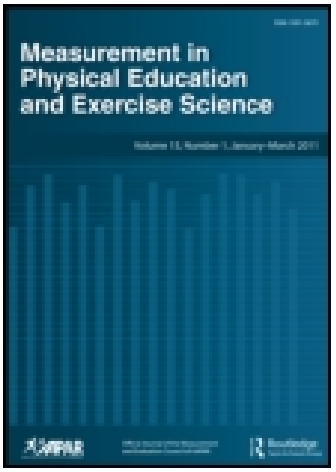


Measurement in Physical Education and Exercise Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmpe20

# Developing a Rasch Measurement Physical Fitness Scale for Hong Kong Primary School-Aged Students

Zi Yan [a] & Trevor G. Bond [b]
[a] Centre for Special Needs and Studies in Inclusive Education , Hong Kong Institute of Education , Tai Po, N.T., Hong Kong
[b] School of Education , James Cook University , Townsville, Queensland, Australia
Published online: 29 Jul 2011.

To cite this article: Zi Yan & Trevor G. Bond (2011) Developing a Rasch Measurement Physical Fitness Scale for Hong Kong Primary School-Aged Students, Measurement in Physical Education and Exercise Science, 15:3, 182-203, DOI: 10.1080/1091367X.2011.590772

To link to this article: http://dx.doi.org/10.1080/1091367X.2011.590772

*Measurement in Physical Education and Exercise Science*, 15: 182–203, 2011

ISSN: 1091-367X print / 1532-7841 online
DOI: 10.1080/1091367X.2011.590772

Routledge
Taylor & Francis Group

# Developing a Rasch Measurement Physical Fitness Scale for Hong Kong Primary School-Aged Students

Zi Yan

*Centre for Special Needs and Studies in Inclusive Education, Hong Kong Institute of Education, Tai Po, N.T., Hong Kong*

Trevor G. Bond

*School of Education, James Cook University, Townsville, Queensland, Australia*



The main purpose of this study was to develop a Rasch Measurement Physical Fitness Scale (RMPFS) based on physical fitness indicators routinely used in Hong Kong primary schools. A total of 9,439 records of students' performances on physical fitness indicators, retrieved from the database of a Hong Kong primary school, were used to develop the Rasch scale. Following a series of iterative Rasch analyses that adopted the "data should fit the model" approach, four physical fitness indicators (i.e., 6-min run, 9-min run, 1-min sit-ups, and dominant handgrip) were successfully calibrated to form the RMPFS. The RMPFS and its scale indicators showed fit to the Rasch model sufficient for the intended purposes of measuring the overall fitness of children. The overall physical fitness measure reflects children's fitness on three key core components of physical fitness (i.e., cardio-respiratory fitness, muscular endurance, and muscular strength). Advantages of the RMPFS are discussed, and recommendations for future research follow. The findings of this study provide a better knowledge basis for interpreting children's physical fitness assessment results.

Key words: Rasch measurement, physical fitness, primary school, data should fit the model

## INTRODUCTION

Given the important role that physical fitness should play in children's lives, fitness assessment/testing is intuitively a crucial part of physical education, which aims to promote a healthy and physically active lifestyle. However, fitness testing in schools has being criticized over decades, and even its necessity for children has been seriously questioned (Liu, 2008). The special issue on youth fitness testing published in *Measurement in Physical Education and Exercise Science* (*MPEES*) in 2008 thoroughly discussed different perspectives on youth fitness testing. For example, a pedagogical perspective argued that fitness tests should be implemented as formative evaluation. Then fitness testing results should be informative for teaching and learning in physical education (Silverman, Keating, & Phillips, 2008). In terms of promoting

Correspondence should be sent to Dr. Zi Yan, Centre for Special Needs and Studies in Inclusive Education, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, N.T., Hong Kong. E-mail: zyan@ied.edu.hk

physical activity, fitness assessments are expected to provide accurate measures, carrying important information about children's health-related fitness levels. Therefore, they could optimize the effectiveness of physical education (Welk, 2008). Moreover, there is no doubt that the use and interpretation of fitness assessment have important educational, pedagogical, and psychological consequences (Mahar & Rowe, 2008). In summary, the editors and authors of the *MPEES* special issue agreed that youth fitness testing can serve a useful purpose in school settings if used in the correct way. This article aims to extend this "correct way" discussion by shedding some light on how to achieve objective physical fitness measurement based on fitness testing scores.

Accurate measures of youth fitness are needed by both researchers and educators, regardless of their purposes (Mahar & Rowe, 2008). The routine practice in traditional approaches is that different components of physical fitness (e.g., body composition, cardio-respiratory fitness, flexibility, muscular endurance, and muscular strength) are assessed using different indicators, and children's abilities in each of these components are reported and interpreted using raw scores (in meters, kilograms, seconds, etc.) or percentile ranks. However, raw scores might not provide a valid *measure*, because they have little inferential value (Wright, 1997; Wright & Mok, 2000). The validity of raw scores in representing fitness levels in this approach is based on an unquestioned assumption; namely, the raw scores are accepted implicitly as being equal interval. Unfortunately, the raw scores themselves (unless used to derive further criterion measures, e.g., estimated $VO_2$max based on scores in the 6/9-min. run test) actually indicate only the ordering of the children's performances but have little inferential value about the size of the differences among scores in terms of "fitness." While meters indicate equal amounts of difference on the length or distance scales, it is an act of faith to conclude that meters indicate equal difference on the cardio-respiratory fitness scale. Meters have only ordinal meaning when they are used as the score units in the 6-min run test; therefore, they might not yield valid measures of the underlying fitness component.

Another deficiency associated with the traditional approaches to physical fitness assessment is that the interpretation of results of physical fitness assessment in norm-referenced framework is often not accurate or comparable because of the sample dependence and indicator dependence of assessments, where ranks or percentiles are provided in interpreting students' performances on physical fitness indicators. Those ranks or percentiles provide only an inexact basis for comparison among students and, rather, should be regarded as indicators of students' relative strengths and weaknesses (Williams, Harageones, Johnson, & Smith, 2000). However, use of raw numbers/counts and the allocation of norm-referenced ratings do not allow for the direct assessment of children's fitness against some objective fitness standard in which the measurement and interpretation of students' fitness levels is independent of sample and indicator.

