim, to establish whether the items operate relatively consistently across different groups, differential item functioning across the groups for Gender, Age, and Perceived level of fitness was examined. Lastly, to address the third aim to provide further evidence of validity (this time, convergent validity), participant scores on the scale were correlated with scores from the same participants on a subscale of an existing scale that measures a construct related to exhaustion and whose validity has been established in the research literature.

The results of these analyses provide information about the validity and reliability of the scale. If these were satisfactory, the person scores (the logit scores or one-dimensional scores) can be used for further analyses as, for example, the comparison of mean scores (person locations) for the different groups of interest, and the investigation of changes in mean locations over time.

Results

Three hundred and seventy-nine athletes agreed to complete the questionnaire with the following demographics. Representing different sports were 187 (49.3%) tri-athletes and 192 (50.7%) cyclists. The sample gender distribution was 211 (55.7%) males and 168 (44.3%) females. The age range frequencies were 31 (8.2%) aged 18–25, 114 (30.1%) aged 26–35, 120 (31.7%) aged 36–45, and 114 (30.1%) aged 46 or older. Frequencies for perceived level of fitness were 20 (5.3%) rated themselves "fair", 69 (18.2%) rated themselves "good", 150 (39.6%) rated themselves "very good", 113 (29.8%) rated themselves "great", and 27 (7.1% rated themselves "exceptional" (though "poor" was a possible rating, it was not chosen by any participants in this sample)).

Using complete information for each person and item, the scale was analysed using the PRM with 10 possible response categories. The results of this PRM analysis are presented in Table 1 and provide a summary of the analyses, and the subsequent section addresses different analytical aspects in more detail.

The results of the 25 items indicated numerous pairs having disordered categories (over the 10-point response scale) and significant residual correlations. The three tests of fit – two statistical (log-residual and item-trait interaction) and one graphical (the Item Characteristic Curves, ICCs) – revealed 14 misfitting items. Three items showed Differential Item Functioning (DIF) for age and one item for perceived level of fitness. Although the Person Separation Index was high at 0.936, this result is likely to have been inflated due to the high number of item dependencies as indicated by significant residual inter-item correlations.

Upon review of the initial analysis a series of tasks were undertaken in order to achieve a more appropriately balanced scale. The following protocol was used to reduce the number of items for the scale in a systematic (and reproducible) manner. Due to specific intentions for the usage of this scale, it was decided that a few well-chosen items would suffice (e.g., 10–15 items) and an easy scoring strategy would be ideal (e.g., items handled in a consistent and comparable manner when calculating an aggregate score).

Table 1
Summary of Rasch analyses for the 25 item HPHEES with DIF according to Gender, Age, perceived Level of fitness.

Scale	Disordered categories	Significant residual correlations (>0.3)	Misfitting items	DIF			PSI[a]
				Gender	Age	Level	
HPHEES	Items 2, 6, 8–11, 13–17, 19, 20, 22, 24, 25	20 Pairs	Items 1, 2, 6, 7, 8, 9, 12, 13, 14, 15, 16, 19, 20, 21	None	Items 10, 19, 21	Item 17	0.936

[a] Person Separation Index.

The strategy presented included both an analysis of the statistics for the items and scale and an analysis of the wording for each item. Items that were flagged by the statistics were examined for wording. If any plausible issue was detected with the item, it was removed from the scale. (Please see the Appendix for the complete set of items.) Prior to aggregating subsets of the Likert-type response categories, all 25 items were examined. Those items that demonstrated poorest fit (Bonferroni adjusted p-values < 0.05) were examined. The fit of responses compared to the ICC curves were examined. Extreme departures from predicted scores (chi-sq < 0.001 for all items) resulted in 5 items being removed: 7 (*How easily can you perform your daily function*), 9 (*How difficult is it to concentrate*), 12 (*How agitated do you feel*), 16 (*How much muscle cramping are you feeling*) and 19 (*How much muscle tightness are you feeling*). Additionally, item 19 showed DIF for age and its wording was deemed to be weak and nondescript and not a good marker for exhaustion. Terms such as weak, tired and painful or achy were thought to be better indicators of a muscle sensation and therefore a physical exhaustive state and were incorporated into other items in the scale.

Fig. 1 shows the ICC for item 12 (one of the items dropped at the first stage). The ICCs show the theoretical probabilities (the continuous curve) of endorsing the item across the range of person locations on the whole set of HPHEES items, and the obtained probabilities for ten class-intervals of person locations (the dots). The dots should follow the theoretical curve closely if fit is good. The chi square statistic represents the deviations of the obtained dots from the theoretical curve. The operation of item 12 is inconsistent and tends not to discriminate as well as the other items amongst persons with different total scores particularly at the highest and lowest person locations (the obtained dots are, respectively, below and above the theoretical curve, primarily those of higher value). This may be because it addresses a feeling, or emotional state of mind, which may not be commonly perceived with acute exhaustion.

At the next stage, examination of the response categories showed a large proportion of disordered categories (most likely due to the fine granularity of a 10-point scale), and therefore were collapsed into 4 new categories: 0 = old 0, 1 = old 1 or 2, 2 = old 3, 4 or 5, and 3 = old 6, 7, 8 or 9. Based on the positive skew of responses for a larger portion of items and the non-negative aspect of the frequency/intensity response scale, this grouping strategy seemed appropriate. Additionally, all items were collapsed in a comparable fashion (with reversed grouping for reverse coded items). Fig. 2 shows the effect graphically with the Category Characteristic Curve (CCC) for item 10 (*How easily could you train some more*) of reversed thresholds and its effect on operation of the categories and Fig. 3 show the effect of collapsing categories (collapsed to 4 response categories) that now appear to be working with each category having a region in which each score has a maximum. The CCC shows the probabilities of responding in each category, across the range of person locations, but the probability structure with 10 categories shows that the 10 categories are not working.

