Psychology, 2015, 6, 2161-2180
Published Online December 2015 in SciRes. http://www.scirp.org/journal/psych
http://dx.doi.org/10.4236/psych.2015.616213

Rasch Rating Scale Analysis of the Arabic Version of the Physical Activity Self-Efficacy Scale for Adolescents: A Social Cognitive Perspective

Sabry M. Abd-El-Fattah[1,2]

[1]Minia University, Minia, Egypt
[2]Sulatn Qaboos University, Muscat, Oman
Email: sabryrahma@hotmail.com

Received 25 November 2015; accepted 28 December 2015; published 31 December 2015

Copyright © 2015 by author and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

This study uses the Rasch rating scale analysis within the framework of the social cognitive theory to examine four psychometric properties of the Arabic version of the Physical Activity Self-efficacy Scale for Adolescents (PASESA-Av): 1) item technical quality to evidence content validity; 2) unidimensionality structure to evidence structural validity; 3) rating scale functioning to evidence substantive validity; and 4) person and item separation and reliability. The analyses showed that the PASESA-Av items represented the physical activity self-efficacy trait adequately. The PASESA-Av demonstrated good fit to the Rasch model and formed a unidimensional structure. The PASESA-Av response categories advanced monotonically. The PASESA-Av demonstrated satisfactory person and item separation and reliability.

Keywords

Rasch Model, Self-Efficacy, Physical Activity, Adolescents

1. Introduction

Physical activity rates for adolescents are insufficient for health benefits and consequently the inactivity-related diseases such as obesity are on the rise in many countries (World Health Organization, 2014). For example, the Omani guidelines suggest that children and adolescents accumulate 90 minutes of moderate to vigorous physical activity per day. However, over 90% of Omani adolescents aged 6 to 20 years are not meeting these guidelines. Furthermore, the obesity

How to cite this paper: Abd-El-Fattah, S. M. (2015). Rasch Rating Scale Analysis of the Arabic Version of the Physical Activity Self-Efficacy Scale for Adolescents: A Social Cognitive Perspective. *Psychology, 6,* 2161-2180.
http://dx.doi.org/10.4236/psych.2015.616213

rates among Omani adolescents are increasing at epidemic proportions. Results from the 2010-2014 national survey suggest that close to 28% of children and adolescents ages 3 - 20 years are overweight according to their heights and weights (Ministry of Health, 2014). To better understand the patterns of activity and inactivity among adolescents, numerous psychological factors must be considered. One of these factors is physical activity self-efficacy.

The concept of self-efficacy is grounded in Bandura's social cognitive theory (Bandura, 1986). Self-efficacy describes "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986: p. 391). The basic principle behind self-efficacy is that individuals are more likely to engage in activities for which they have high self-efficacy and less likely to engage in those they do not (Stajkovic & Luthans, 1998), thus, self-efficacy functions as a self-fulfilling prophecy. Bandura (1997) proposed that self-efficacy beliefs are shaped by the cognitive processing and integration of four main sources of information: 1) performance attainments and failures—what we try to do and how well we succeed or not; 2) vicarious performances—what we see other people do; 3) verbal persuasion—what people tell us about what we are able or not able to do; and 4) imaginal performances—what we imagine ourselves doing and how well or poorly we imagine ourselves doing it.

Self-efficacy beliefs are presumed to have actual task ability as an important underlying determinant. In other words, someone who typically does well on a task knows that he or she does well and shows this knowledge in his or her self-efficacy ratings. However, beliefs in one's self-efficacy are not just based on a simple knowledge of one's level of ability. Those self-beliefs go beyond actual capability, being "instrumental in determining what individuals do with the knowledge and skills they have" (Pajares & Miller, 1995: p. 190). According to Bandura (1997), self-efficacy has "effects on thought, affect, action, and motivation" (p. 46). Thus, someone high in self-efficacy might do better because that person approached a task with a different mindset than does someone low in self-efficacy, even though both people might beat the same level of ability.

1.1. Measuring Self-Efficacy

Bandura (1997: p. 42) maintained that self-efficacy "is not a contextless global disposition [to be] assayed by an omnibus test." Instead, proper self-efficacy measures "must be tailored to domains of functioning". Such domains can refer to any activity, or class of activities, where individuals can differ in their success rates and, more important, in their beliefs about their success rates. The domain might be related to good performance in a course in mathematics, biology, or language. The domain might concern doing well at tasks involving physical strength, eye-hand coordination, or memory. The domain could even represent maintaining successful relationships, being a good parent, or sticking to a diet. Within any one domain of performance, self-efficacy beliefs can be measured with respect to diverse arrays of accomplishments differing in breadth. Consider the domain of physical fitness. At a very narrow level, one could measure self-efficacy for performing a specific physical exercise. At a broader level, self-efficacy could be assessed with regard to passing Standing Stork Tests (tests of an athlete's ability to maintain a state of equilibrium in a static position). Even broader yet would be to evaluate beliefs about one's physical fitness aptitude (Warner, Parschau, Schwarzer, Wolff, Wurm, & Schuez, 2014).

However, there are even some measures of self-efficacy that are so broad in scope that they do not refer to any specific performance domain. Instead, such global measures refer to general competence and life coping skills, with items related to accomplishing goals in general and performing effectively on different tasks (e.g., Bandura, 2005). Bandura (1997) has maintained that global self-efficacy measures "violate the basic assumption of the multidimensionalities of self-efficacy beliefs" (p. 48) and that "undifferentiated, contextless measures of personal efficacy have weak predictive value" (p. 49). Ideally, according to Bandura, a self-efficacy measure should match, in level of generality, the performance criterion of interest (Bandura, 2002, 2008). So if the criterion is, for example, an obtained score on a Standing Stork Test, then the self-efficacy measure should represent a person's beliefs about his or her performance on that narrow task. On the other hand, if the criterion is overall physical fitness, then the self-efficacy measure should be broader, referring perhaps to a person's expectations about his or her performance on a physical fitness aptitude test.

1.2. Physical Activity Self-Efficacy

One important domain to examine individual's self-efficacy beliefs is the domain of physical activity because much research has shown that self-efficacy is a critical antecedent to physical activity. High self-efficacy has been linked to better performance on physical activity tasks, expending more effort on physical activity tasks, and persevering when difficulties arise (Gao, Lee, Kosma, & Solmon, 2010; Gao, Lodewyk, & Zhang, 2009). For example, Gao et al. (2010) found that self-efficacy predicted 54% of the variance in physical activity among 207 middle school students in physical education classes. Gao, Lochbaum, & Podlog (2011) found that self-efficacy predicted 27% of the variance in physical activity among 194 middle school students in physical education classes when it was set as the sole predictor, and it predicted 28% of the variance in physical activity when it was set as a predictor along with a mastery-approach goal and mastery motivational climate variables.

1.3. The Physical Activity Self-Efficacy Scale for Adolescents

Bandura (2004) proposed that self-efficacy beliefs must be measured against gradations of challenges to successful performance. For example, in assessing personal efficacy to stick to an exercise routine, individuals judge their efficacy to get themselves to exercise regularly in the face of different obstacles, such as being under pressure, tired, depressed, or anxious. Wu, Robbins, & Hsieh (2011) developed the Physical Activity Self-Efficacy Scale for Adolescents (PASESA**)** using a sample of 105 boys and 101 girls ranging in age from 11 - 14 years and enrolled in physical education or gym classes at a public middle school in the United States. The 11-item PASESA was used to determine each participant's confidence in his or her ability to overcome specific barriers to physical activity (e.g., "*I'm sure that I can still do my exercise even if I am too busy*."). The scale items were preceded by the phrase "*Please show how true each statement is regarding how sure you are that you can exercise, be active, or do sports when you face.*" Participants responded to each item using a 4-point Likert scale: "Not at all true", "Not very true", "Sort of true", and "Very true". The Likert scale received the scores 1, 2, 3, and 4 respectively because all items were positive. Scores for each participant were determined by computing the mean of a participant's item scores. Higher scores corresponded to a higher level of self-efficacy to overcoming barriers to physical activity. The results of Wu et al.'s study revealed that all item correlations were within .29 to .55. The corrected item–total correlation coefficients ranged from .39 to .64. Confirmatory factor analysis showed a one-factor model achieved modest fit to the data after correlated errors on two pairs of items (Items 4 and 8; Items 2 and 10). The PASESA has satisfactory internal consistency with Cronbach's alpha coefficient of .86, and test-retest reliability of .61 (2 weeks apart).

A review of the related literature reveals that there are four studies that have examined the validity of the factorial structure of the original English version of the PASESA. For example, Simpson & Rustin (2012) reported that a confirmatory factor analysis of responses from a sample of 172 Canadian adolescents aged (12 - 14 years) demonstrated that a single factor model fitted the data adequately. Gui & Smith (2012) reported that a confirmatory factor analysis of responses from a sample of 140 British adolescents aged (13 - 14 years) demonstrated that a single factor model fitted the data adequately after correlating the error terms associated with Item 4 and Item 8. Eaton & Crcook (2013) found that an exploratory factor analysis of responses from a sample of 133 Scottish adolescents aged between 11 and 13 years retained a single factor that accounted for 59% of the total variance extracted with item loadings ranging from .54 to .72. Peterson & Shavelson (2014) reported that a confirmatory factor analysis of responses from a sample of 155 Singaporean adolescents aged (12 - 14 years) demonstrated that a single factor model fitted the data adequately after correlating the error terms associated with Item 2 and Item 8. Furthermore, a study by Mohammad-Hassun & Ahmed-Kali (2014) has examined the validity of the factorial structure of the translated version of the PASESA within a Pakistani context. A confirmatory factor analysis of responses from 190 adolescents aged between 11 - 13 years in the western territory of Pakistan showed that a single-factor structure of the Pakistani (Urdu) version of the PASESA achieved a modest fit to the data. They concluded that "The structure of the PASESA should be examined with more meticulous procedures that can explore its items functions" (p. 16).