Furthermore, it is time consuming to use the traditional approach to administering all fitness tests a whole class with 40 or more students. Since a single total score might not provide a meaningful summary of different fitness indicators, multi-faceted profiles that contain scores for each component of physical fitness are often regarded as more appropriate (Marsh, 1993). A consequent by-product is that assessment tasks in the physical education curriculum increase teachers' workloads and occupy resources that could be put into teaching. There is little doubt that physical fitness is a multi-faceted concept, but the extent to which any set of multi-dimensional indices used in traditional approaches should disqualify a uni-dimensional fitness index still remains open for discussion, as well as evidence-based empirical investigation. The question addressed in this article is to what extent is it possible to generate a uni-dimensional index of physical fitness,

which provides interval scale fitness measures for children, independent of sample and indicator, for estimating differences between groups of children and for tracking changes in fitness levels over time.

The Rasch model (Andrich, 1988; Rasch, 1960) provides ways to address the deficiencies inherent in traditional approaches to physical fitness assessment. First, Rasch analysis can transform non-linear raw scores into logit scale measures that have constant interval meaning and provide objective and linear measurement from ordered category responses (Linacre, 2000, 2006a, p. 12). Second, the feature of "parameter separation" or "invariance of parameters" (Bond & Fox, 2007, p. 71; Wright & Masters, 1982, p. 34) of the Rasch model implies that the calibration of fitness indicators is sample distribution free and the calibration of persons is indicator distribution free along the fitness continuum. The sample-distribution-free calibration of fitness indicators means that the difficulty estimates of indicators (e.g., 6-min run, 1-min sit-ups, etc.) should be invariant, within measurement error, no matter which sample is used to calibrate those indicators. The indicator-distribution free calibration of persons means that the fitness estimate of any person should remain invariant, within measurement error, no matter which particular fitness indicators are used to measure that person's fitness. Therefore, direct person–person, item–item, and person–item comparisons can be conducted easily, based on their locations on the common logit scale. Finally, an overall fitness measure can be provided for a student, even if he/she had not performed on all of the physical fitness indicators that have been calibrated onto the fitness trait continuum.

Unlike more general multi-dimensional or Item Response Theory (IRT) models and other (true score) statistical techniques that adopt a "the model fits the data" approach, manipulating the different parameters to accommodate the idiosyncrasies of any dataset, the Rasch model requires that "data fit the model" (Andrich, 2004) for the purpose of achieving objective measurement. This is one of the key differences between Rasch-based studies and other quantitative studies in the human sciences. The Rasch model is held as being able to solve the basic measurement problem common to all social sciences (Andersen, 1995), and it has been applied in sport sciences and physical education studies by a growing number of researchers whose reviews provide more detail (e.g., Strauss, Büsch, & Tenenbaum, 2007; Tenenbaum, Strauss, & Büsch, 2007). For example, Rasch analysis has been utilized to calibrate physical function or competence (Zhu & Kurz, 1994), perception of sports games (Kang & Kang, 2006), and difficulty levels of physical fitness indicators (Zhu & Safrit, 1993). Studies have applied the Rasch model to develop or evaluate instruments used in exercise studies. Hands and Larkin (2001) studied children's performance on different motor tasks and developed two separate uni-dimensional Rasch scales of motor abilities for boys and girls, respectively. Zhu, Timm, and Ainsworth (2001) modified an exercise barriers instrument and validated it using the Rasch model framework. Heesch, Masse, and Dunn (2006) used Rasch analysis to re-evaluate three commonly used scales, including the Physical Activity Enjoyment Scale, the Benefits of Physical Activity Scale, and the Barriers to Physical Activity Scale. Büsch et al. (2009) used a mixed Rasch model to investigate the construct validity of the German general motor fitness and coordination test for children. They found that two qualitatively different classes of children could be distinguished. Members of the first class were characterized by high running ability and low throwing ability, whereas members of the second class were characterized by low running and high throwing abilities.

Tenenbaum and colleagues (2007) claimed that the application of Rasch model in physical education and sport sciences is promising from both a methodological and a content-related

perspective. A number of advantages of Rasch model analyses have been echoed in previous studies. In calibrating a gross motor skills instrument with the many-faceted Rasch model, Zhu and Cole (1996) demonstrated the advantages of the Rasch model over the traditional norm-referenced interpretation, including benefits of parameter separation, sharing the same metric among items and examinees, and providing linear measures. They also pointed out that the person measures, together with Standard Error (SE) and fit statistics, provided useful diagnostic information to identify strengths and weaknesses of examinees. Bowles and Ram (2006) revealed that Rasch analyses of volleyball players' performances on three skills (serve, serve receive, and attack) produced an equal-interval scale that provided more objective and consistent information about volleyball players' abilities than could be obtained by traditional instruments. Zhu (2001) found that the Rasch model could accurately equate different motor function tests so that cross-test scores could be interpreted in a common measurement framework, an important outcome that remains unachievable in traditional approaches to motor function assessment. Büsch and Strauss (2005) used the Rasch measurement model to study 503 participants' performances on 6 gross-motor coordination tasks, categorized as precision and time-pressure tasks. They found that persons performing gross-motor coordination tasks could be differentiated based on the coordination strategy they used. The results displayed the advantages of Rasch model in identifying strategies used by persons in completing gross-motor tasks and distinguishing between person and item characteristics.

The Rasch model has also been applied in attempts to combine closely related scales to assess single uni-dimensional physical functioning constructs. An interesting study conducted in the health care domain combined two separate but related scales into one uni-dimensional scale (Hsueh, Wang, Sheu, & Hsieh, 2004). The 10-item Barthel index (BI) assessing activities of daily living (ADL) and the 15-item Frenchay activities index (FAI) assessing instrumental ADL were administered to 245 patients one year after stroke. The data from these two scales were combined and analyzed using the Rasch model, and the result indicated that all but two FAI items fit the uni-dimensional Rasch model very well, indicating that the BI and the FAI assess a single underlying uni-dimensional ADL construct. Further analyses of the 23-item uni-dimensional scale revealed that it had quite high person reliability (.94) and that the range of item difficulties was well targeted to the patient sample. A "look-up" conversion table was then offered to transform combined BI and FAI raw scores into Rasch ADL interval measures. Thus, a clinically useful instrument was developed by combining the BI and the FAI scales, and the new scale had improved range and sensitivity for assessing comprehensive ADL function.