Paired residual correlations were examined to detect potentially redundant items. Using relatively high correlations as indicators of items for closer examination, the wording of highly correlated pairs were examined. Following this, 2 items were selected for removal. Item 11 (*How knackered/bushed are you*) was removed because of conceptual similarity with 15 (*How easily can you complete sub-maximal training associated with your sport*), 17 (*How easily can you compete in your sport*) and 20 (*How easily can you complete maximal training associated with your sport*). Additionally, this item demonstrated disordered thresholds (this criterion was used because it was desired to have every one of the new Likert-type categories to be modal for some level of fatigue/exhaustion). Item 20 (*How easily can you complete maximal training associated with your sport*) was removed due to it being flagged in multiple residual pairs and the item's wording. Item 15 (*How easily can you complete sub-maximal training associated with your sport*) closely resembled item 20 and was not removed at this at stage but closely monitored in subsequent examination stages.

With the remaining items, the ICCs were examined for items with a Bonferroni adjusted p-value < 0.01. Only one item was flagged that also had wording issues. This was item 17 (*How easily*

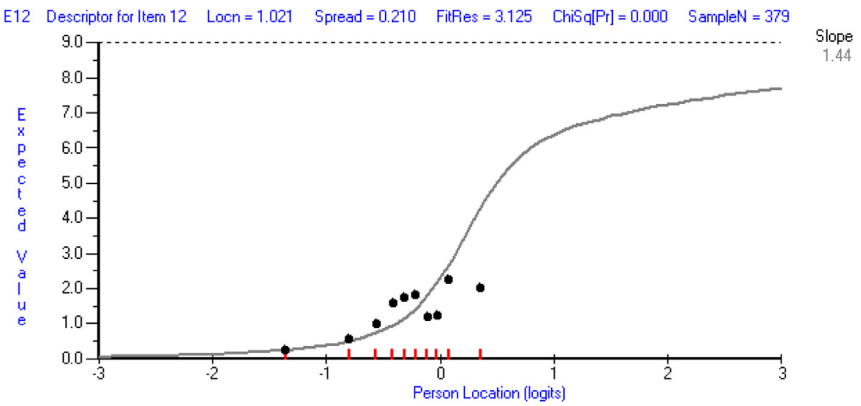

Fig. 1. ICC of item with poor fit: 12 (*How agitated do you feel*).

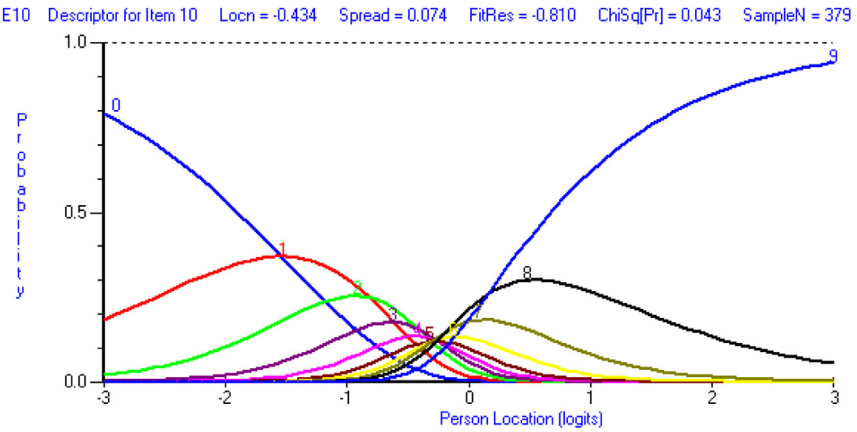

Fig. 2. 10 response Category Characteristic Curve (CCC) for item 10 (*How easily could you train some more*).

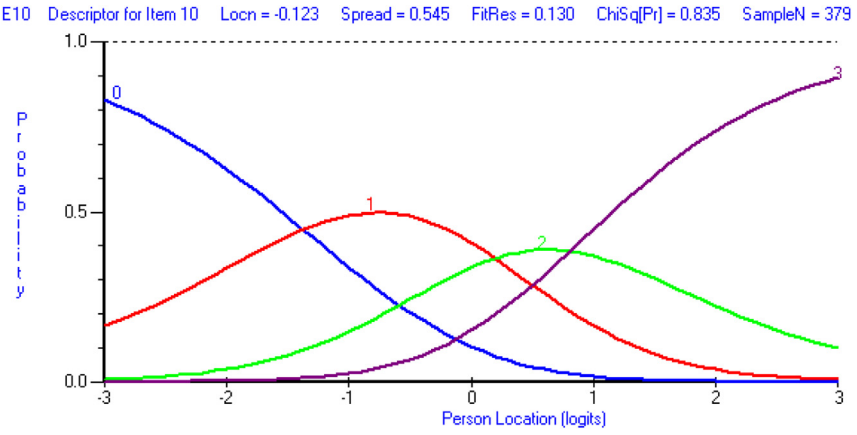

Fig. 3. 4 response Category Characteristic Curve (CCC) for item 10 (*How easily could you train some more*).

can you compete in your sport), and it was dropped because the item was determined to focus on "sports teams," an attribute that may not apply to all people intended to use this scale (i.e., exercise only) and as a level of talent and ability. Additionally, no theoretical basis for the DIF could be justified.

Table 2
Locations (in increasing order) and Chi Square fit statistic for 17 item HPHEES.