1.4. Rasch Rating Scale Model

The Rasch model is one of the Item Response Theory (IRT) models that have been used to analyze ordinal data in order to provide linear measures by the use of logarithmic transformation procedures. The Rasch analysis converts the raw item scores into equal-interval measures using a logarithmic transformation of the odds probabilities of responses. The logarithmic transformation simultaneously results in an estimation of item location (difficulties) and person measures (abilities) separately to produce estimates for each parameter that are sample and item independent respectively. The Rasch model uses location parameters (item location) to model item characteristics. The location parameters specify the position of the item and the item's response categories on the continuum of the latent variable. Thus, item parameters and person measures are directly comparable because they are on the same metric (Bond & Fox, 2007; Hambleton & Swaminathan, 2010; Linacre, 2015; van der Linden & Hambleton, 2010).

The basic Rasch model is a dichotomous response model (Rasch, 1960; Wright & Stone, 1979) that represents the conditional probability of a binary outcome as a function of a person's (n), trait level (β) and an item's difficulty (δ). When $\beta_n > \delta_i$, $\beta_n = \delta_i$, and $\beta_n < \delta_i$, P_{ni} (Probability of an endorsed response "yes" response to an item) >50 per cent, =50 per cent, and <50 per cent, respectively. Andrich (1978, 1988) is credited for extending Rasch dichotomous response model to the rating scale model. Generally, the simple dichotomous response model can be extended to provide an appropriate model for use with polytomous response categories by the addition of an additional difficulty parameter; either a second δ parameter or a τ parameter. For polytomously scored items, that is when there are m + 1 possible ordered response categories for each item (coded as $x = 0,1,\cdots,m$), following the rating scale model (Andrich, 1978), this probability is given by

$$P\left(X_{ni} = x\right) = \frac{\exp\left\{x\left(\beta_n - \delta_i\right) - \sum_{j=0}^{x}\tau_j\right\}}{\sum_{k=0}^{m}\exp\left\{k\left(\beta_n - \delta_i\right) - \sum_{j=0}^{k}\tau_j\right\}} \tag{1}$$

where $\tau_0 = 0$, β_n identifies the *ability* of the person n, δ_i, the *mean difficulty* of the item i and τ_j, called *threshold*, is the point of equal probability of categories $j - 1$ and j. Thresholds add up to zero,

$$\sum_{j=1}^{m} t_j = 0$$

Masters advocated that the rating scale model can be used to analyse "questionnaires in which a [Likert-type]fixed set of responses alternatives like "strongly disagree", "disagree", "agree", "strongly agree" was used with every item on the questionnaire" (Masters, 1982: p. 105). Smith, Rush, Fallowfield, Velikova, & Sharpe (2008) implied that thresholds are derived for each adjacent response category in a scale. For k response categories, there are $k - 1$ thresholds. Each threshold (Fk) has its own estimate of difficulty. Thus, the rating scale model describes the probability, P_{ni} of a person with ability β_n choosing a given category with a threshold F_k and item difficulty δ_i.

It is important to note that the Likert scale can be modelled with either the rating scale or the partial credit model (Masters, 1988; Wright & Masters, 1982). The partial credit model allows the item format and the number of categories to vary from item to item (e.g., some items are scored with a 5-point scale and others with a 6-point scale). When the item format is inconsistent from item to item, the partial credit model is useful in providing estimates of the psychological distance between each set of the ordinal categories (Masters, 1988). However, the rating scale model restricts the step structure to be the same for all items (Linacre, 2000; Wright & Masters, 1982). The rating scale model is the recommended model when all items share a common rating scale. In fact, Linacre indicated that strong evidence would be needed to use a model other than the rating scale model (e.g., partial credit model) where all items have the same rating scale. However, in essence, the rating scale models are a subset of the partial credit models (Andrich, 1988).

1.5. Problem and Rationale for the Present Study

Cultural values and norms in relation to physical activity in different countries may account for the cross-cultural variations in levels of physical activity reported by adolescents (Lee & Martinek, 2009). In fact, a number of studies have examined physical activity among specific cultural groups (Nakamura, 2002; Vertinsky, Batth, & Naidu, 1996) and found that social norms of ethno-cultural communities play a significant role in exposure to and attitudes towards physical activities and these factors in turn affect actual physical activity. Furthermore, culture values and norms may affect not only the type of information provided by the various sources of self-efficacy (Bandura, 1997), but also what information is selected and how it is weighted and integrated into a person's self-efficacy judgments. For example, people in an individualist culture may focus their self-appraisals on information about their personal attainments. On the other hand, for people in a collectivist culture, evaluation of their performance by members of their cultural group may be the most important source of efficacy formation. In an individualist society, when approaching a new task, an individual's self-appraisal of efficacy is likely to be affected by his or her previous performance on similar tasks. In contrast, in a collectivist society, an individual's self-appraisal of efficacy is likely to be affected by the beliefs of the cultural group. Does the group think the individual has the capability to perform the task? Would other members of the group be likely to do the task better (Bandura, 1986)?

In line with these concerns about the effects of cultural values and norms on individuals' actual physical activity and self-efficacy, the psychometric properties of the PASESA needs to be assessed in a non-Western context because it is possible that instruments developed in the West might not work in the same manner in non-Western settings due to cultural differences (Maneesriwongul & Dixon, 2004). In fact, we know little about self-efficacy to overcome specific barriers to physical activity and stick to excise routine amongst native Omanis, especially adolescents, because apparently no study has been conducted amongst native Omanis using the PASESA. It also is possible that the paucity of physical activity self-efficacy research in Oman has, in part, been due to the lack of Arabic language measures with acceptable psychometric properties and also to the fact that many Omanis do not have an adequate command of the English language for the use of English language questionnaires. Thus, there is a lack of a rapidly applicable and reliable measure of physical activity self-efficacy in the Arabic-speaking context; a tool not only valid but also useful for providing information about individuals' judgment of their efficacy to get themselves to exercise regularly in the face of different obstacles. Taking together, it seems that more research is *needed to examine* the psychometric properties of the PASESA within an Omani context.

A review of previous *research that* evaluated the psychometric properties of *the PASESA* (e.g., Eaton & Crcook, 2013; Mohammad-Hassun & Ahmed-Kali, 2014; Peterson & Shavelson, 2014; Simpson & Rustin, 2012; Wu et al., 2011) revealed that these research has mainly used the Classical Testing Theory (CTT) despite the fact that the CTT suffers several limitations. For example, the CCT approaches assume an interval data and that all items in a scale are equally difficult. These approaches do not allow for a separation of the evaluated persons and items—these are both interpreted in the context of the other. The methods of analysis that has been used within the CTT to evaluate the psychometric properties of tests suffer the limitation that they are sample dependent. The focus of the CTT is on the total score (i.e., the sum of the responses to each item on the scale) not individual items. Furthermore, the CTT lacks of additivity of rating scale data (i.e., equal-interval property of the scale).

An alternative approach to examine the psychometric properties of the Arabic version of the PASESA (PASESA-Av) within an Omani context is to use the Rasch rating scale analysis. To date, it appears that no studies available have used a Rasch analysis approach to examine the psychometric properties of PASESA within Western or non-Western contexts (Bond & Fox, 2007; Boone, Staver, & Yale, 2014; de Ayala, 2009; Hambleton & Swaminathan, 2010). The Rasch model is one specific application of the Item Response Theory (IRT). The IRT, also referred to as Latent Trait Test Theory (LTTT), constitutes a mathematical model of the process through which persons interact with test items, depicting the probability of various outcomes that may result from that interaction. Specifically, the IRT explains a person's performance on a test using an exponential function comprising item characteristics and the person's latent ability measured by responses to items. Built around a dichotomous logistic response model (suitable for Yes/No response choices), the Rasch model specifies that each item response is taken as an outcome of the linear probabilistic interaction of a person's "ability" and an item's "difficulty". The Rasch model estimates each item's difficulty as well as each person's ability on the same metric, allowing for meaningful comparisons of the two. The Rasch model generates reliability and validity estimates of both persons and items that are independent of the sample distribution. These estimates can be used for in-depth monitoring of test functioning. For example, the PASESA-Av items and persons demonstrating poor fit to the Rasch model have unexpected response patterns given the item's estimated difficulty and the respondent's estimated level of physical activity self-efficacy. This information can be useful in identifying items that do not contribute to a valid measure of the underlying trait, as well as potential response biases related to respondent characteristics contexts (Bond & Fox, 2007; Fox & Jones, 1998; Boone et al., 2014; de Ayala, 2009; Hambleton & Swaminathan, 2010; van der Linden & Hambleton, 2010). When the Rasch rating scale model is used in the present study, *student ability* refers to physical activity self-efficacy, and *item difficulty* to the level of endorsement to an item. The easier the items, the more it is endorsed by students.

1.6. Aims of the Study

The aim of this study is to examine the psychometric properties of the PASESA-Av within an Omani context using the Rasch rating scale model. Specifically, the present study intended to:

(1) Examine the unidimensional factorial structure of the PASESA-Av as supported in Wu et al.'s (2011) original study and other studies conducted internationally (i.e., (a) whether items exhibit expected response patterns given each participant's estimated level of physical activity self-efficacy, (b) whether participants exhibit expected response patterns given each item's estimated level of difficulty, and (c) whether the majority of the variance explained by a single underlying construct).

(2) Examine the rating scale functioning of the PASESA-Av [not at all true = 1, not very true = 2 sort of true = 3, and very true = 4] (i.e., whether the PASESA-Av response categories logically reflect less/more physical activity self-efficacy).

(3) Examine whether the PASESA-Av can separate participants into sufficient number of distinct levels of physical activity self-efficacy.

(4) Examine whether the PASESA-Av items construct a well-defined variable and form a reproducible item difficulty hierarchy.