However, this kind of combining attempt is seldom found in physical education literature. Traditional approaches conceptualize physical fitness as a multi-faceted construct, hence, psychometrically multi-dimensional. However, a single *overall* fitness score is still preferable, even necessary, in many situations, especially for the interpretation of students' comprehensive physical ability. In most cases, an overall fitness score is obtained by simply summing or averaging the scores for different components of physical fitness. It is obvious that such averaged overall fitness scores lose quite large amounts of information about specific fitness aspects, and therefore, one should be very cautious in interpreting that kind of overall score (Fleishman, 1964). Furthermore, as argued by Büsch et al. (2009), item homogeneity must be checked before using a sum score to estimate general fitness. Marsh (1993) recommended constructing, if necessary, a weighted summary score that assigns an optimal weight to each component based on theoretical and empirical research. But it remains a considerable challenge to derive optimal weights for different

fitness components, because the weights might need modification according to particular criteria, particular research purposes, or even the predisposition of the particular investigator.

Thus, the main purpose of this study was to develop, to the extent that it was both useful and possible, a Rasch Measurement Physical Fitness Scale (RMPFS), combining all, or at least some, of the indicators routinely used in Hong Kong primary schools. A successful scale would then calibrate person ability (students' overall physical fitness levels) and item difficulty (difficulty levels of each of the physical fitness indicators) in a single, stable fitness measurement framework. Given the review of the quantitative approaches open for adoption in such a research project, the position taken in this research is the primacy of the requirement to produce scientifically repeatable measures based on the principles espoused in Rasch measurement. As a consequence, this particular research explicitly adopts the Rasch "data-fit-the-model" approach for the empirical investigation of the construction of physical fitness measures for children.



## METHOD

### Data

Fitness data used in this study were retrieved from the physical fitness assessment records database of a large, regional Hong Kong primary school, a government-subsidized primary school located in the northeastern new territories of Hong Kong. This school routinely has five classes at each year level from primary 1 (6 years old) to primary 6 (12 years old) with an annual enrollment of over 1,000 students.

The dataset covers this school's students' physical fitness records for the academic years 2002–2003 to 2006–2007. There are two rounds of students' records for each academic year, except 2002–2003, for which the records for the second semester were not entered into the school's database. Initially, 10,512 student records were included in the potential data pool for this study, and finally, 9,439 records were kept for scale development after excluding exceptional and unreasonably extreme data. It is worth pointing out that each record does not necessarily refer to an independent student, since this is a longitudinal dataset over five years, and most students would have several records (potentially up to nine) over time in the dataset. Of the records, there are 5,149 records (54.6%) for males and 4,290 records (45.4%) for females. The age range for all records extends from 6 to 13 years ($M = 8.53$, $SD = 1.73$). Only four records did not include age information. The details of the sample used in this study are presented in Table 1.

### Physical Fitness Indicators

The partner school of this study administers the physical fitness testing recommended by the School Physical Fitness Award Scheme (Hong Kong Education and Manpower Bureau, 2005) except that body composition is estimated by BMI and not the skinfold method, because the equipment required for skinfold testing is not available in this school. The other eight fitness indicators include the 6-min run, the 9-min run, 1-min sit-ups, standard push-ups, modified push-ups, right handgrip, left handgrip, and sit-and-reach. It is worth noting that the 6-min run test is administered to students in grades 1 to 3 only, and the 9-min run test is administered to students

TABLE 1
Details of the Sample

| *Academic year* | *2002–2003* | | *2003–2004* | | *2004–2005* | | *2005–2006* | | *2006–2007* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Semester* | *First* | *Second* | *First* | *Second* | *First* | *Second* | *First* | *Second* | *First* | *Second* | *Total* |
| Male | 510 | 0 | 556 | 551 | 572 | 574 | 592 | 590 | 606 | 598 | 5,149 |
| Female | 458 | 0 | 472 | 468 | 492 | 489 | 488 | 487 | 468 | 468 | 4,290 |
| Total | 968 | 0 | 1,028 | 1,019 | 1,064 | 1,063 | 1,080 | 1,077 | 1,074 | 1,066 | 9,439 |
| *Age* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | *13* | *Missing* | | *Total* |
| Male | 837 | 900 | 877 | 845 | 813 | 779 | 94 | 4 | 0 | | 5,149 |
| Female | 666 | 701 | 727 | 742 | 717 | 672 | 61 | 0 | 4 | | 4,290 |
| Total | 1,503 | 1,601 | 1,604 | 1,587 | 1,530 | 1,451 | 155 | 4 | 4 | | 9,439 |



in grades 4 to 6 only. The standard push-ups test is administered only to male students in grades 3–6, and the modified push-ups test is administered to all students in grades 1 and 2, as well as female students in grades 3 to 6.

## Data Analysis

The software package used for Rasch analyses in the present study is WINSTEPS 3.0 (Linacre, 2006a); the partial credit model (PCM) was specified for these analyses.[1] The PCM is a sound option for Rasch analyses with the physical fitness data in this study, considering that the definition of the rating scale is unique for each of the physical fitness indicators. The "partially correct response(s)" between incorrect and completely correct item responses provided in the PCM are in accordance with the different levels of performances between minimum and maximum raw scores on each physical fitness test.