Item	Location	ChiSq	Prob[a]	Item content
E14	−1.156	31.933	0.002	How recovered do you feel
E21	−1.133	26.021	0.028	How energetic do you feel
E18	−0.726	15.140	1.000	How refreshed to you feel
E5	−0.663	3.771	1.000	How easily can you run
E1	−0.587	14.619	1.000	How tired do your legs and/or arms feel
E8	−0.552	15.342	1.000	How physically drained to you feel
E25	−0.516	9.576	1.000	How easily can you replicate your last game, event, or competition
E10	−0.146	9.541	1.000	How easily could you train some more
E22	−0.130	13.332	1.000	How weak do your legs and/or arms feel
E15	0.011	27.413	0.016	How easily can you complete sub-maximal training associated with your sport
E24	0.073	9.244	1.000	How much do your muscles ache
E23	0.456	12.860	1.000	How mentally sharp do you feel
E4	0.752	20.391	0.219	How relaxed do you feel
E6	0.892	19.005	0.352	How uncomfortable do you feel
E13	1.073	15.321	1.000	How mentally drained do you feel
E3	1.085	6.370	1.000	How easily can you walk
E2	1.268	24.758	0.045	How mentally cloudy do you feel

[a] Bonferroni adjusted p-values < 0.05.

At this stage a 17-item scale remained. The 17-item scale had a PSI of 0.893, item 14 (*How recovered do you feel*) displayed poor fit to the model, three paired correlations remained, item 3 showed DIF for perceived level of fitness, item thresholds were operating as required, and item locations (the lowest locations indicate the easiest items to agree with, meaning that even participants with relatively low levels of exhaustion are likely to agree with these. Conversely, the highest locations indicate the items that require relatively high levels of exhaustion to agree with) ranged (in logits, the Rasch unit of measurement) from −1.156 to 1.268.

Though assessment of various fit indices suggested a reasonable unidimensional scale, it was decided to attempt to drop a few more items. A review of item difficulty location indicated clusters of unevenly spaced items (Table 2). From these clusters, items were examined for possible removal. This resulted in 3 items being removed: items 6 (*How uncomfortable do you feel*) and 15 (*How easily can you complete sub-maximal training associated with your sport*) were removed because of wording concerns and proximity of item-difficulty rating to other items; item 1 (*How tired do your legs and/or arms feel*) was removed because of proximity to other item-difficulty rating for a cluster of items and because there was substantive overlap with item 22 (*How weak do your legs and/or arms feel*). Item 3 (*How easily can you walk*) was retained at this stage but was monitored due to its DIF.

The final scale contained 14 items (Table 3). The PSI was 0.881. Bonferroni adjusted fit (with 4 categories) for each item indicated only 1 potential misfitting item (item 2). However, as this was at the

extreme end of the item-difficulty scale, it was decided to retain it as its performance could be monitored in future research. The ICC for this item is shown in Fig. 4. Item 3 still showed DIF for perceived level of fitness, but again, was retained due to its high item-difficulty, and wording, which provides a good indication of physical exhaustion. The person-item distributions (Fig. 5) indicate a good scale spanning the range of potential person-responses. There were two paired residual correlations (items 13 and 2; items 13 and 23). It was decided to retain these items as their performance could be monitored in future research.

Correlation with existing scales

When comparing the final set of 14 items against the response scales of 10 choices (the original survey instrument) or 4 collapsed choices (0-11-222-3333), the logits for each 1 parameter logistic-item response are very highly correlated ($r = .965$, $n = 379$, $p < .001$) with 93% shared variance. The logit scores from each model (10 and 4 response choices) are plotted in Fig. 6. Fig. 7 shows the logit scores from the 4 response category model compared to the aggregated score obtain from averaging the raw scores from the original 10 response categories (with appropriate reverse coding).

To assess the construct validity of the new scale, the logit scores from the 4 response category model were compared with an existing scale (the 13-item fatigue subscale of the Lee Fatigue Scale). Though comparable (yet slightly higher) results were obtained when using estimated scores based on model parameters for these individuals, the more conservative estimates obtained without these 3 individuals are reported here. The model estimated person logits from each scale were moderately correlated with $r = .666$ ($n = 366$, $p < .001$) and resulted in 44% shared variance. Fig. 8 shows the scatterplot of the two scales for visual comparison (with the 3 extreme responders' values estimated based on model parameters).

Discussion

The purpose of this study was to describe the development and psychometric analysis of a new scale aimed at measuring acute onset exhaustion in athletes. Identifying individuals who may develop the accumulative influences of acute exhaustion which may lead to or be related to disease or pathology recognised in chronic exhaustion is important. The scale was designed to be simple, appropriate for its intended use, utilised rapidly after an event or exercise session and include a clear and interpretable scoring system in order to increase compliance (Connelly, 2011; Kelley, Clark, Brown, & Sitzia, 2003).

Table 3
Locations (in increasing order) and Chi Square fit statistic for final 14 item HPHEES.

Item	Location	ChiSq	Prob[a]	Item content
E14	−1.156	31.933	0.002	How recovered do you feel
E21	−1.133	26.021	0.030	How energetic do you feel
E18	−0.726	15.140	1.000	How refreshed to you feel
E5	−0.663	3.771	1.000	How easily can you run
E8	−0.552	15.342	1.000	How physically drained to you feel
E25	−0.516	9.576	1.000	How easily can you replicate your last game, event, or competition
E10	−0.146	9.541	1.000	How easily could you train some more
E22	−0.130	13.332	1.000	How weak do your legs and/or arms feel
E24	0.073	9.244	1.000	How much do your muscles ache
E23	0.456	12.860	1.000	How mentally sharp do you feel
E4	0.752	20.391	0.220	How relaxed do you feel
E13	1.073	15.321	1.000	How mentally drained do you feel
E3	1.085	6.370	1.000	How easily can you walk
E2	1.268	24.758	0.046	How mentally cloudy do you feel

[a] Bonferroni adjusted p-values < 0.05.

Through the use of the PRM for psychometric analysis we achieved three aims of validating the Hecimovich–Peiffer–Harbough Exercise Exhaustion Scale (HPHEES). Furthermore, as there have been only a small amount of studies in the field of exercise science using the Rasch model for psychometric analysis our findings are novel.