1.7. Significance of the Study

The present study makes a useful contribution for three main reasons:

(1) The growing research interest in the effect of culture and norms on individuals' physical activity and trait self-efficacy, coupled with a prior lack of empirical evidence for psychometric properties of PASESA-Av within non-Western contexts (such as Oman) highlighted the need for this type of investigation.

(2) The results of the present study are expected to facilitate and promote physical activity self-efficacy research within the Omani context by making available rapidly applicable and valid and reliable measure of physical activity self-efficacy for adolescents.

(3) The Rasch rating scale model represents a meticulous procedure to validate PASESA-Av within non-Western contexts (such as Oman) that offer solutions to many problems of the conventional CTT.

1.8. Questions of the Study

The present study intended to answer the following research questions:

(1) **Content validity**: What is the technical quality of the PASESA-Av items, and more specifically, do the PASESA-Av items correlate to form a dimensional structure (i.e., physical activity self-efficacy)?

(2) **Structural Validity**: Do the items in the PASESA-Av support a unidimensional underlying construct; that is: (a) Do the PASESA-Av items' response patterns across the participants demonstrate acceptable goodness-of-fit to the Rasch model (i.e., do the items exhibit expected response patterns given each participant's estimated level of physical activity self-efficacy)? (b) Do the participants' response patterns on the PASESA-Av items demonstrate acceptable goodness-of-fit to the Rasch model (i.e., do the participants exhibit expected response patterns given each item's estimated level of difficulty)? (c) Is the majority of the variance explained by a single underlying construct?

(3) **Substantive validity**: What is the structure of the PASESA-Av response scale, and more specifically, do the response categories [not at all true = 1, not very true = 2 sort of true = 3, and very true = 4] logically reflect less/more physical activity self-efficacy?

(4) **Persons reliability**: Does the PASESA-Av separate participants into a sufficient number of distinct levels of physical activity self-efficacy?

(5) **Items reliability**: Do the PASESA-Av items construct a well-defined variable and form a reproducible item difficulty hierarchy?

2. Methods

2.1. Participants

Participants of the present study included 260 Year 9 students enrolled in seven public schools in three governorates in Sultanate of Oman; Muscat (3 schools), Al-Batinah South (2 schools), and Ash Sharqiyah (2 school). All schools were from metropolitan areas and had single-gender population. The means and the standard deviations of the whole sample age were 14.73 and .72 for boys (n = 135) and 14.42 and .47 for girls (n = 125). Only students with complete data were retained for the present study. The percentage of missing data was 2%. Those students left several item blank on the PASESA-Av. The analysis of demographic data showed that Arabic was the first language for all participant students and that 98% of them belong to the middle socioeconomic class.

2.2. Measures

2.2.1. Physical Activity Self-Efficacy Scale for Adolescents

Wu et al. (2011) developed the PASESA using a sample of 206 middle school students ranging in age from 11 - 14 years and enrolled in physical education or gym classes at a public middle school in the United States. The PASESA intended to measure each participant's confidence in his or her ability to overcome specific barriers to physical activity. The PASESA consisted of 11 items that were preceded by the phrase "Please show how true each statement is regarding how sure you are that you can exercise, be active, or do sports when you face." Participants responded to each item of the PASESA using a 4-point Likert scale; "Not at all true," "Not very true," "Sort of true," and "Very true". The Likert scale received the scores 1, 2, 3, and 4 respectively because all items were positive. Scores for each participant were determined by computing the mean of a participant's item scores. Higher scores corresponded to a higher level of perceived self-efficacy to overcoming barriers to physical activity.

2.2.2. Translation of the PASESA

The author translated the PASESA from English into Arabic using the back-translation method. Three qualified translators who held a doctoral degree in English and a minor in psychology, working without referencing to the English version of the PASESA, independently translated the Arabic version back to English. Three other qualified translators who also held a doctoral degree in English and a minor in psychology independently compared the original English version of the PASESA to the new English version that was translated back from Arabic, and rated the match between every two corresponding versions on a scale from 1 to 5. A score of 1 represented poor match, whereas a score of 5 represented perfect match. The average percentage of match was 99% which could be considered acceptable (see, Sousa & Rojjanasrirat, 2011).

Krippendorff's alpha was also used to address the reliability of ratings among the panel of translators. Krippendorff's alpha is a statistical measure of the agreement achieved when coding a set of units of analysis in terms of the values of a variable. The value of Krippendorff's alpha would normally goes between 0 and +1. When raters agree perfectly, Krippendorff's alpha will be equal to 1. When raters disagree perfectly, Krippen-dorff's alpha will be equal to zero. Krippendorff's alpha has the advantage of being applicable to any number of observers, not just two, any number of categories, scale values, or measures, any metric or level of measurement, incomplete or missing data, and large and small sample sizes alike, not requiring a minimum (Hayes & Krippendorff, 2007; Krippendorff, 2004). In the present dataset, Krippendorff's alpha was .97 (LL95%CI = .94; UL95%CI = .99). Although there is no universally acceptable cutoff score for Krippendorff's alpha, the values reported in the present study may suggest satisfactory levels of agreement among the raters (Krippendorff, 2004).

2.3. Procedures

Approval was obtained to conduct the research investigation at the schools prior to data collection. Students were recruited to participate in the present study during their normal physical education classes at their schools. Only certain classes in each school participated in the present study depending on students' classroom schedules on the day and time of the administration of the measures. Students were informed that participation was voluntary and that confidentiality of their answers would prevail at all times. All students gave assent by signing a consent form prior to their participation in the study that they are willing to respond to the measure of the present study. The PASESA-Av was administered to participant students by trained experimenters according to standardized instructions during the second week of the first semester of the 2015/2016 school year. Students completed the PASESA-Av in 5 - 10 minutes. Apparently, the items of the PASESA-Av were within the age-equivalent reading level of the participant students because they did not indicate any difficulty understanding their content.

2.4. The Statistical Analysis

A Rasch rating scale model was applied to the PASESA-Av in the sample of the present study, because the generic scale used in PASESA-Av is expected to function in similar manner across all items. The WINSTEPS program, version 3.81.0 (Linacre, 2014) was used to run all Rasch analyses.

3. Results

(1) Content validity

What is the technical quality of the PASESA-Av items, and more specifically, do the items of the PASESA-Av correlate to form a dimensional structure (*i.e.*, *physical activity self-efficacy*)?

Content validity provides the evidence of content applicability and the representativeness of the construct being measured (Messick, 1989). One component of the content aspect of validity surrounds the technical quality of the items on an instrument (Dimitrov, 2012). The point measure correlations were estimated using Rasch measurement, which is analogous to the item-total correlation in CTT. Positive point-measure correlations (>.30) provide evidence for content validity (Linacre, 2015; Smith, 2003; Wolfe & Smith, 2007a, 2007b). In the present study, the point-measure correlations were used for assessing content validity of the PASESA-Av.

Point-measure correlations. **Table 1** shows that the observed point-measure correlations were positive for all items of the PASESA-Av and they ranged from .74 to .87. Similarly, the expected value of the point-measure correlations when the data fit the Rasch model with the estimated measures was positive for all items of the PASESA-Av and they ranged from .73 to .86. These findings support the content validity of the PASESA-Av because a point-measure correlation reflects the extent to which person responses on items are in accord with the Rasch requirement; that higher category scoring corresponds to the presence of more of the latent variable (Linacre, 2015).

Category point-measure correlations. The item category options reflect correlations with the measures that support the Rasch measurement model (Dimitrov, 2012). The point-measure correlation for high and low category-option groupings was from −.41 to .37 for the low category group (1, 2) and was −.47 to .55 for the high category group (3, 4) which shows the accordance of category options' correlations with person measures. These values provide evidence for the content validity of the PASESA-Av.

(2) Structural validity

Do the items in the PASESA-Av support a unidimensional underlying construct?

Structural validity relates to the alignment of the scoring structure of an instrument to the structure of the construct around which the instrument is designed (Messick, 1995a, 1995b). In Rasch modeling terms, unidimensional measurement means that all of the non-random variance found in the data can be accounted for by a single dimension of difficulty and ability (Bond & Fox, 2015). In the present study, several diagnostic tools were used for assessing category functioning of the PASESA-Av. These tools include (Linacre, 2015):

Item fit statistics. Item fit statistics are a measure of how well the model predicts the data based on residuals (Abd-El-Fattah, 2007). For the data to fit the model adequately, it was generally recommended that the two fit statistics ranged from 0.72 to 1.30 logits. The infit and outfit mean squares can be converted to an approximately

Table 1. Rasch point-measure correlations of the PASESA-Av.

Items	Point-measure correlations	
	Correlation	Expected
SE_1: I feel self-conscious or concerned about my looks when I exercise.	.87	.86
SE_2: I am not motivated or feeling too lazy at exercise at the time.	.82	.83
SE_3: I am too busy.	.85	.84
SE_4: I have to exercise alone.	.78	.77
SE_5: I am afraid to fail.	.80	.79
SE_6: The weather is bad.	.75	.74
SE_7: I have minor aches and pains from activity.	.79	.75
SE_8: I am tired.	.83	.82
SE_9: I have a bad day at school.	.74	.75
SE_10: It is very hard work.	.82	.73
SE_11: I am sure that I can still exercise, be active, or do sports even if I face certain barriers or problems.	.76	.77

Note: *N* = 260.

normalized *t*-statistic using the Wilson-Hilferty transformation with the critical values are usually set at ±2. This means that the acceptable values for *t* values should be >±2 ($p < .05$) (Bond & Fox, 2015). The analysis showed that no item has been deleted from the calibration procedure because the values of the INFIT and the OUTFIT MNSQ statistics for all items fall within the accepted range of 0.72 to 1.30 logits. Also, the *t* values associated with these two statistics were >±2.0. **Table 2** shows that INFIT MNSQ values ranged from 0.92 to 1.20 (INFIT *t* values ranged from 3.15 to 7.22), and that OUTFIT MNSQ values ranged from 0.90 to 1.23 (OUTFIT *t* values ranged from 3.45 to 7.40).