## Logarithmic Transformation of Raw Scores

Although it is not a problem for WINSTEPS to handle data with a large number of ordered response categories (it accommodates up to 255 category levels per item), using many more than the necessary category levels is likely to introduce challenges to the meaningful interpretation of the results. From a practical perspective, it is unlikely that primary school-aged students'

[1] Although the Rasch Poisson Counts Model has been used to measure physical fitness (e.g., Zhu & Safrit, 1993), the appropriateness of the Rasch Poisson Counts Model in time-limited psychomotor performances, such as 1-min sit-ups test scores, is dubious, because some model requirements are not satisfied by such data (Zhu & Safrit, 1993). For example, the Rasch Poisson Counts Model assumes that examinees should complete the repetitions at a constant speed through the whole performance. However, the effect caused by fatigue in the 1-min sit-ups test violates this basic assumption; repetition speed is usually slower and slower as examinees complete greater numbers of sit-ups during the 1-min period.

performances on physical fitness indicators have more than about ten useful qualitatively different levels; it is unlikely that 10 m in a 6-min run test or 1 cm in a sit-and-reach test indicate meaningful differences in overall physical fitness levels, even if such a small difference could move a child's fitness estimate from a lower to a higher response category for that one indicator. Thus, re-expressing raw data into a reasonable number of ordered categories would help the interpretation of the results and the detection of departures from fit to the model more clearly (Linacre, 2000). A Poisson logarithmic transformation was used to transform the raw scores into a dataset with more even distribution and more meaningful category structure. The transformation can be expressed as

$$\text{Scored category} = 1 + 8^* \frac{\log(\text{observation} + 1) - \log(L+1)}{\log(H + 1) - \log(L+1)}, \quad (1)$$

where $L$ is the lowest value of the observations, and $H$ is the highest value of the observations. The number 8 was chosen just because after some initial investigation analyses, a nine-category structure was selected as the appropriate transformation target.

## Iterative Sequence of Analytical Steps

Given that the RMPFS would be developed from a Rasch measurement perspective, each fitness indicator where data violated Rasch measurement requirements was excluded, in turn, from the scale. More specifically, this study took the strong data-fit-the-model approach in developing this physical fitness scale. Seven criteria were utilized in the procedure of scale development to investigate the quality of those indicators and to decide whether an indicator should be retained in or excluded from subsequent analyses.

- *Investigations from a practical perspective*. Practical considerations undertaken before undertaking statistical analyses might uncover some important factors detrimental to scale development.
- *Fit statistics for indicators*. Since mean square (MNSQ) statistics are relatively more stable than their standardized forms (ZSTD) in the rating scale model and PCM analyses (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008), MNSQs are used as fit criteria for scale quality assurance.
- *Point-measure correlations for indicators*. The point-measure correlation coefficient of an indicator refers to the correlation between the estimate for any particular fitness indicator and the overall measure of the fitness trait under measurement (Linacre, 2006b). Normally, a point-measure correlation coefficient higher than +.4 indicates acceptable consistency of indicator polarity in a scale.
- *Rasch reliability*. Rasch measurement provides both person reliability and item reliability indices. Rasch person reliability refers to the consistency of person ordering along the trait continuum measured by the scale (Smith, 2001; Wright & Masters, 1982, p. 114), and Rasch item reliability indicates replicability of item placements along the trait continuum if the same set of items were administered to another similar sample of persons (Bond & Fox, 2007, p. 41).

- *Variance explained by measures*. Variance explained by measures refers to the proportion of variance of observations that could be explained by the item difficulties, person abilities, and rating scale structures in Rasch analyses (Linacre, 2006a, p. 221). A higher proportion of variance explained by Rasch measures means that the Rasch model has better capacity for predicting performances of both items and persons.
- *Response category structure*. Successful implementation of polytomous Rasch measurement requires well-functioning performance categories for each indicator in the scale.
- *Influence of under-fitting persons*. The impact of extremely mis-fitting persons, especially under-fitting (erratically performing) persons, on fitness scale quality would be investigated.



## RESULTS

Through a theory-driven iterative developmental procedure guided by Rasch model measurement perspectives as well as practical considerations, eventually four physical fitness indicators, including the 6-min run (R6), 9-min run (R9), 1-min sit-ups (SU), and dominant (not left or right) handgrip (DH) were successfully calibrated to form the RMPFS, thereby integrating three key components of physical fitness—cardio-respiratory fitness, muscular endurance, and muscular strength—to provide a single person measure of overall physical fitness suitable for use with primary school children in Hong Kong.

The indicator properties of the RMPFS are presented in Table 2. The difficulty levels (i.e., item measures) for the four fitness indicators range from –1.59 logits to +1.25 logits, associated with standard errors of .02 or .03 logits. These small standard errors imply that the indicator difficulty estimations are quite precise, primarily due to the large calibration sample. The infit and outfit MNSQs range from .85 to 1.13, indicating sufficient fit to the Rasch model for practical measurement purposes, especially for such low-stakes decisions as monitoring children's fitness levels in school settings. The point-measure correlations approximate .8, supporting the claim that all the indicators function in the same direction as a part of the physical fitness latent trait under measurement. The Rasch item reliability is 1.00, and the Rasch person reliability is lower, but acceptable, at .77, a consequence of retaining only four indicators in the RMPFS.

Figure 1 presents the Wright map of the 4-indicator RMPFS. Students are placed on the left side of the scale according to physical fitness, and the fitness indicators are shown on the right

TABLE 2
Scale Properties of the RMPFS

| | *Measure (Logits)* | *S.E.* | *Infi MNSQ* | *Outfit MNSQ* | *Point-Measure Correlation* |
|---|---|---|---|---|---|
| MTB<6-min Run | −0.61 | 0.03 | 0.93 | 0.96 | .78 |
| 9-min Run | 1.25 | 0.03 | 0.85 | 0.88 | .86 |
| 1-min Sit-ups | −1.59 | 0.02 | 0.95 | 1.00 | .79 |
| Dominant handgrip | 0.96 | 0.02 | 1.11 | 1.13 | .79 |
| Person | Separation: | 1.83 | Reliability: | 0.77 | |
| Item | Separation: | 43.16 | Reliability: | 1.00 | |