The first aim was to establish internal consistency and reliability of the HPHEES. Our results demonstrate, with amendments, the HPHEES provided internally consistent and reliable measures of exhaustion for participants. The final scale includes 14 items with good fit and adds to the lack of available scales to assess acute onset exhaustion and fatigue in exercise and sport. This is important because entities of exercise-induced exhaustion and fatigue have been an area of interest for many physiologists and psychologists and represent psychological entities, which will sooner or later introduce changes in behaviour (Ament & Verkerke, 2009). The final scale addresses both entities as items are pertinent to affect on functioning and associated symptoms.

After initial analyses, an exploratory series of tasks was used to demonstrate the presence of necessary elements for a measurement tool, namely that the scale did not have redundant items and the (number of) categories worked correctly for all items (e.g., each category for each item would be the modal response for some level of fatigue/exhaustion). One of the tasks involved collapsing the 10 original categories to 4 aggregated categories. While the post-hoc collapsing of categories used here was exploratory in nature, the rational for collapsing is based on subtle variations among participants self-assessment of exhaustion, and the nature of a frequency scale. Thus, future research should examine if this proposed 10-to-4 category instrumental analysis works comparably for other populations. For example, items pertaining to muscle tightness (item 19), muscle cramping (item 16), ease of daily functions (item 7), levels of concentration (item 9), and feelings of agitation (item 12), were the least well-fitting and removed. Item 19 was eliminated due to its poor fit and resemblance to other items, which performed better, as a description of muscle sensation (tightness, aches, tired). Item 7 was initially developed as a result of the athlete and expert panel feedback but may have been too broad and non-specific and therefore eliminated. Further analysis, which included examination of paired residual correlations, review of item wording and disordered thresholds, resulted in additional item removal. These included items referred to ease of completing maximal and sub-maximal training (i.e., item 20, *How easily can you complete maximal training associated with your sport*), respectively; ease at competing in the participants' sport; sensation of being knackered/bushed (item 11, *How knackered/bushed are you*), and; an item which referred to level of talent (item 17, *How easily can you compete in your sport*).

The development and wording of these items highlights the need to carefully consider how each item is worded and whether the meaning mirrors the content of a particular area of interest. The wording of the items 15 and 20, which referred to sub-maximal and maximal training, may be difficult for the participant to respond to due to a possible lack of knowledge concerning these levels. Developing exhaustion items for athletes or those in exercise and physical training (i.e., military) is challenging due to the overlying relationship between exhaustion and other similar constructs such as burnout. This is apparent with Raedeke's (1997) definition of athlete burnout as a syndrome being characterized by: a) emotional and physical exhaustion, b) sport devaluation, and c) a reduced sense of accomplishment (e.g. Cresswell & Eklund, 2006; Raedeke, Lunney, & Venables, 2002). In Raedeke's and Smith's Athlete Burnout Questionnaire (ABQ) (Raedeke & Smith, 2001), which contains 15 items designed to measure: 1) reduced sense of accomplishment, 2) devaluation, and 3) emotional/physical

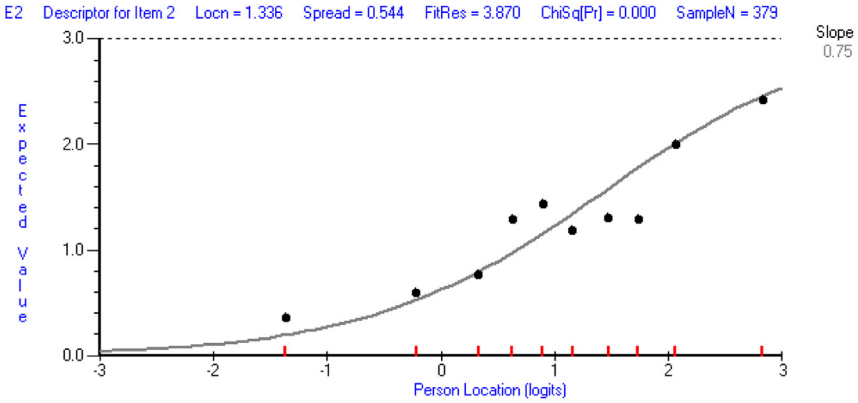

Fig. 4. ICC of item with poor fit: 2 (*How mentally cloudy do you feel*).

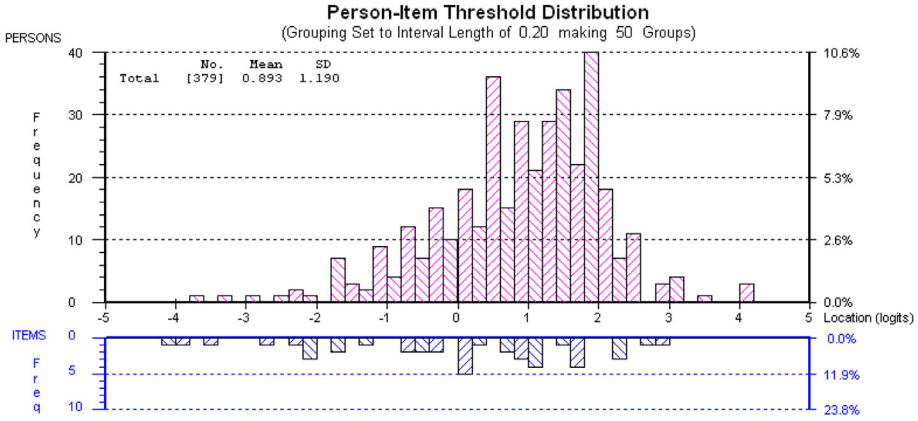

Fig. 5. Person-Item Threshold Distribution for final 14 item HPHEES.

exhaustion, it encompasses broader components than only exhaustion. With the HPHEES the items focus on associated symptoms, which are physical, and affect of function, which may have a psychological perspective depending on the respondent.