Person fit statistics. Person fit statistics are statistical methods used to detect improbable item-score patterns given an IRT model. They give the performance level of each student on the total scale. From a statistical Rasch perspective, persons and items are exactly the same. They are merely parameters of the Rasch model. So the fit criteria would be exactly the same (Bond & Fox, 2015; Linacre, 2015). As such, it is recommended that the INFIT MNSQ and OUTFIT MNSQ statistics ranged also from 0.72 to 1.30 logits and that the acceptable values for infit *t* and outfit *t* values range from −2 to +2 ($p < .05$) (Abd-El-Fattah, 2007; Bond & Fox, 2015). The analysis showed showed that the INFIT and the OUTFIT MNSQ statistics for all persons fall within the accepted range of 0.72 to 1.30 logits and that the *t* values associated with these two statistics were >±2 except for two persons. The INFIT MNSQ values ranged from 0.90 to 1.17 (INFIT *t* ranged from 4.20 to 8.62). The OUTFIT MNSQ values ranged from 0.87 to 1.20 (OUTFIT *t* values ranged from 3.30 to 6.55). For the first outlier person, the INFIT MNSQ value was 0.45 (INFIT $t = 1.12$) and OUTFIT MNSQ value was 0.66 (OUTFIT $t = 1.35$). For the second outlier person, the INFIT MNSQ value was 0.39 (INFIT $t = 0.95$) and ONTFIT MNSQ value was 0.54 (OUTFIT $t = 1.10$). These values violates the guideline that the INFIT and OUTFIT MNSQ should range from 0.72 to 1.30 logits with associated $t > \pm2.0$ (Bond & Fox, 2015). Both students were excluded from the calibration procedure.

Rasch principal components analysis (*Rasch-PCA*) *of residuals*. The Rasch-PCA of residuals is an un-rotated technique with an underline hypothesis that there is only one dimension, called the *Rasch dimension*, captured by the model so that the residuals do not contain other significant dimensions (Boone et al., 2014; Linacre, 2015). Three criteria were used to suggest unidimensionality: (1) at least 50% of the total variance should be explained by the first latent variable/dimension (Rasch dimension); (2) The first contrast should *not* have an eigenvalue > 2.0 because an eigenvalue of 2.0 represents the smallest number of items that could represent a second dimension (2 items); and (3) the ratio of the percent of raw variance explained by the measures (persons and items) to the percent of total variance explained in the first contrast should exceed three (Linacre, 2015; Bond & Fox, 2015).

The results of the Rasch-PCA, summarized in **Table 3**, support the unidimensionality of the PASESA-Av. The Rasch dimension explained 84.6% of the total variance, with 75.7% explained by persons and 8.9% explained by items. This indicated that a dominant first factor was present. There were small amounts of unexplained variances in the components which came from the residuals (3.2%, 2.3%, 1.6%, and 1.2% for the first,

Table 2. Item fit statistics of the PASESA-Av.

Items	INFIT MNSQ	ITFIT *t*	OUTFIT MNSQ	OUTFIT *t*
1	0.99	4.23	1.12	3.55
2	1.15	3.75	0.96	4.22
3	1.10	3.15	1.20	6.30
4	1.20	5.12	1.23	7.40
5	0.92	4.60	0.94	3.45
6	1.20	6.30	0.90	4.40
7	1.12	4.20	1.13	5.20
8	0.95	5.10	0.97	3.90
9	1.10	7.22	0.87	4.60
10	0.98	4.19	1.15	5.50
11	1.15	3.80	1.17	4.70

Note: $N = 260$.

Table 3. Rasch-PCA for the PASESA-Av.

	Eigenvalue Units	Observed %		Expected %
Total raw variance in observations	65	100		100
Raw variance explained by measures	55	84.6		84.4
Raw variance explained by persons	49.2	75.7		75.5
Raw variance explained by items	5.8	8.9		8.9
Raw unexplained variance (total)	10	15.4	100%	15.6
Unexplained variance in 1st contrast	1.6	3.2	20.7	
Unexplained variance in 2nd contrast	1.2	2.3	14.9	
Unexplained variance in 3rd contrast	.80	1.6	10.3	
Unexplained variance in 4th contrast	.50	1.2	7.7	

Note: *N* = 260.

second, third, and fourth contrasts, respectively). Residual factor loadings suggested that the data closely approximated the Rasch model, and there were no meaningful components beyond the primary dimension of measurement. The Rasch-PCA demonstrated, by and large, that there were no extraneous dimensions. Furthermore, the first contrast eigenvalue was 1.6, which showed that the unexplained variance did not have the strength of more than two items. The ratio of the the percent of raw variance explained by the measures (persons and items) [84.6%%] and the percent of total variance explained in the first contrast [3.2%] exceeded the widely-accepted minimum 3:1 ratio.

Disattenuated correlations. Disattenuated correlations can be calculated for person measures on clusters of items on the instrument. The WINSTEPS program partitions items into three clusters, obtains person measures on each of the three clusters, and reports disattenuated correlations between person measures on each of the three clusters. Correlations approaching 1.0 indicate empirically that the clusters of items are measuring the same thing and that the measure is likely unidimensional (Boone et al., 2014; Linacre, 2015). It was found that ***the*** disattenuated first contrast person-measure correlations on the item clusters were .98 (item cluster 1 - 3) and .99 (item clusters 1 - 2 and 2 - 3). These findings indicate that the clusters of items are measuring the same thing and that the PASESA-AV is likely unidimensional.

Local independence. An underlying assumption of Rasch measurement is the notion of local independence. Local independence requires that a participant's response on one item is not dependent on his or her response to another item. Unidimensionality can be defined through local independence. Local independence means that once the latent trait level is controlled for, there are no significant correlations left among the items (Bond & Fox, 2015; van der Linden & Hambleton, 2010). In the present study, local independence of the PASESA-AV was evaluated using WINSTEPS program by investigating the standardized residual item correlations, which were obtained via ICORFILE from the output file menu. Except for one item (*I am too busy and It is very hard work* = 0.10), the residual item correlations were all negative or zero, which suggests that the items reflect local independence (Linacre, 2015). The small positive value of the correlation between the two items was likely far too small to present a dependency concern, as values below .30 are generally not considered to indicate a violation of local independence (Christensen, Kreiner, & Mesbah, 2013). However, Christensen et al. added that residual item correlations on instruments of fewer than 20 items (the PASESA-Av has 11 items) may not be as confidently interpreted as those with a large number of items. To overcome this limitation, they suggest comparing the magnitude of the residual item correlation to the average residual correlation for all items in a set rather than depending solely on cut-off value criteria. The 0.10 correlation between the two items on the PASESA-Av were not particularly larger than the average correlation on PASESA-Av of −.15. Thus, there is strong evidence that the items on the PASESA-Av exhibit local independence.

(3) Substantive validity

What is the structure of the PASESA-Av response scale, and more specifically, do the response categories [*not at all true* = 1, *not very true* = 2 *sort of true* = 3, *and very true* = 4] *logically reflect less/more physical activity self-efficacy*?

One component of substantive validity that is particularly salient in Rasch-based arguments is scale functioning (Dimitrov, 2012). Substantive validity (Messick, 1989, 1995a, 1995b) refers to the observed consistency and patterns in responses of examinees or raters. Reliable and valid measurements require that rating categories be substantively different and meaningful for respondents (Linacre, 2002). Essential to the functioning of a scale is the advancement monotonically of the categories at higher levels of trait (Dimitrov, 2012). The step measure parameter defines the boundaries between categories which should increase monotonically with categories. Disordering of step measures occurs when the rating scale does not function properly. If this occurs, collapsing response categories is suggested to minimize this problem (Bond & Fox, 2015). Rasch measurement can evaluate the effectiveness of rating scale structure. This provides an instrument's author insight into the substantive meaning of categories interpreted by respondents in order to take decisions about category labels or category number (van der Linden & Hambleton, 2010). In the present study, several diagnostic tools were used for assessing category functioning of the PASESA-Av. These tools include:

Category use distributions and outfit MnSq of categories. **Figure 1** shows that the distributions of category responses for the items of the PASESA-Av were generally unimodal, which suggest that there is no unusual category usage. Outfit Mean Square (MnSq) of categories was evaluated to determine if the category is functioning expectedly. Values greater than 2.0 indicate a large amount of unexplained noise in the data and significant misinformation in the category (Linacre, 2002, 2015). **Table 4** shows that the PASESA-Av has an optimal category structure; that is categories advanced monotonically because the Outfit MnSq values were between 0.95 and 1.05 for all categories. These findings support the notion that the PASESA-Av has an optimal category structure.

Average measure. This is the ability estimates for all participants who choose a particular category, which is expected to advance monotonically with categories along the rating scale (Linacre, 2002, 2015). The average measure relative to item difficulty difference, $\theta_n - \delta_i$, for all persons can be used to evaluate the rating scale across categories (Because the rating scale structure is the same across all items τ_k is a constant that can be neglected) (Boone et al., 2014). **Table 4** shows observed average measures relative to item difficulties and the expected average measures relative to item difficulties for the PASESA-Av. The average measures demonstrated monotonicity across the PASESA-Av as measured on the logit scale. The expected values for the average measures in logits did not reflect any concerning differences from their observed values.

Step difficulties. The number of response categories should be large enough to identify along a wide range of the variable on an instrument but small enough so that respondents can conceptualize substantive differences of meaning between the category labels (Linacre, 2002, 2015). The step difficulties between two categories k and $k - 1$ should advance by at least 1.0 logit, which indicates a meaningful dichotomy between label k and label $k - 1$. Step difficulties should advance by no more than 5.0 logits, or a loss of precision and information results (Bond & Fox, 2015). **Table 4** shows that each advance in step difficulties (Andrich thresholds; logits) was large enough (>1.0 logit) to support an interpretation that movement across the threshold meaningfully indicated successfully overcoming the impediment from category $k - 1$ to category k, and that each advance was small enough (<5.0 logit) to prevent a loss of precision in measurement.