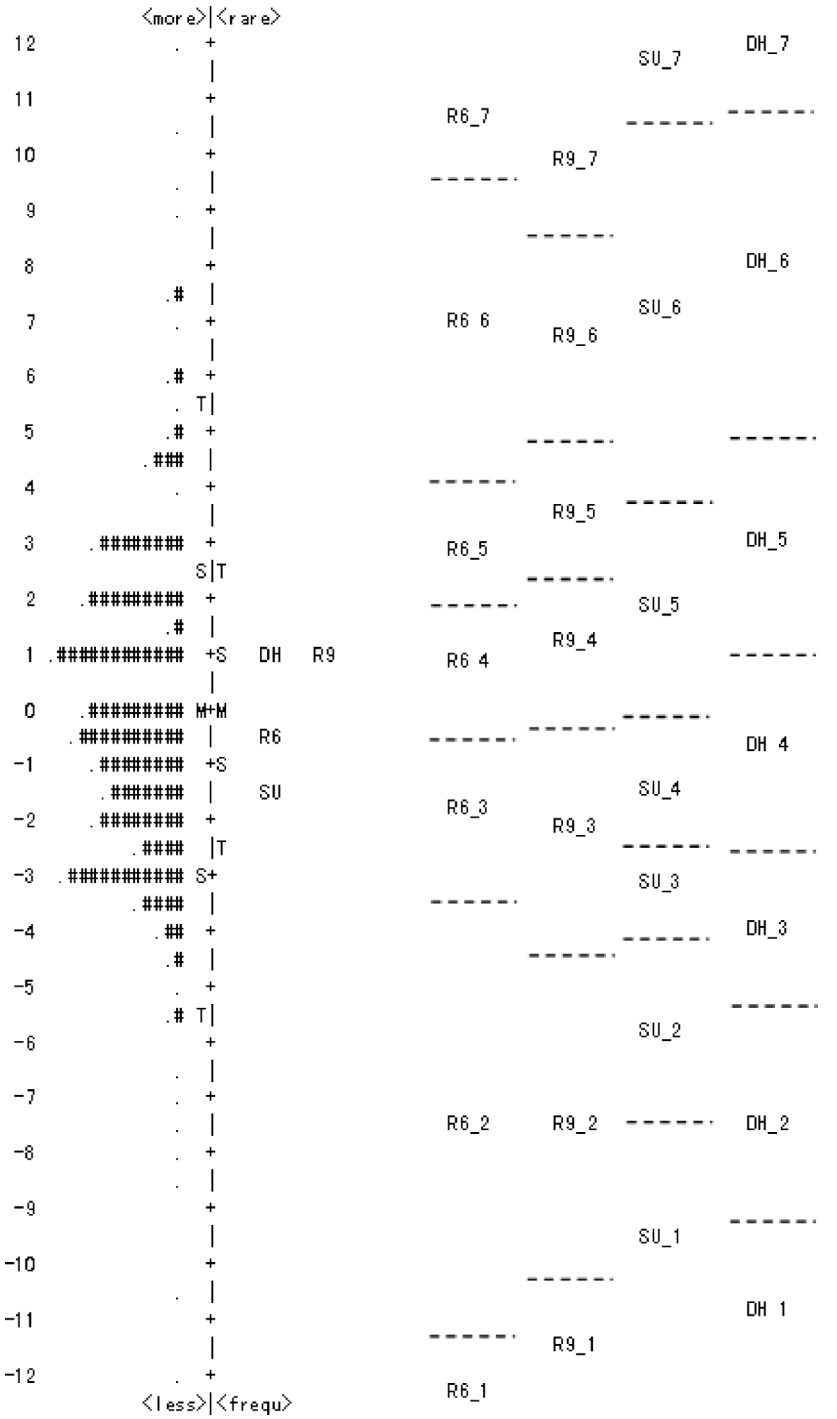


FIGURE 1 Wright map of the RMPFS. Each # represents 64 children (R6: 6-min run, R9: 9-min run, SU: 1-min sit-ups, DH: dominant handgrip).

side. The students with the highest fitness levels and the fitness indicators with highest difficulty levels are located at the top of the map, while the students with the lowest fitness level and the easiest fitness indicators are located at the bottom. The means of student measures and indicator difficulty calibrations are shown as the corresponding *M*s on the map. *S* and *T* represents ±1 and ±2 standard deviations of the student and indicator distributions, respectively. It can be seen that the difficulty levels of the RMPFS physical fitness indicators ($M = .00$, $\underline{SD} = 1.16$) are appropriate for these students' fitness levels ($M = -.21$, $SD = 2.78$). The range of indicators' difficulty (–1.59 to 1.25 logits) is much smaller than the range of students' ability (–12.86 to 11.17 logits). However, the ranges of difficulty levels of the response categories (categories 1 to 7) for each indicator, as presented on the right-hand side of the map (from R6_1 at −11.31 logits to DH_7 at +10.94 logits), reveals that the indicators, overall, provide good coverage of the fitness of the primary school-aged students in this sample.

The item characteristic curves (ICCs) and category probability curves provide further support for the valid functioning of the scale. Figure 2 presents the empirical and expected ICCs for the four indicators. It can be seen that the empirical ICCs match the theoretical ICCs reasonably well, especially for students' with median fitness levels located around the middle of the curves. There are larger discrepancies between the empirical and theoretical ICCs for the most able and least able students located at the extremes of the curves.

The category probability curves for each of the four indicators presented in Figure 3 show that each performance category has a distinct peak in the graph for all indicators. That means each category for each indicator was the most probable performance level for given groups of persons with a specific level of physical fitness. There is no evidence of category threshold disordering (Linacre, 2002), and the threshold calibrations advance monotonically with category, indicating that higher performance categories correspond to higher measures of physical fitness.

## DISCUSSION

Initially, a total of nine physical fitness indicators were used to develop the RMPFS, but only four were retained to form the final interval-level measurement scale. The other five indicators were excluded or replaced in the development procedure based on the seven criteria described in the "Method" section. The development procedure is presented in Table 3. Scales 1 to 5 displayed in Table 3 are intermediate scales before Scale 6—the final version of RMPFS—was finally established.

### Consideration of BMI

BMI was excluded for conceptual and practical reasons. First, BMI is a rough index appropriate for reporting adiposity at the population level but not optimal for use with individuals because of prediction error (Heyward, 2002, p. 183; Stratton & Williams, 2007). Second, BMI is a trait with an inverted U-shaped (∩) distribution. A higher BMI score does not necessarily stand for a better level of physical fitness, nor does a lower BMI score. This is a distinctive feature that sets BMI apart from other fitness indicators. Combining BMI together with other indicators in the Rasch measurement scale would contradict one of the requirements of Rasch model: all items in the same scale should function in the same (linear) direction along the latent trait under measure.