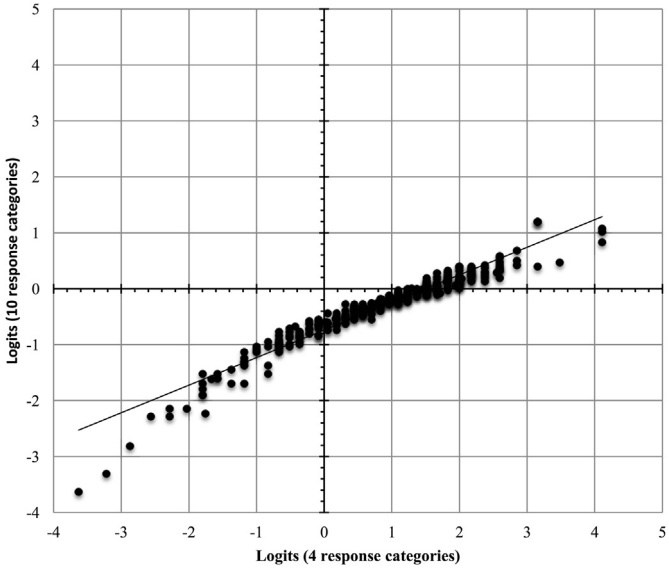

Fig. 6. Person logit scores for the final set of 14 items using response categories with 10 choices compared to 4 collapsed categories.

Thus, if those items in the HPHEES consistently demonstrate high scores within a specific group or individual respondent, we suggest a corresponding scale, such as the ABQ, be utilized.

For exhaustion or fatigue, which can be distinguished as primarily physiological (e.g., muscle strength, and exercise tolerance) or self-report (i.e., patients' perceptions of fatigue and its consequences) (Lai et al., 2011), careful consideration on content for the HPHEES was taken, for example, an extensive literature review, incorporating feedback from athletes and an expert panel, in the formulation of items. Although content validity was not tested, this feedback provides the scale with important face validity which is simply the judgements that the athletes and experts made (Brown, Bull, & Pendlebury, 1997) and viewed as a measure of credibility (Matsell, Wolfish, & Hsu, 1991).

The high PSI provided statistical evidence of reliability. The HPHEES could benefit from items dealing with moderate and high levels of exhaustion and its associated symptoms. This might involve developing items that focus on associated symptoms pertaining to the muscular system (muscular aches and weakness), and motivation (i.e., 10, *How easily can you train some more*) as these are located in the mid-range and items which pertain to ease of walking (i.e., *How easily can you walk*) and mental sensation (i.e., *How mentally cloudy do you feel*) as these are the most difficult to agree with and may help guide the development of additional items which can measure this level of acute onset exhaustion.

The second aim was to verify whether each scale had similar psychometric properties across different groups of participants. There was evidence of differential item functioning (DIF) for item 3

with perceived level of fitness (*How easily can you walk*). Overall this item is valuable as it showed high item-difficulty and its wording spread across motivation and associated symptoms (i.e., inability to walk easily can be a sign of physical ability) and therefore the decision to retain it. However, using person's perceived fitness level does pose risks in that people may create positive perceptions of their abilities by distorting reality in a direction that enhances their self-esteem and self-efficacy, thereby promoting an optimistic view of their future (Dunning & Story, 1991). Because of a tendency to create positive illusions, self-perceptions of fitness levels may also be exaggerated in a positive direction (Asendorpf & Ostendorf, 1998; John & Robins, 1994; Robins & Beer, 2001; Taylor, 1989). However, Germain and Hausenblas (2006) examined the moderating influence of gender, age, and perceived fitness measure for the perceived-actual fitness relationship and found a medium effect size indicating people had accurate perceptions of their actual physical fitness. Additionally, Germain and Hausenblas (2006) found that standardized perceived fitness measures had a significantly larger effect size than author-developed measures. Lamb (1992) stated that there is no simpler method of assessing fitness than asking people about it, and many researchers did that through the use of an author-developed, non-validated survey. For example, Young (1985) asked "How would you rate your present physical fitness level?" and provided five options (very good to very poor). Some researchers qualified items by asking participants to rate their fitness relative to someone their own age (Lamb, 1992; Marsh, 1993; Shephard & Bouchard, 1995). While Lamb (1992) is correct in arguing that these methods are simple, it appears they are too simple, and that they do not address the multidimensionality of physical fitness. The result is a reduced (albeit still significant) effect size between perceived and actual fitness. Whether the reason for DIF in this item was due to its spanning motivation and physical ability or accuracy of a person's perception of fitness level would need to be assessed and in the future it is advised to monitor its performance and determine if the current result is a consistent anomaly.

The researchers' choice to collapse the 10 response scale to 4 categories was one of convenience in that it was desired to obtain a scale in which all of the response categories were operating comparably across all items and each response category would be the modal response for each item for some level of exhaustion for some level of difficulty. However, it is acknowledged that many researchers are inclined to use the raw scores (for ease of calculation, lack of familiarity with latent variable modelling, or simply to obtain a quick measure). The similarity of the person estimates obtained from the two response scales (4 vs. 10) were comparable with 93% shared variance. In the scatterplot (Fig. 7) comparing the logit scores for each level, the only notable deviation from the linear transformation between scales was observed for the very lowest of exhaustion levels. Furthermore, the results indicate that the use of the averaged raw scores form the original 10 categories produces a nearly linear transformed estimate of the model estimated logits from the 4 category scale. As such, when targeting a sample over a specific range of exhaustion levels, the averaged 10-response scores appear to be a viable proxy for the model-based estimates. Of course, it is recommended that researchers employ a more sophisticated latent variable model to accommodate the measurement error in such scales when such accuracy or specificity is required. However, under reasonable circumstances, the averaged raw score will most likely suffice.