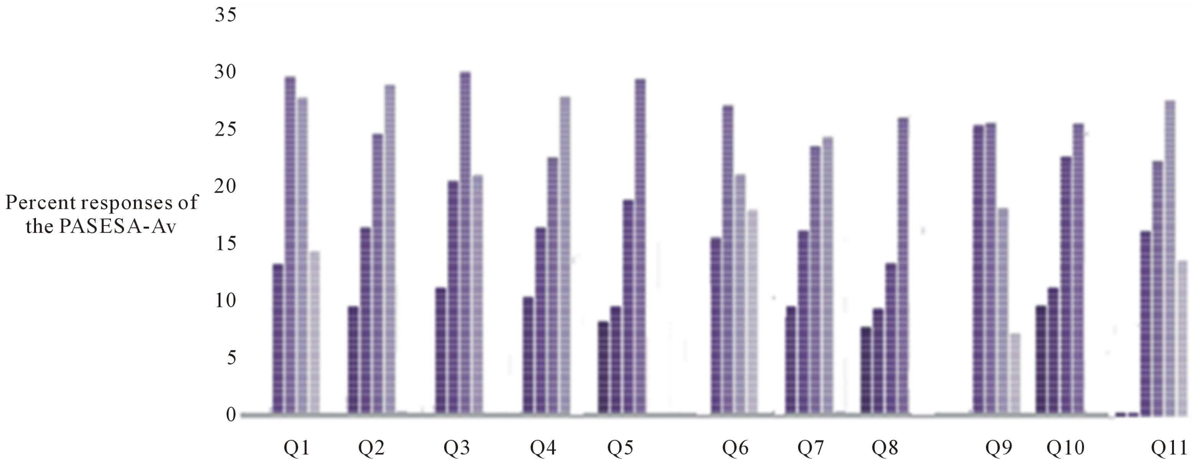


Figure 1. Distribution of responses by category for the PASESA-Av items.

Table 4. Outfit MNSQ of categories, average measures (observed & expected), Andrich thresholds, and coherence between measures and category of the PASESA-Av.

Statistic	Category label			
	1 not at all true	2 not very true	3 sort of true	4 very true
Outfit MNSQ of categories	1.05	0.98	0.95	1.02
Average measures (Observed values in logits)	−3.16	−1.39	2.11	3.85
Average measures (Expected values in logits)	−3.12	−1.43	2.15	3.88
Andrich thresholds (Step difficulties)	None	−3.66	1.95	4.84
Measure-implies-category percentage coherence	72	69	65	74
Category-implies-measure percentage coherence	70	66	64	71

Note: *N* = 260.

Coherence between measures and category observations. Coherence expresses the number of measures that were expected to produce observations in a category as a proportion of those that actually did. Additionally, coherence expresses the proportion of observations in a category that were produced by measures corresponding to the category. A general coherence threshold is that it is acceptable above 40% (Dimitrov, 2012; Linacre, 2002). Coherence percentages are shown in **Table 4** and they were generally above the desirable level.

Step calibrations. Category probability functions are a visual representation of the probabilistic relationship between category difficulty and student location on the latent variable (Bond & Fox, 2015). Andrich (1988) emphasized that as the measures of persons increase, the probability of observing a person in a higher category should increase as well. Each curve represents an individual rating scale category, and the curves always appear in ascending order so that the curve representing the lowest category is farthest to the left and the curve for the highest category is farthest to the right. For example, for a person with low ability, the probability of observing the person in category zero must be higher than the probability of observing the person in category four. Visual examination of the probability characteristics curves can be used to verify the ordering of step calibrations. Each category should appear to have a peak at some point where it is the most probable category to be endorsed (Bond & Fox, 2015). **Figure 2** shows that increasing student location on the latent variable (i.e., physical activity self-efficacy) is associated with an increasing probability for observed ratings in higher categories.

(4) Person separation and reliability

Does the PASESA-Av separate participants into a sufficient number of distinct levels of physical activity self-efficacy?

Person separation and reliability. Person separation indicates the extent to which an instrument's scale discriminates well between persons (Smith, 2003). The real person separation accounts for any error that arises from model misfit (Bond & Fox, 2015; Smith, 2003). Real person separation greater than 2.0 with person reliability greater than .80 implies that the instrument is likely sensitive enough to accurately classify between those of high ability and those of low ability on the instrument. The person reliability expresses the probability that persons on the high range of ability do, indeed, have a high ability, while those of low ability are, indeed, likely to be found on the lower range of ability on another measure of the same variable (Linacre, 2015).

The analysis showed that the real person separation index was 5.65 and the real person reliability was .95. The real person separation index indicates that the PASESA-Av is sensitive enough to distinguish between high and low students with regard to physical activity self-efficacy trait. This is especially important to an instrument such as the PASESA-Av, which purports to be able to provide information about adolescents' self-confidence to to overcome specific barriers to physical activity. The real person reliability of 0.95 indicates that the PASESA-Av placed persons reliably on its respective measurement continuum, which suggested that person placement on another instrument measuring the same construct will likely show the same students above or below others as on the PASESA-Av.

Person/item map. The Rasch person-item map for the PASESA-Av is depicted in **Figure 3**. Self-reported ratings of physical activity self-efficacy in response to the PASESA-Av items are shown on the left hand side of the map, while the thresholds of the items of the overall PASESA-Av are on the right hand side. Numerical values on the extreme left hand side of the map which range from −3 to +4 are expressed as a log odd unit interval or logit which is the natural unit of the Rasch scale (Fox & Bond, 2007).

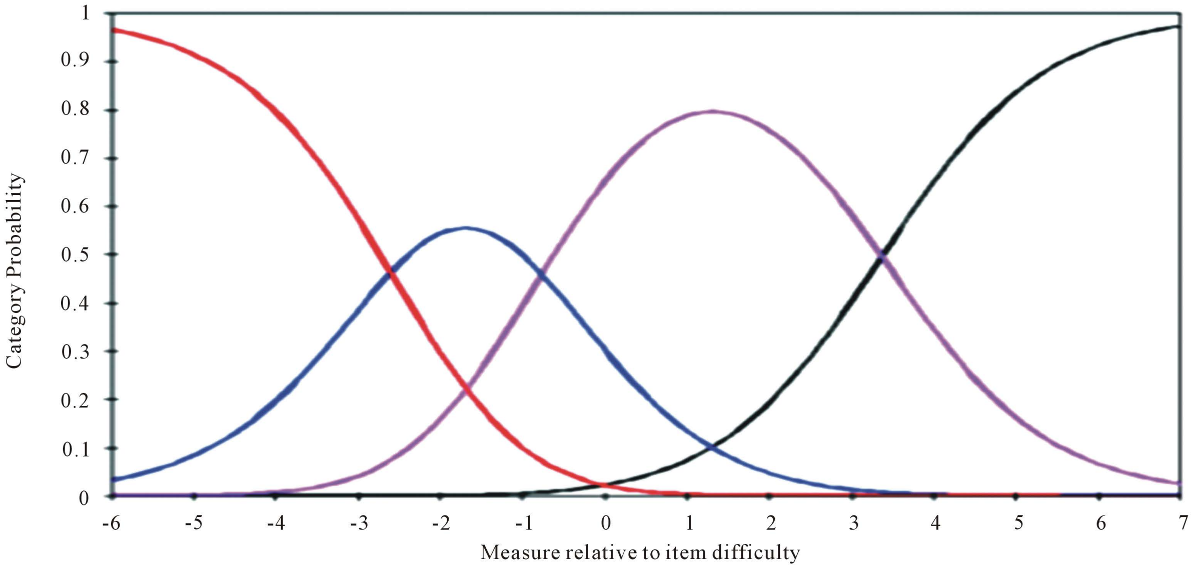


Figure 2. Category probability curve for the PASESA.

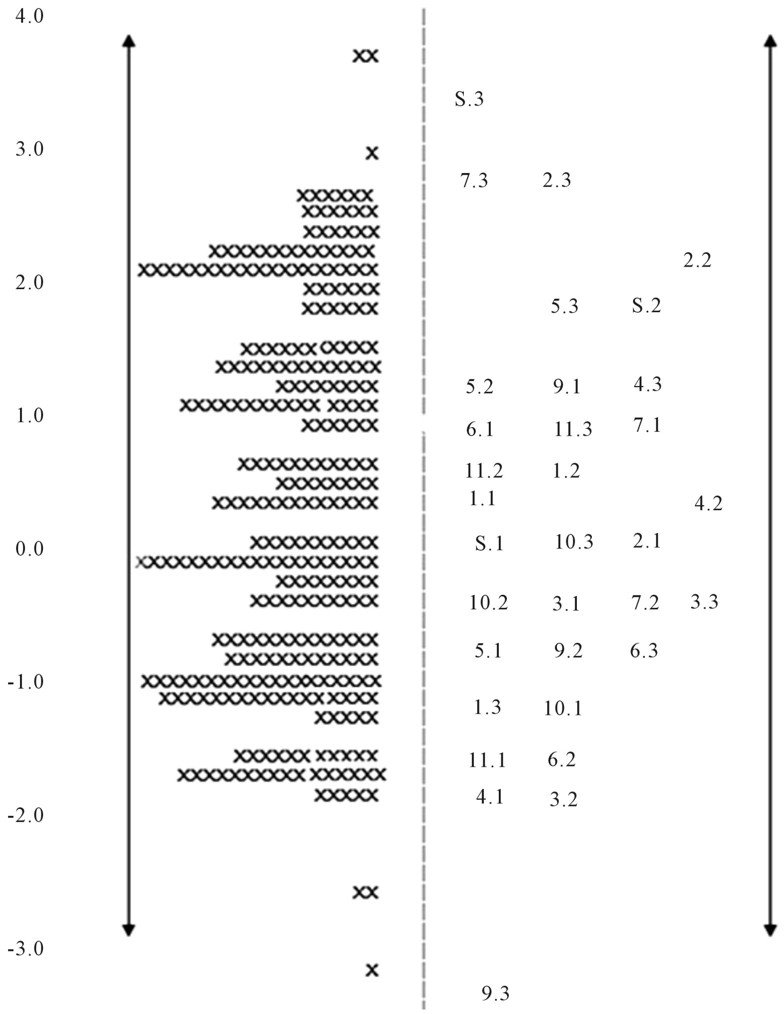


Figure 3. The person-item map of the PASESA-Av.