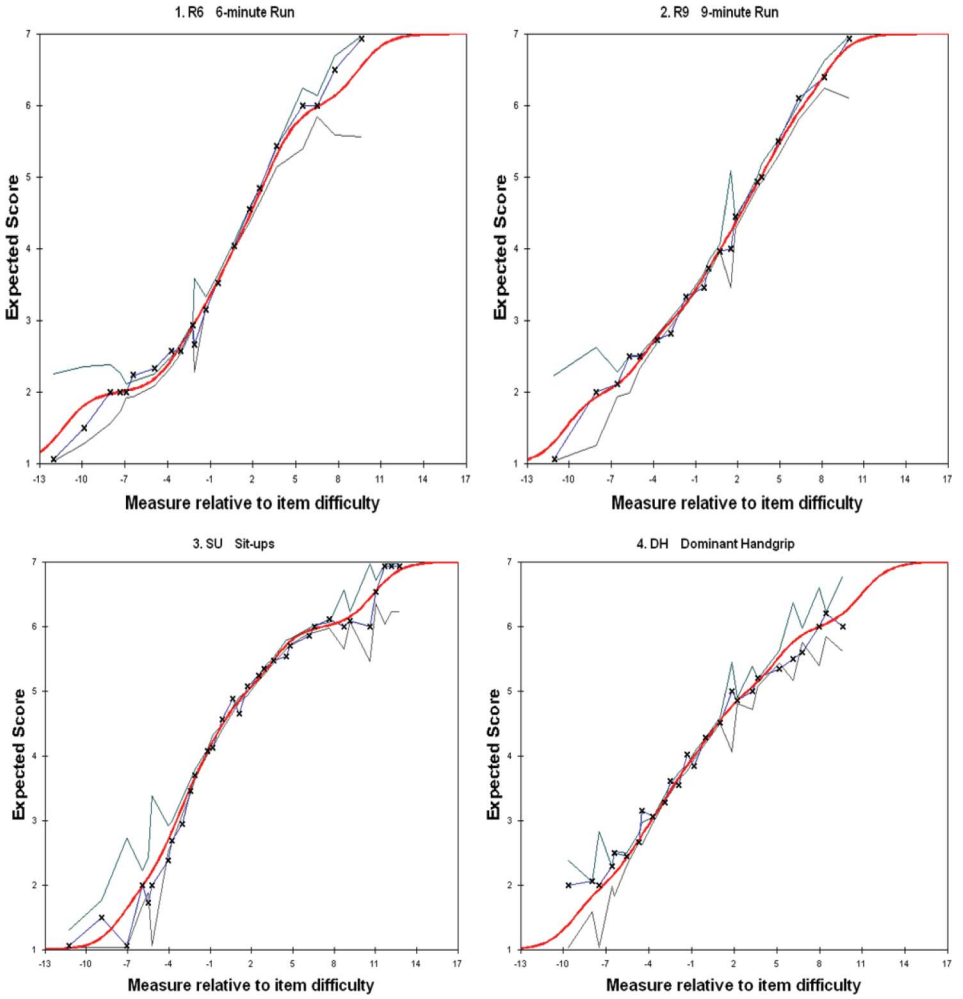


FIGURE 2 Empirical (blue) and expected (red) ICCs for RMPFS indicators (R6: 6-min run, R9: 9-min run, SU: 1-min sit-ups, DH: dominant handgrip) (color figure available online).

## Consideration of Sit-and-Reach

Sit-and-reach, which is used to assess flexibility, is distinct from the other indicators in some important ways. Students' performances for other indicators increase monotonically with students' age, but this is not the case for sit-and-reach. Furthermore, the correlation matrix shows that the flexibility component has relatively low correlations with other indicators of physical fitness. This is consistent with the findings of other studies; Marsh and Redmayne (1994) found that the correlations involving the flexibility component are smaller than the correlations involving other components of physical fitness. Therefore, sit-and-reach was excluded from the RMPFS to see if there was any subsequent improvement in scale properties. The results in Table 3 show

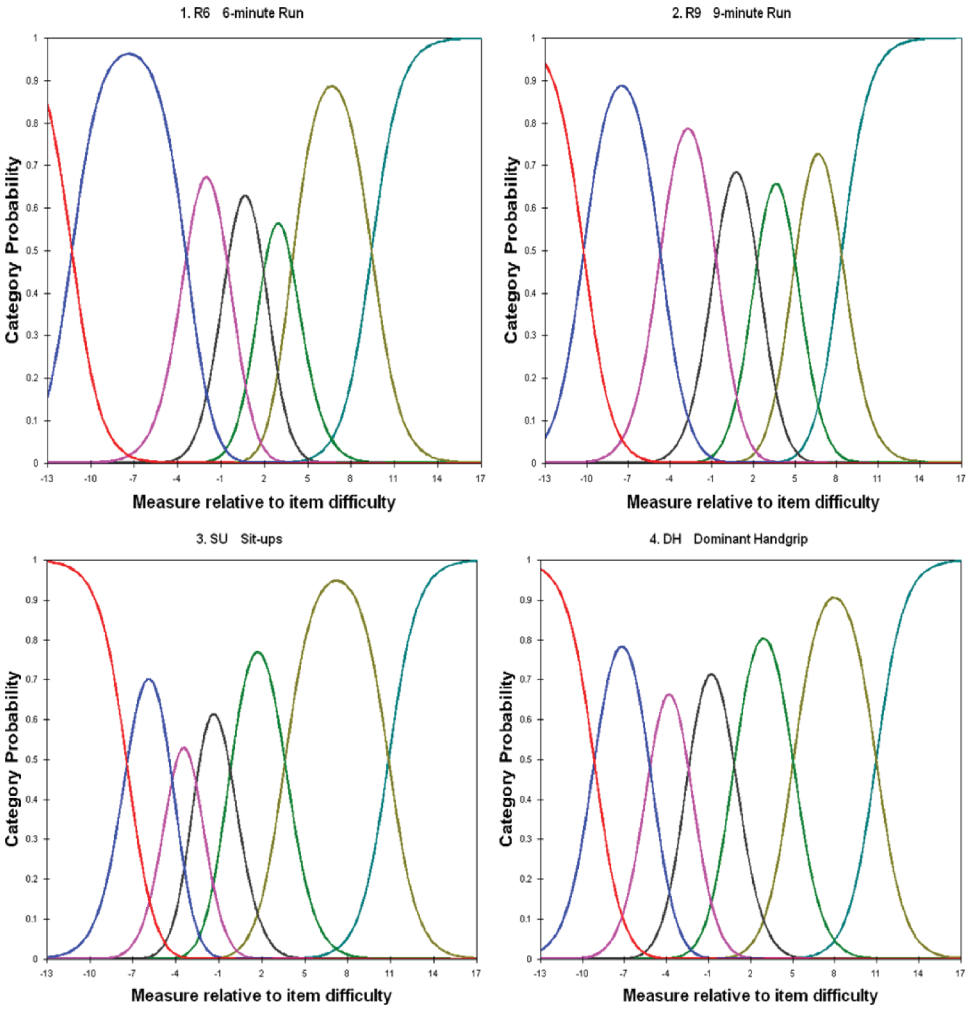


FIGURE 3 Category probability curves for RMPFS indicators (R6: 6-min run, R9: 9-min run, SU: 1-min sit-ups, DH: dominant handgrip) (color figure available online).

that the Rasch person reliability increased appreciably from .52 to .66, even though the raw score range of the scale was reduced.