The third aim of the study was to investigate the construct validity of the new scale using a well-established scale. The results obtained from the new scale shared a substantial amount of shared variance ($R^2 = 44\%$) with the LFS 13-item fatigue subscale. As the scales are comparable, but not necessarily measuring the exact same unidimensional construct, these results provide a reasonable indication that the new scale is indeed measuring a construct relating to exhaustion.

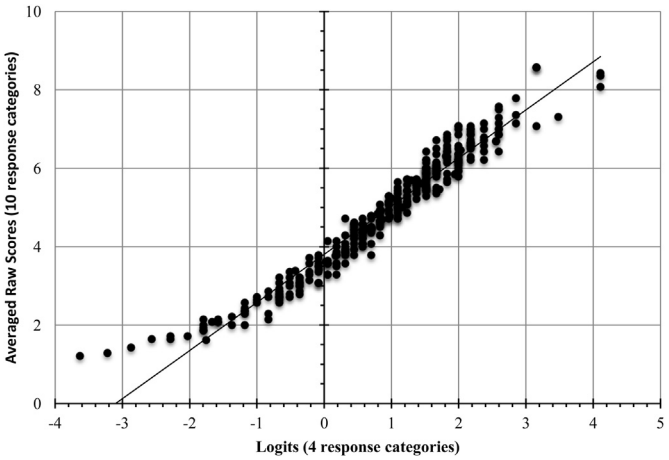

Fig. 7. Person logit scores for the final set of 14 items using 4 response categories compared with the averaged raw scores for the same 14 items using the original 10 response categories.

Limitations include the use of tri-athletes and cyclist which did not allow for a balanced proportion of different kinds of sports and therefore participants. Future studies need to consider a wider range of sports, activities or occupations (i.e., military personnel, fire fighters). In reference to various components of validity, the Rasch analysis does provide evidence of internal consistency which is an aspect of construct validity. Face validity was addressed by the assessment of a panel of experts while construct validity was assessed by the scores on the new scale being correlated with scores from the same participants on an existing valid and reliable scale. Although IRT researchers have not explicitly put forward a way to check the consequential aspect of validity, a type of validity evidence that addresses the intended and unintended consequences of test interpretation and use (Messick, 1989, 1995), issues like item bias and examination of differential item functioning (DIF) or a close examination of the person-item map reveals information on the basis of which decisions for action are taken and can provide helpful evidence to decide about the consequential aspect of construct validity of a test. Future studies may want to utilise

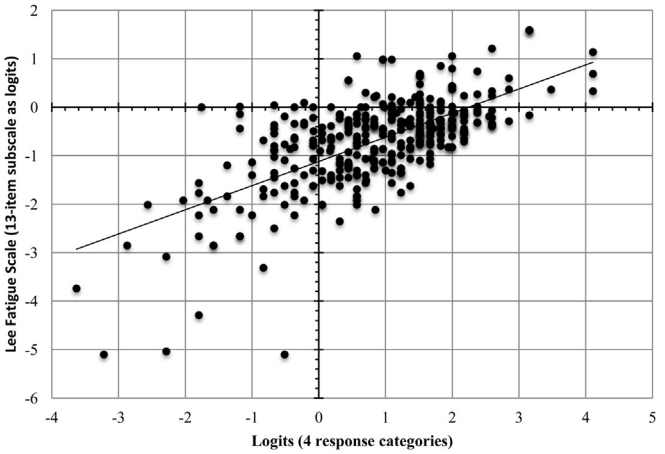

Fig. 8. Person logit scores for the final set of 14 items using 4 response categories compared with the person logit scores for the 13-item fatigue subscale of the Lee Fatigue Scale.

exploratory and confirmatory analysis which can be thought of as two ends of a spectrum.

Future studies may want to consider using the new scale to measure the correlation between perceived acute onset exhaustion and physiological and biochemical markers of exhaustion such as maximal incremental cycle ergometer test (W_{max}) with continuous ventilatory measurements and blood lactate values, basal blood parameter tests, hormones, neuro-endocrine stress test, combined anterior pituitary test (Rietjens et al., 2005) and others. Also, measuring the relationship in rate of perceived exertion (RPE) and acute onset exhaustion (with the HPHEES), for example with the Borg CR-10 scale (RPE scale), which evaluates PRE in exercise testing, training, and rehabilitation and has been validated against objective markers of exercise intensity (Borg, 1985; Noble, Borg, Jacobs, Ceci, & Kaiser, 1983), may assist in identifying individuals' variations in training levels over time. The RPE scale measures overall perceived exertion during an exercise bout and integrates signals from the peripheral working muscle as well as the central cardiovascular, respiratory, and nervous systems (Borg, 1982). These signals also play a key role in perceived levels of exhaustion (Rietjens et al., 2005).

In summary, the HPHEES appears to be a valuable tool for the assessment of exercise-induced acute onset exhaustion. In addition to its valid theoretical structure and sound psychometric properties, the scale has advantages over other exhaustion or fatigue scales as it is not disease-specific. The utility of the HPHEES is further increased by its limited amount of items (14) and the use of a single response format and the brevity of most items. As a result, the HPHEES may be easier to complete and less burdensome than other similar scales, which is often an advantage when assessing athletes; however, no empirical data are available to support this claim. Nevertheless, the HPHEES may be an attractive option for researchers, clinicians, and coaches seeking to measure the levels of exhaustion in individuals.

Acknowledgements

I would like to thank Associate Professor Irene Styles, and Dr Josh McGrane, Pearson Psychometric Laboratory, The Graduate School of Education, The University of Western Australia, who assisted with using RUMM2030 and the Rasch model.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.psychsport.2014.06.003.