The Rasch person-item map is used to compare the range and position of the person measure distribution on the left hand side of **Figure 3** to the item measure on the right hand side. Persons represented in the map as an X appear in ascending order of PASESA-Av from the bottom of the figure to the top. Items on the right are represented by item numbers, with a decimal representing the response scale boundary or threshold of each of the ratings (Adams & Khoo, 1993; Curtis & Boman, 2007). Items at the top of the scale on the right hand side are harder for adolescents to rate (e.g., Item 8. I am tired.), while items become easier for adolescents further down the scale (e.g., Item 9. I have a bad day at school). Adolescents with higher physical activity self-efficacy at the top of the scale are more likely to rate the PASESA-Av items as being "Very true"; students with lower physical activity self-efficacy at the bottom of the scale are more likely to rate the items as "Not at all true".

The vertical scale is an interval scale. Spaces between items, between persons, and between items and persons have substantive meaning in terms of the underlying variable (Callingham & Bond, 2006). The physical activity self-efficacy of each adolescent while rating the statements is referred to as the 'person measure' and the level of physical activity self-efficacy while performing each item with a criterion level of difficulty is called 'item measure' (Adams & Khoo, 1993). Items should be located at each point on the scale to measure meaningful differences and must cover all the areas on the scale to measure the physical activity self-efficacy of all adolescents. Rasch rating scale structure parameters, the step calibrations or Tau's, are related directly to category probabilities. These probabilities relate to the probability of a category being observed, not to the substantive order of achievement of the categories (Linacre, 1999).

Furthermore, the difficulty of the PASESA-Av items ranged from −.60 to .86. Item 8 (response 3) is seen as the most difficult item in the PASESA (only 23% reported "not at all true" to this item) while Item 9 (response 3) is the easiest item (91 of respondents "not at all true" to this item). It was notable that two students at the higher end of the scale do not have any corresponding items, implying that while they have high levels of physical activity self-efficacy, the actual level cannot be estimated accurately because of the paucity of item thresholds. There are no items that discriminated between them at that measure level. Three students at the lower end of the scale who do not have any corresponding items from the PASESA-Av have a low level of physical activity self-efficacy, which has not been estimated with corresponding items.

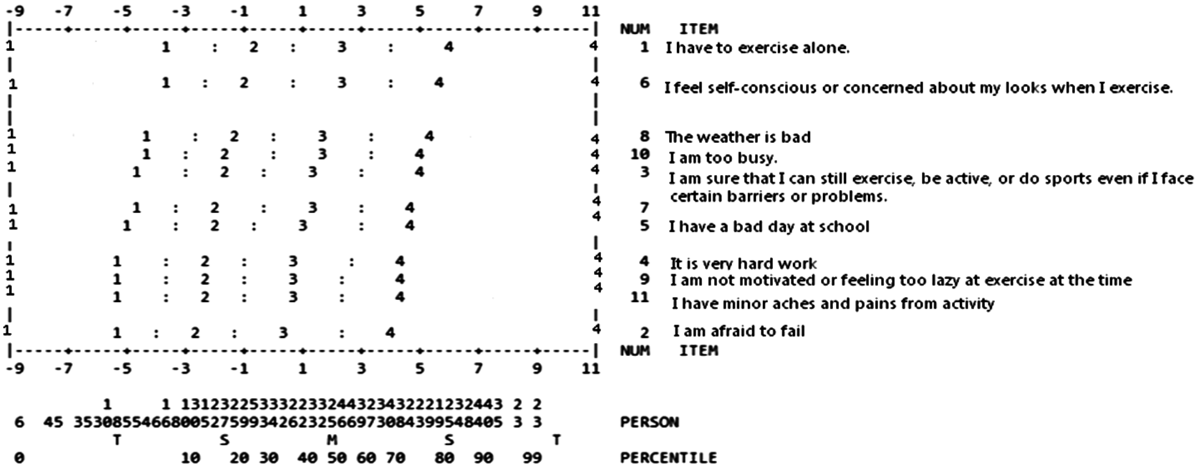
Most importantly, in **Figure 3**, both persons and items appear along the same scale with the 11 items of the PASESA-Av forming a unidimensional scale. The range of item difficulties approximately matches the range of students' scores, implying that the scale is appropriate for this group of students. Furthermore, Bond & Fox (2015) described the analysis of data using the Rasch model as "an estimate of what our construct might look like if we were to create a ruler to measure it (p. 8). In the present analysis, the fact that the item-person map spread along the continuum of the logit scale and that persons spread along the continuum of the ability estimates, the scale provide evidence for the content and substantive validity of the PASESA-Av. This means that the PASESA-Av provides "well-spaced items that [cover] a substantial length of the construct" (Green & Frantom, 2002: p. 27), which provides evidence for content and substantive validity.

(5) Item separation and reliability

Do the PASESA-Av items construct a well-defined variable and form a reproducible item difficulty hierarchy?

Item separation and reliability. Item separation confirms the item difficulty hierarchy. Item separation is particularly useful in providing evidence for content validity (Linacre, 2015). Values of item separation lower than 3.0 with item reliability less than .90 implies a lack of items at a wide enough range of difficulties to provide evidence for content validity; i.e., the items potentially do not offer a range of difficulties that cover the range of the construct and might not represent a well-defined variable (Smith, 2003). Item reliability is a measure of the extent to which the items on an instrument can be precisely located along the latent variable. Values of item reliability (<0.90) suggest that the items are not representative of a wide range of difficulty or that the sample size was too small (Bond & Fox, 2015; Linacre, 2015; Smith, 2003). The analysis showed that the real item separation was 6.20 and the real item reliability was .97. These values indicate that the items of the PASESA-Av create a well-defined variable and that the items formed a reproducible item hierarchy (Bond & Fox, 2007).

Key form. The key form provided a visual representation of the measure and the rating scale difficulty calibrated to the qualitative aspects of the scale—the items in hierarchical order. The key form, illustrates how logits and rating scale units (along the horizontal axis) were linked to the qualitative content-items in hierarchical order (along the right vertical axis) (Linacre, 2015). The location of rating units is arranged along the difficulty continuum expressed in logits. Each rating scale unit on the key form estimates the level of difficulty of achieving that rating for a specific item, as it related to other items and other rating units (Bond & Fox, 2007). In the key form, the



```
-9    -7    -5    -3    -1     1     3     5     7     9    11
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
1               1    :   2   :    3     :     4             4    1  I have to exercise alone.
|                                                           |
1               1   :   2    :    3     :    4              4    6  I feel self-conscious or concerned about my looks when I exercise.
|                                                           |
|                                                           |
1             1    :   2   :    3     :    4                4    8  The weather is bad
1             1   :   2    :    3     :    4                4   10  I am too busy.
1            1    :   2   :    3    :     4                 4    3  I am sure that I can still exercise, be active, or do sports even if I face
|                                                           |       certain barriers or problems.
1            1   :   2    :    3    :    4                  4    7
1           1    :   2   :    3     :    4                  4    5  I have a bad day at school
|                                                           |
1          1    :   2   :    3     :    4                   4    4  It is very hard work
1          1    :   2   :    3    :     4                   4    9  I am not motivated or feeling too lazy at exercise at the time
1          1    :   2   :    3    :     4                   4   11  I have minor aches and pains from activity
|                                                           |
1          1   :   2   :    3     :    4                    4    2  I am afraid to fail
|-----+-----+-----+-----+-----+-----+-----+-----+-----+-----|  NUM   ITEM
-9    -7    -5    -3    -1     1     3     5     7     9    11

          1     1 131232253332233244323432221232443 2 2
 6   45 35308554668005275993426232566973084399548405 3 3       PERSON
           T          S          M          S         T
 0                 10   20 30  40 50 60 70   80  90   99       PERCENTILE
```

Figure 4. Key form for the PASESA-Av.

most difficult items to endorse at higher option categories are shown at the top, while the items easier to endorse at higher option categories are shown at the bottom (Linacre, 2015). **Figure 4** is the key form that represents the item hierarchy for the PASESA-Av. The alignment of the items in such an order supports the underlying conceptual nature of the construct (i.e., physical activity self-efficacy) in terms of an increasing difficulty to endorse higher option categories for higher-placed items (item difficulty hierarchy)

4. Discussion

The aim of this study was to investigate the psychometric properties of the Arabic version of the Physical Activity Self-efficacy Scale for Adolescents (PASESA-Av) within a non-Western context using a sample of Omani adolescents. All analyses were based upon the Rasch rating scale because the generic scale used in PASESA-Av was assumed to function in similar manner across all items (Wright & Masters, 1982). Four psychometric properties were examined for the PASESA-Av: 1) item technical quality to evidence content validity; 2) unidimensionality to evidence structural validity; 3) rating scale functioning to evidence substantive validity, and 4) person and item separation and reliability.