## Consideration of Handgrip

Rasch factor analyses of fit residuals show that right handgrip and left handgrip have quite high loadings on the first contrast factor, and the correlation between their residuals is .52; i.e., they share about 27% of their variance in common. That suggests there is probably a separate fitness sub-dimension comprising of right handgrip and left handgrip and that there is local dependency

TABLE 3
Developmental Procedure for RMPFS

| | *Infit MNSQ* | *Outfit MNSQ* | *Point-Measure Correlation* | *Rasch Reliability (Person/Item)* | *Variance Explained by Measures* |
|---|---|---|---|---|---|
| Eight indicators, nine categories (Scale 1) | | | | | |
| 6-min Run | 1.03 | 1.03 | .58 | .52/1.00 | 62.1% |
| 9-min Run | 1.09 | 1.09 | .65 | | |
| 1-min Sit-ups | 0.93 | 0.91 | .63 | | |
| Right handgrip | 0.76 | 0.75 | .73 | | |
| Left handgrip | 0.74 | 0.72 | .73 | | |
| Sit-and-reach | 1.21* | 1.27* | .42 | | |
| Standard push-ups | 1.01 | 1.01 | .70 | | |
| Modified push-ups | 1.49* | 1.48* | .47 | | |
| Seven indicators, nine categories (Scale 2) | | | | | |
| 6-min Run | 1.10 | 1.10 | .61 | .66/1.00 | 60.6% |
| 9-min Run | 1.15 | 1.15 | .68 | | |
| 1-min Sit-ups | 1.07 | 1.05 | .64 | | |
| Right handgrip | 0.78 | 0.78 | .76 | | |
| Left handgrip | 0.75 | 0.75 | .76 | | |
| Standard push-ups | 1.01 | 1.02 | .74 | | |
| Modified push-ups | 1.57* | 1.58* | .51 | | |
| Six indicators, nine categories (Scale 3) | | | | | |
| 6-min Run | 0.93 | 0.92 | .70 | .60/1.00 | 62.6% |
| 9-min Run | 0.95 | 0.95 | .75 | | |
| 1-min Sit-ups | 0.90 | 0.90 | .69 | | |
| Dominant handgrip | 1.11 | 1.10 | .65 | | |
| Standard push-ups | 0.88 | 0.88 | .78 | | |
| Modified push-ups | 1.26* | 1.27* | .6 | | |
| Four indicators, nine categories (Scale 4) | | | | | |
| 6-min Run | 0.92 | 0.91 | .73 | .63/1.00 | 66.9% |
| 9-min Run | 0.90 | 0.90 | .79 | | |
| 1-min Sit-ups | 0.97 | 0.98 | .70 | | |
| Dominant handgrip | 1.09 | 1.08 | .70 | | |
| Four indicators, seven categories (Scale 5) | | | | | |
| 6-min Run | 0.92 | 0.95 | .72 | .62/1.00 | 68.7% |
| 9-min Run | 0.90 | 0.91 | .79 | | |
| 1-min Sit-ups | 0.95 | 0.99 | .73 | | |
| Dominant handgrip | 1.10 | 1.10 | .71 | | |
| Four indicators, seven categories without under-fitting persons (Scale 6/RMPFS) | | | | | |
| 6-min Run | 0.93 | 0.96 | .78 | .77/1.00 | 81.5% |
| 9-min Run | 0.85 | 0.88 | .86 | | |
| 1-min Sit-ups | 0.95 | 1.00 | .79 | | |
| Dominant handgrip | 1.11 | 1.13 | .79 | | |

*Misfitting item.

between these two indicators. From the Rasch perspective, one promising solution is to use dominant handgrip instead of right handgrip and left handgrip. In this case, the higher score of right handgrip and left handgrip was chosen as the dominant handgrip result for each student. As well as for right and left handgrip, local dependence is likely to occur between the 6- and

9-min run or between standard push-ups and modified push-ups, considering their very similar nature. However, this is not a concern for the current analyses, since no single case in the dataset has scores on both the 6-min run and 9-min run or standard push-ups and modified push-ups.

## Consideration of Push-Ups

The properties of Scale 3 (see Table 3) show that the standard push-ups and modified push-ups have poor fit to the Rasch model. There are two reasons that probably introduce noise to these two indicators. The first, these two indicators are usually used for secondary school-aged students in Hong Kong but not for primary school-aged students. The partner school of this study used these two indicators just as supplementary tests for a small portion of students (14.4% for standard push-ups and 20.2% for modified push-ups) before academic year 2005–2006. The second reason is related to the nature of the push-ups test. These two tests have no time limit but have an assumption about students' willingness to participate; i.e., students were assumed to try their best to complete as many push-ups as possible (until they cannot do any more). But this does not always seem to be the case in practice, especially for supplementary tests to which students often attach less importance. Considering the misfit shown by these two indicators and the possibility of measurement noise introduced by them, it is reasonable to exclude them from further RMPFS development. As indicated in Table 3, the properties of a 4-indicator scale (Scale 4) are much better than those of previous versions. The Rasch person reliability increased from .60 to .63. The variance explained by measures increased considerably from 62.6% to 66.9%.