References

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43*, 561–573.
Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch unidimensional measurement model (RUMM2030) computer program*. Perth, Western Australia: RUMM Laboratory.
Andrich, D., & Styles, I. (2004). *Final report on the psychometric analysis of the Early Development Instrument (EDI) using the Rasch Model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI)*. Perth, WA: Murdoch University.
Ament, W., & Verkerke, G. J. (2009). Exercise and fatigue. *Sports Medicine, 39*(5), 389–422.
Asendorpf, J. B., & Ostendorf, F. (1998). Is self-enhancement healthy? Conceptual, psychometric, and empirical analysis. *Journal of Personality and Social Psychology, 74*, 955–966.
Bartlett, F. C. (1953). Physiological criteria of fatigue. In W. F. Floyd, & A. T. Welford (Eds.), *Fatigue* (pp. 1–15). London: Lewis.
Bond, T. G., & Fox, C. M. (Eds.). (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
Borg, G. (1982). Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise, 14*(5), 377–381.
Borg, G. (1985). *An introduction to Borg's RPE-scale*. Ithaca, NY: Movement Publications.
Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.
Connelly, L. M. (2011). Surveys, surveys, and more surveys. *MedSurg Nursing, 20*(2), 61.
Cox, R. H. (2007). *Sport psychology: Concepts and applications*. New York, NY: McGraw-Hill.
Cresswell, S. L., & Eklund, R. C. (2006). Changes in athlete burnout over a thirty-week "rugby year". *Journal of Science and Medicine in Sport, 1*(2), 125–134.
Dunning, D., & Story, A. L. (1991). Depression, realism, and the overconfidence effect: are the sadder wiser when predicting future actions and events? *Journal of Personality and Social Psychology, 61*, 521–532.
Embretson, S. E., & Reise, S. P. (Eds.). (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
Fawzy, F. I., Cousins, N., Fawzy, N. W., Kemeny, M. E., Elashoff, R., & Morton, D. (1990). A structured psychiatric intervention for cancer patients. *Archives of General Psychiatry, 47*, 720–725.
Galambos, S. A., Terry, P. C., Moyle, G. M., & Locke, S. A. (2005). Psychological predicators of injury among elite athletes. *British Journal of Sports Medicine, 39*, 351–354.
Gay, C. L., Lee, K. A., & Lee, S. Y. (2004). Sleep patterns and fatigue in new mothers and fathers. *Biological Research in Nursing, 5*(4), 311–318.
Germain, J. L., & Hausenblas, H. A. (2006). The relationship between perceived and actual physical fitness: a meta-analysis. *Journal of Applied Sport Psychology, 18*(4), 283–296. http://dx.doi.org/10.1080/10413200600944066.
Goodger, K., Wolfenden, L., & Lavallee, D. (2007). Symptoms and consequences associated with three dimensions of burnout in junior tennis players. *International Journal of Sports Psychology, 38*, 342–364.
Greenberg, D. B., Sawicka, J., Eisenthal, S., & Ross, D. (1992). Fatigue syndrome due to localized radiation. *Journal of Pain and Symptom Management, 7*(1), 38–45.
Gustafsson, H., Hassmen, P., Kentta, G., & Johansson, M. (2008). A qualitative analysis of burnout in elite Swedish athletes. *Psychology of Sport and Exercise, 9*, 800–816.
Harquist, C., Bruce, M., & Gustavsson, J. (2009). Using the Rasch model in nursing research: an introduction and illustrative example. *International Journal of Nursing Studies, 46*, 380–393.
Heywood, N., Sabado, S., & De Leon, B. (2012). Reduction of fear by intense aerobic exercise approaching physical exhaustion. *Psychology, 3*(8), 613–615.
John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology, 66*, 206–219.
Kant, I. J., Bultmann, U., Schroer, K. A., Beurskens, A. J. H. M., van Amelsvoort, L. G. P. M., & Swaen, G. M. H. (2003). An emidemiological approach to study fatigue in the working population: the Maastricht Cohort Study. *Occupational and Environmental Medicine, 60*(Suppl. I), i32–i39.
Kelley, K., Clark, B., Brown, V., & Sitzia. (2003). Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care, 15*(3), 261–266.
Kellman, M. (2002). Underrecovery and overtraining: different concepts—similar impact? In M. Kellman (Ed.), *Enhancing recovery: Preventing underperformance in athletes* (pp. 103–118) Champaign, IL: Human Kinetics.
Kenttä, G., & Hassmen, P. (1998). Overtraining and recovery: a conceptual model. *Sports Medicine, 26*, 1–16.
Kobashi-Schoot, J. A. M., Hanewald, G. J. F. P., Van Dam, F. S. A. M., & Bruning, P. F. (1985). Assessment of malaise in cancer treated with radiotherapy. *Cancer Nursing, 8*, 306–314.
Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology, 46*, 1121–1123.
Lai, J. S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., et al. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: a PROMIS Fatigue Item Bank Example. *Archives of Physical Medicine and Rehabilitation, 92*(10 0), S20–S27. http://dx.doi.org/10.1016/j.apmr.2010.08.033.
Lamb, K. L. (1992). Correlates of self-perceived fitness. *Perceptual & Motor Skills, 74*, 907–914.
Lee, K. A., Hicks, G., & Nino-Murcia, G. (1991). Validity and reliability of a scale to assess fatigue. *Psychiatry Research, 36*(3), 291–298.
Lee, K. A., Lentz, M. J., Taylor, D. L., Mitchell, E. S., & Woods, N. F. (1994). Fatigue as a response to environmental demands in women's lives. *Journal of Nursing Scholarship, 26*, 149–154.
Lee, K. A., Portillo, C. J., & Miramontes, H. (1999). The fatigue experience for women with human immunodeficiency virus. *Journal of Obstetrics, Gynecology, and Neonatal Nursing, 28*(2), 193–200.
Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.
Linacre, J. M. (2000). Item discrimination and infit mean-squares. *Rasch Measurement Transactions, 14*, 743.
Marsh, H. S. (1993). The multidimensional structure of physical fitness: invariance over gender and age. *Research Quarterly for Exercise and Sport, 64*, 256–273.
Matsell, D. G., Wolfish, N. M., & Hsu, E. (). Reliability and validity of the objective structured clinical examination in paediatrics. *Medical Education, 25*, 293–299.
Meijman, T. F., & Schaufeli, W. B. (1996). Psychische vermoeidheid en arbeid. Ontwikkelingen in de A&Opsychologie [Mental fatigue and work. Developments in work and organizational psychology]. *De Psycholoog, 31*, 236–241.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York: American Council on Education.
Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741–749.
Miaskowski, C., Paul, S. M., Cooper, B. A., Lee, K., Dodd, M., & West, C. (2008). Trajectories of fatigue in men with prostate cancer before, during, and after radiation therapy. *Journal of Pain and Symptom Management, 35*(6), 632–643.
Noakes, T. D. (2000). Physiological models to understand exercise fatigue and the adaptations that predict or enhance athletic performance. *Scandinavian Journal of Medicine and Science in Sports, 10*, 123–145.
Noble, B. J., Borg, G., Jacobs, I., Ceci, R., & Kaiser, P. (1983). A category-ratio perceived exertion scale: relationship to blood and muscle lactate and heart rate. *Medicine and Science in Sports and Exercise, 15*, 523–528.
Olson, K. (2007). A new way of thinking about fatigue: a reconceptualization. *Oncology Nursing Forum, 34*, 93–99.
Piper, B. (1989). Fatigue: current bases for practice. In S. Funk, E. Tomquist, M. Champagne, L. Copp, & R. Weise (Eds.), *Key aspects of comfort* (pp. 187–189). New York: Springer.
Raedeke, T. (1997). Is athlete burnout more than just stress? A sport commitment perspective. *Journal of Sport and Exercise Psychology, 19*(4), 396–417.
Raedeke, T. D., Lunney, K., & Venables, K. (2002). Understanding athlete burnout: coach perspectives. *Journal of Sport Behavior, 25*, 181–206.
Raedeke, T. D., & Smith, A. L. (2001). Development and preliminary validation of an athlete burnout measure. *Journal of Sport & Exercise Psychology, 23*, 281–306.
Ramp, M., Khan, F., Misajon, R. A., & Pallant, J. F. (2009). Rasch analysis of the Multiple Sclerosis Impact Scale (MSIS-29). *Health and Quality of Life Outcomes, 7*, 58. Available from: URL http://www.hqlo.com. Last accessed 08.07.12.
Expanded edition (1980). In Rasch, G. (Ed.), *Probabilistic models for some intelligence and attainment tests*. Chicago, Illinois: University of Chicago Press.
Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133–144.
Rietjens, G. J., Kuipers, H., Adam, J. J., Saris, W. H., van Breda, E., van Hamont, D., et al. (2005). Physiological, biochemical and psychological markers of strenuous training-induced fatigue. *International Journal of Sports Medicine, 26*(1), 16–26.
Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80*, 340–352.
Sharp, M., & Wilks, D. (2002). Fatigue. *British Medical Journal, 325*(7362), 480–483.
Shephard, R. J., & Bouchard, C. (1995). Relationship between perceptions of physical activity and health-related fitness. *Journal of Sports Medicine and Physical Fitness, 35*(3), 149–158.
Smets, E. M. A., Garssen, B., Bonke, B., & De Haes, J. C. J. M. (1995). The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *Journal of Psychometric Research, 39*(5), 315–325.
Strauss, B., Büsch, D., & Tenenbaum, G. (2012). Rasch modeling in sports. In G. Tenenbaum, R. Eklund, & A. Kamata (Eds.), *Handbook of measurement in sports* (pp. 75–80). New York: Human Kinetics.
Taylor, S. E. (1989). *Positive illusions*. New York: Basic Books.
Tiesinga, L. J., Dassen, T. W. N., & Halfens, R. J. G. (1998). DUFS and DEFS: development, reliability and validity of Dutch Fatigue Scale and the Dutch Exertion Fatigue Scale. *International Journal of Nursing Studies, 34*, 115–123.
Vertommen, H., & Leyssen, J. (1988). Vermoeidheid: Van onhanteerbaar symptoom tot diagnostisch waardevolle gemoedstoestand [Fatigue: from an unmanageable symptom to a diagnostically valuable state of mind]. *Tijdschrift voor de Klinische Psychologie, 18*, 35–59.
Weinberg, R. S., & Gould, D. (2007). *Foundations of sport and exercise science* (4th ed.). Champaign, IL: Human Kinetics.
Wessely, S., & Powell, R. (1989). Fatigue syndromes: a comparison of chronic 'post viral' fatigue with neuromuscular and affective disorders. *Journal of Neurology, Neurosurgery, and Psychiatry, 52*, 940–948.
Williams, J. M. (2006). *Applied sport psychology: Personal growth to peak performance*. New York: McGraw-Hill.
Winwood, P. C., Winefield, A. H., Dawson, D., & Lushington, K. (2005). Development and validation of a scale to measure work-related fatigue and recovery: the Occupational Fatigue Exhaustion Recovery Scale (OFER). *Journal of Occupational and Environmental Medicine, 47*, 594–606.
Wood, & Zhu. (2006). *Measurement theory and practice in kinesiology*. Champaign, IL: Human Kinetics.
Yang, C. M., & Wu, C. H. (2005). The Situational Fatigue Scale: a different approach to measuring fatigue. *Quality of Life Research, 14*, 1357–1362.
Young, M. L. (1985). Estimation of fitness and physical ability, physical performance, and self-concept among adolescent females. *Journal of Sports Medicine and Physical Fitness, 25*, 144–150.