For the technical quality of the PASESA-Av items, the observed and the expected point-measure correlations were positive for all items of the PASESA-Av. Values ranged from .74 to .87 for the observed correlations and from .73 to .86 for the expected correlations. These values were all greater than the recommended minimum criterion of .30, suggesting high technical quality of the PASESA-Av items. The positive values of the point-measure correlations suggest that the items were oriented in the same direction as the measure. These findings support the content validity of the PASESA-Av because a point-measure correlation reflects the extent to which person responses on items are in accord with the Rasch requirement that higher category scoring corresponds to the presence of more of the latent variable (i.e., physical activity self-efficacy) (Linacre, 2015; Smith, 2003; Wolfe & Smith, 2007a, 2007b). The category point-measure correlations were also in line with the measures. In other words, for each item, lower category options (1 and 2) showed more-negative point-measure correlations, while more-positive point-measure correlations were seen with higher category options (3 and 4).

The findings of the present study supported the unidimensionality of the PASESA-Av as a measure of a single underlying latent trait of physical activity self-efficacy. This pattern of results is consistent with Wu et al's (2011) original findings of the PASESA, and also replicates the findings from research conducted internationally by means of CTT on the factorial structure of the original English version of the PASESA in Canada (Simpson & Rustin, 2012), England (Gui & Smith, 2012), Scotland (Eaton & Crcook, 2013), and Singapore (Peterson & Shavelson, 2014) as well as the translated versions of the PASESA (Mohammad-Hassun & Ahmed-Kali, 2014). In the present study, the Rasch-PCA indicated that the 11-item version of the PASESA-Av explained 84.6% of the total variance, with 75.7% explained by persons and 8.9% explained by items suggesting that a dominant first factor was present (unidimensionality of the PASESA-Av). The ratio of the the percent of raw variance explained by the measures (persons and items) [84.6%%] and the percent of total variance explained in the first contrast [3.2%] exceeded the widely-accepted minimum 3:1 ratio**.** The components which came from the residual explained trivial amounts of variance offering further support for the unidimensionality of the PASESA-Av because no meaningful components beyond the primary dimension of measurement were possible. In fact the first contrast eigenvalue was 1.6, which showed that the unexplained variance did not have the strength of more than two items.

Another evidence for the unidimensionality of the PASESA-Av was embedded within the item and the person fit statistics (Abd-El-Fattah, 2007). Item fit statistics and their normalized *t*-statistic were within the accepted guideline for all items of the PASESA-Av. This means that the PASESA-Av response patterns across the participants demonstrate acceptable goodness-of-fit to the Rasch model (Yim, Abd-El-Fattah, & Lee, 2007). That is, the PASESA-Av items exhibit expected response patterns given each participant's estimated level of physical activity self-efficacy. The person fit statistics and their normalized *t*-statistic was also within the accepted guideline for all persons except for two cases (persons). This means that the participants' response patterns on the PASESA-Av items demonstrate acceptable goodness-of-fit to the Rasch model. That is, the participants exhibit expected response patterns given each item's estimated level of difficulty.

The two cases that did not fit the Rasch model were excluded from the calibration procedure. However, before they were excluded, an examination of the dataset showed that one student had a zero score and the other had a perfect score. Cases with zero scores are cases that had minimum possible scores on the scale, while cases with perfect scores are cases that have maximum possible scores on the scale (Alagumalai, Curtis, & Hungi, 2005). Because the Rasch model is a stochastic model rather than a deterministic model, some amount of misfit is tenable and expected (Bond & Fox, 2015). Callingham & Bond (2006) indicated that as many as 5% of the persons can exhibit misfit without causing concern in the social sciences (although they noted that for a high-stakes situation a 5% misfit rate on items must be addressed), while Linacre (2015) suggested that up to 10% misfit can be expected and not particularly interfere with Rasch measurement. These suggested proportions were chosen as guidelines for the analysis of person goodness-of-fit of the PASESA-Av to the Rasch model.

The local independence provide further evidence for the unidimensionality of the PASESA-Av because local independence means that once the latent trait level is controlled for, there are no significant correlations left among the items (Boone, et al., 2014). Except for one positive residual correlations on the PASESA-Av (I am too busy and It is very hard work = 0.10), the residual item correlations were all negative or zero, which suggests that the items reflect local independence. The small positive value of the correlation between the two items was likely far too small to present a dependency concern, as values below .30 are generally not considered to indicate a violation of local independence (Christensen et al., 2013). Overall, the findings from the person and item fit statistics, Rasch-PCA, and local independence supported the unidimensionality of the PASESA-Av that reflected structural validity of the measure. There are at least three related reasons why unidimensionality is important to consider in the PASESA-Av. Firstly, unidimensionality is a basic assumption for valid calculation of total scores according to both classic and modern test theories. Secondly, explicit interpretation requires scores to represent a single defined attribute of physical activity self-efficacy. That is, scores on a scale that is used to measure one variable should not be noticeably influenced by varying levels of one or more other variables. Thirdly, if scores do not represent a common line of inquiry, it is unclear if two individuals with the same score can be considered comparable. The interpretation of any differences between individuals will be ambiguous since it is unknown in what way(s) they actually differ (Nunnally & Bernstein, 1994; Smith Jr., 2002; Stout, 1987).

In addition to the technical quality and unidimensionality of the PASESA-Av items, the validity of the four-category rating scale was assessed. Specifically, the aim was to assess whether successive response categories for each item of the PASESA-Av were located in the expected order. Several criteria supported the utility of the four-category rating scale and that they are located in the expected order. First, the Outfit Mean Square (MnSq) of categories was <2.0 indicating the existence of a trivial amount of unexplained noise in the data and a significant amount of information in the category (Linacre, 2015). That is, the PASESA-Av has an optimal category structure with categories advanced monotonically. Also, the distributions of category responses for the items of the PASESA-Av were generally unimodal, which suggested that there was no unusual category usage (Andrich, 1988; Linacre, 2002).

Furthermore, the average measure showed that the ability estimates for all participants who choose a particular category advanced monotonically with categories along the rating scale, which indicated successful use of the rating scale categories by respondents. Each advance in step difficulties (Andrich thresholds; logits) was large enough (>1.0 logit) to support an interpretation that movement across the threshold meaningfully indicated successfully overcoming the impediment from category $k-1$ to category k, and that each advance was small enough (<5.0 logit) to prevent a loss of precision in measurement (Bond & Fox, 2015). This means that the step difficulties on the PASESA-Av showed that movement across thresholds reflected a meaningful advancement across impediments to higher category options. Across all scales, step difficulties were wider than 1.0 logit and narrower than 5.0 logits, which provided precision to the measures.

The validity of the PASESA-Av categories was pictorially supported through the category probability curve wherein increasing student location on the latent variable (i.e., physical activity self-efficacy) was associated with an increasing probability for observed ratings in higher categories. Taken together, these findings offered empirical evidence to the substantive validity of the PASESA-Av; that substantively different and meaningful categories of the PASESA-Av were interpreted by respondents. These findings also provides the PASESA-Av authors insight into the substantive meaning interpreted by respondents to ensure support for its use rather than relying on conjectural or anecdotal information to make decisions that the category labels and category number of the PASESA-Av are informative.

Finally, the item and the person separation and reliability of the PASESA-Av were investigated in the present study. For the item separation and reliability, the analysis showed that the real item separation was 6.20 and the real item reliability was .97. This means that there is enough items on the PASESA-Av to express enough range of difficulties hierarchy that cover the range of the construct (i.e., physical activity self-efficacy) and represent a well-defined variable (Boone, et al., 2014). Furthermore, the alignment of the items in such an order as depicted in the key form graph supports the underlying conceptual nature of the construct (i.e., physical activity self-efficacy) in terms of an increasing difficulty to endorse higher option categories for higher-placed items (item difficulty hierarchy). These findings offer further support for the PASESA-Av content validity. In addition, the high item reliabilities indicated the replicable position of items in terms of item difficulty; thus, the item reliabilities confirmed the item hierarchies of the scales and add evidence for the substantive aspect of construct validity (Linacre, 2015).

For the person separation and reliability, the analysis showed that the real person separation index was 5.65 and the real person reliability was .95. The real person separation index indicates that the PASESA-Av is sensitive enough to distinguish between high and low students with regard to physical activity self-efficacy trait. The real person reliability of .95 indicates that the PASESA-Av placed persons reliably on its respective measurement continuum, which suggested that person placement on another instrument measuring the same construct will likely show the same students above or below others as on the PASESA-Av (Linacre, 2015). The fact that the item-person map spread along the continuum of the logit scale and persons spread along the continuum of the ability estimates the scale provide evidence for the content and substantive validity of the PASESA-Av (Yim et al., 2007). This means that the PASESA-Av provides "well-spaced items that [cover] a substantial length of the construct" (Green & Frantom, 2002: p. 27), which provides evidence for content and substantive validity.

Overall, high person reliabilities indicated two characteristics of the PASESA-Av: (a) persons rated high on a scale had a high probability of indeed having a higher measure of the behavioral attributes on the scale and (b) persons were successfully placed into groups of varying performance levels. Importantly, this showed that students were clearly placed along a continuum of the measure rather than dichotomized into just high or low groups. Such a continuous measure provides opportunities for decisions to be based on a wide variety of ability levels to ensure that students can be exposed to experiences tailored to their developmental level (Bond & Fox, 2007; Linacre, 2015).

5. Conclusion

In summary, the development of the Arabic version of the PASESA is one of the strengths of this study. The Rasch measurement analysis demonstrated content validity of the PASESA-Av as a unidimensional measure of trait physical activity self-efficacy. The scale rating categories were found to advance monotonically and increase uniformly. In addition, the scale item and person separation and reliability were satisfactory. Although future studies are needed to replicate these results in additional settings, our findings suggest that researchers and practitioners can be confident in their interpretation of the PASESA-Av scores when used within an Arab context.

References

Abd-El-Fattah, S. M. (2007). Is the Aggression Questionnaire Bias Free? A Rasch Analysis. *International Education Journal, 8,* 237-248.

Adams, R. J., & Khoo, S. T. (1993). *QUEST: The Interactive Test Analysis System*. Hawthorn: Australian Council for Education Research.

Alagumalai, S., Curtis, D. D., & Hungi, N. (2005). *Applied Rasch Measurement: A Book of Exemplars*. Dordrech: Springer.

Andrich, D. (1978). Rating Formulation for Ordered Response Categories. *Psychometrika, 43,* 561-573. http://dx.doi.org/10.1007/BF02293814

Andrich, D. (1988). *Rasch Models for Measurement.* Beverly Hills, CA: Sage Publications.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York: Freeman.

Bandura, A. (2002). Social Cognitive Theory in Cultural Context. *Applied Psychology: An International Review, 51,* 269-290. http://dx.doi.org/10.1111/1464-0597.00092

Bandura, A. (2004). Health Promotion by Social Cognitive Means. *Health Education and Behaviour, 31,* 143-164. http://dx.doi.org/10.1177/1090198104263660

Bandura, A. (2005). The Evolution of Social Cognitive Theory. In K. G. Smith, & M. A. Hitt (Eds.), *Great Minds in Management* (pp. 9-35). Oxford: Oxford University Press.

Bandura, A. (2008). Toward an Agentic Theory of the Self. In H. W. Marsh, R. G. Craven, & D. M. McInerey (Eds.), *Self-Processes, Learning, and Enabling Human Potential* (pp. 15-49). Advances in Self-Research, Greenwich, CT: Information Age Publishing.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Mahwah, NJ: L. Erlbaum.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Mahwah, NJ: L. Erlbaum.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer. http://dx.doi.org/10.1007/978-94-007-6857-4

Callingham, R., & Bond, T. (2006). Research in Mathematics Education and Rasch Measurement. *Mathematics Education Research Journal, 18,* 1-10. http://dx.doi.org/10.1007/BF03217432

Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds.) (2013). *Rasch Models in Health.* Hoboken, NJ: John Wiley & Sons, Inc.

Curtis, D. D., & Boman, P. (2007). X-Ray Your Data with Rasch. *International Education Journal: Comparative Perspectives, 8,* 249-259.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Dimitrov, D. M. (2012). *Statistical Methods for Validation of Assessment Scale Data in Counselling and Related Fields*. Alexandria, VA: American Counseling Association.

Eaton, L., & Crcook, S. (2013). Factorial Validity and Reliability of the Physical Activity Self Efficacy Scale for Adolescents. *North American Journal of Educational Studies, 1,* 25-37.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch Modelling in Counselling Psychology Research. *Journal of Counselling Psychology, 45,* 30-45. http://dx.doi.org/10.1037/0022-0167.45.1.30

Gao, Z., Lee, A. M., Kosma, M., & Solmon, M. A. (2010). Understanding Students' Motivation in Middle School Physical Education: Examining the Mediating Role of Self-Efficacy on Physical Activity. *International Journal of Sport Psychology, 41,* 199-215.

Gao, Z., Lochbaum, M., & Podlog, L. (2011). Self-Efficacy as a Mediator of Children's Achievement Motivation and In-Class Physical Activity. *Perceptual and Motor Skills, 113,* 969-981. http://dx.doi.org/10.2466/06.11.25.PMS.113.6.969-981

Gao, Z., Lodewyk, K., & Zhang, T. (2009). The Role of Ability Beliefs and Incentives in Middle School Students' Intentions, Cardiovascular Fitness, and Effort. *Journal of Teaching in Physical Education, 28,* 3-20.

Green, K., & Frantom, C. (2002). *Survey Development and Validation with the Rasch Model*. Presentation Manuscript, Charleston, SC. https://portfolio.du.edu/downloadItem/115525

Gui, B., & Smith, P. (2012). Validation of the Physical Activity Self-Efficacy Scale for Adolescents in a Sample of British

Youth. *Journal of Psychology and Gymnastics, 2,* 72-85.

Hambleton, R. K., & Swaminathan, H. (2010). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff Publishing.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures, 1,* 77-89. http://dx.doi.org/10.1080/19312450709336664

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.) Thousand Oaks, CA: Sage Publications.

Lee, O., & Martinek, T. (2009). Navigating Two Cultures: An Investigation of Cultures of a Responsibility Based Physical Activity Program and School. *Research Quarterly for Exercise and Sport, 80,* 230-240. http://dx.doi.org/10.1080/02701367.2009.10599557

Linacre, J. M. (1999). Category Disordering vs. Step (Threshold) Disordering. *Rasch Measurement Transactions, 13,* 675.

Linacre, J. M. (2000). Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions, 14,* 768.

Linacre, J. M. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J. M. (2014). *WINSTEPS Rasch Measurement* (Version 3.81.0). Beaverton, OR: Winsteps.com.

Linacre, J. M. (2015). *Winsteps Rasch Measurement Computer Program User's Guide*. Beaverton, OR: Winsteps.com.

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument Translation Process: A Methods Review. *Journal of Advanced Nursing Research, 48,* 175-186. http://dx.doi.org/10.1111/j.1365-2648.2004.03185.x

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47,* 149-174. http://dx.doi.org/10.1007/BF02296272

Masters, G. N. (1988). The Analysis of Partial Credit Scoring. *Applied Measurement in Education, 1,* 279-297. http://dx.doi.org/10.1207/s15324818ame0104_2

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). New York: Macmillan.

Messick, S. (1995a). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice, 14,* 5-8. http://dx.doi.org/10.1111/j.1745-3992.1995.tb00881.x

Messick, S. (1995b). Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist, 50,* 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Ministry of Health (2014). Annual Report on Obesity and Physical Activity in Oman. http://www.moh.gov.om/en/nv_menu.php?o=hr/majorprojects.htm&SP=1.pdf

Mohammad-Hassun, A., & Ahmed-Kali, K. (2014). A Confirmatory Factor Analysis of the Urdu Version of the Physical Activity Self-Efficacy Scale for Adolescents. *Journal of Education in South Asia, 4,* 56-68.

Nakamura, Y. (2002). Beyond the Hijab: Female Muslims and Physical Activity. *Women in Sport Physical Activity Journal, 11,* 21-48.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill, Inc.

Pajares, F., & Miller, D. M. (1995). Mathematics Self-Efficacy and Mathematics Performance: The Need for Specificity of Assessment. *Journal of Counselling Psychology, 42,* 190-198. http://dx.doi.org/10.1037/0022-0167.42.2.190

Peterson, S., & Shavelson, T. (2014). The Adaptation of the Physical Activity Self-Efficacy Scale for Adolescents within a Singaporean Context. *South Pacific Psychology Bulletin, 2,* 45-60.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks Paedogogiske Insitut.

Simpson, S., & Rustin, F. (2012). Psychometric Properties of the Physical Activity Self Efficacy Scale for Adolescents in a Canadian Context. *Canadian Journal of Psychology and Education, 3,* 38-52.

Smith Jr., E. V. (2002). Detecting and Evaluating the Impact of Multidimensionality Using Item Fit Statistics and Principal Component Analysis of Residuals. *Journal of Applied Measurement, 3,* 205-231.

Smith, A., Rush, R., Fallowfield, L., Velikova, G., & Sharpe, M. (2008). Rasch Fit Statistics and Sample Size Considerations for Polytomous Data. *BMC Medical Research Methodology, 8,* 1-11. http://dx.doi.org/10.1186/1471-2288-8-33

Smith, R. M. (2003). *Rasch Measurement Models: Interpret in WINSTEPS/ BIGSTEPS and FACETS Output*. Maple Grove, MN: JAM Press.

Sousa, V., & Rojjanasrirat, W. (2011). Translation, Adaptation and Validation of Instruments or Scales for Use in Cross-Cultural Health Care Research: A Clear and User Friendly Guideline. *Journal of Evaluation in Clinical Practice, 17,* 268-274. http://dx.doi.org/10.1111/j.1365-2753.2010.01434.x

Stajkovic, A. D., & Luthans, F. (1998). Self-Efficacy and Work-Related Performance: A Meta-Analysis. *Psychological Bulletin*, *124*, 240-261. http://dx.doi.org/10.1037/0033-2909.124.2.240

Stout, W. (1987) A Nonparametric Approach for Assessing Latent Trait Unidimensionality. *Psychometrika, 52,* 589-617. http://dx.doi.org/10.1007/BF02294821

van der Linden, W. J., & Hambleton, R. K. (2010). *Handbook of Modern Item Response Theory*. New York: Springer.

Vertinsky, P., Batth, I., & Naidu, M. (1996). Racism in Motion: Sport, Physical Activity and the Indo-Canadian Female. *Avante, 2,* 1-23.

Warner, L. M., Parschau, L., Schwarzer, R., Wolff, J. K., Wurm, S., & Schuez, B. (2014). Sources of Self-Efficacy for Physical Activity. *Health Psychology, 33,* 1298-1308. http://dx.doi.org/10.1037/hea0000085

Wolfe, E. W., & Smith, E. V. (2007a). Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part I—Instrument Development Tools. *Journal of Applied Measurement, 8,* 97-123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II—Validation Activities. *Journal of Applied Measurement, 8,* 204-234.

World Health Organization (2014). Physical Activity. http://www.who.int/topics/physical_activity/en/

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago, IL: Mesa Press.

Wu, T. Y., Robbins, L. B., & Hsieh, H. F. (2011). Instrument Development and Validation of Perceived Physical Activity Self-Efficacy Scale for Adolescents. *Research and Theory for Nursing Practice, 25,* 39-54. http://dx.doi.org/10.1891/0889-7182.25.1.39

Yim, H. Y. B., Abd-El-Fattah, S. M., & Lee, L. W. M. (2007). A Rasch Analysis of the Teachers Music Confidence Scale. *International Education Journal, 8,* 260-269.