## Optimizing Response Category

The results of Rasch analyses adopting 9-category dataset showed that the response category structure was not optimal because (a) the distribution of respondents among categories was not even, (b) there were some reversed average measures and threshold calibrations, and (c) the category probability curves for some categories were submerged by others. Therefore, it is appropriate to collapse some adjacent and potentially redundant categories in order to obtain a meaningful and interpretable category structure for each indicator. Two principles were followed in the process of combining adjacent categories. The first was to ensure that each category had a reasonable number of respondents, and the second was to ensure that average measures for categories and threshold difficulties increase monotonically and with reasonable increments. Finally, a 7-category structure was developed, and the category functioning effectiveness was examined in detail according to the guidelines suggested by Linacre (2002). The point-measure correlation coefficients of all four indicators range from .71 to .79. The number of observations of all categories for each indicator ranges between 14 and 3,496 with a mean of 879. The observation distributions across categories for all indicators are uni-modal distributions, peaking in a central category and showing smooth decreases to categories 1 and 7, respectively. The average measures of categories for all indicators advance monotonically with category. The outfit and infit MNSQs for all categories range between .79 and 1.44, and most of them are very close to 1.0. The threshold calibrations of categories for all indicators advance monotonically. The measure-to-category coherence and category-to-measure coherence for most of the categories except categories 1 to 7 are acceptable. The distances between adjacent threshold calibrations are all larger than 1.0 logit and less than

5.0 logits with only two exceptions. Therefore, Scale 5 based on the 7-category data replaced Scale 4, which had been constructed from 9-category data (see Table 3).

## Influence of Under-Fitting Persons on the RMPFS

Verhelst and Glas (1995) stated that there are two methods to improve Rasch measurement scale construction. The one is to eliminate "bad" items; the other is to exclude temporarily some test takers whose performances do not fit the Rasch model. At this point, eliminating further items from the scale is not the preferable option because only four indicators are retained in Scale 5 (see Table 3), and all of them are "acceptably good" items from both practical and Rasch measurement perspectives. Consequently, the alternative—temporarily eliminating misfitting persons who introduce unexpected noise to the measurement—was carried out in order to investigate possible improvement in the measurement characteristics of the scale. Bond and Fox (2007) pointed out that under-fitting persons (MNSQ » 1.0) are more detrimental to calibrating a measurement scale than are over-fitting persons (MNSQ « 1.0). Linacre (2002) further stated that MNSQs higher than 2.0 indicate more noise than useful information provided by the observations. Consequently, persons were excluded from the scale construction if either their outfit MNSQ or infit MNSQ was higher than 2.0 on Scale 5. Finally, a total of 1,185 cases were excluded, and the final version of the RMPFS was established based on the retained 8,469 cases, which had at least 1 score for any of the 4 indicators (6-min run, 9-min run, 1-min sit-ups, and dominant handgrip). The results in Table 3 show that the scale constructed without under-fitting persons exhibits significant improvement in both Rasch person reliability (increased from .62 to .77) and variance explained by measures (increased from 68.7% to 81.5%).

## Properties of the RMPFS with Subsamples

As described before, the data used to develop the RMPFS came from a longitudinal dataset collected over five years, and most students would have several records over time in the dataset. That means some records might be considered as dependent on each other, as they are the performances of the same student at different time points. The reason for including all data in the primary calibration analysis was to develop the RMPFS with as much good quality data as possible. At this point, the concerned reader might have some reservations due to the nature of the complete sample. Rasch modeling requires performances from persons who are independent of each other. Otherwise, the attractive feature of sample-distribution-free measurement as well as the property of local independence of Rasch models might be lost. In the complete analysis, each record has been treated as that of an individual person. To rule out any concern in that regard, separate Rasch analyses has been completed for the 4-item test (RMPFS) using subsamples in which each person has only one record (e.g., potentially for the nine separate subsamples for each measurement point in time (approximately 1,000 pupils each). This approach examines the robustness of the RMPFS to these apparently dependent records. Eventually, six subsamples with records for all four RMPFS indicators were used. The subsamples for the first semester of academic year 2002–2003, the second semester of academic year 2003–2004, and the second semester of academic year 2004–2005 were excluded, since the records for one or more RMPFS indicators were missing. The results are presented in Table 4.

TABLE 4
Scale Properties of the RMPFS for Subsamples

| *Sample* | *Measure (Logits)* | *S.E.* | *Infit MNSQ* | *Outfit MNSQ* | *Point-Measure Correlation* | *Person Reliability/Separation* | *Item Reliability/Separation* |
|---|---|---|---|---|---|---|---|
| Overall | | | | | | | |
| R9 | 1.25 | .03 | 0.85 | 0.88 | .86 | .77/1.83 | 1.00/43.16 |
| DH | 0.96 | .02 | 1.11 | 1.13 | .79 | | |
| R6 | −0.61 | .03 | 0.93 | 0.96 | .78 | | |
| SU | −1.59 | .02 | 0.95 | 1.00 | .79 | | |
| 2003/1 | | | | | | | |
| R9 | 2.63 | .08 | 0.79 | 0.81 | .85 | .78/1.87 | 1.00/22.27 |
| R6 | −0.06 | .08 | 0.90 | 0.99 | .75 | | |
| DH* | −1.13 | .06 | 1.12 | 1.12 | .78 | | |
| SU | −1.44 | .06 | 0.93 | 0.97 | .77 | | |
| 2004/1 | | | | | | | |
| R9 | 2.33 | .09 | 0.80 | 0.84 | .86 | .80/2.03 | 1.00/20.54 |
| DH | −0.18 | .06 | 0.99 | 1.00 | .82 | | |
| R6 | −0.31 | .08 | 1.01 | 1.04 | .72 | | |
| SU | −1.84 | .06 | 0.98 | 1.05 | .76 | | |
| 2005/1 | | | | | | | |
| R9 | 2.15 | .09 | 0.87 | 0.89 | .82 | .76/1.78 | 1.00/21.41 |
| DH | 0.45 | .06 | 1.00 | 1.01 | .81 | | |
| R6 | −0.45 | .08 | 0.96 | 1.00 | .74 | | |
| SU | −2.14 | .06 | 1.01 | 1.00 | .76 | | |
| 2005/2 | | | | | | | |
| R9 | 3.37 | .08 | 0.74 | 0.75 | .87 | .80/2.02 | 1.00/27.82 |
| DH | −0.45 | .07 | 1.16 | 1.19 | .77 | | |
| R6 | −0.88 | .07 | 0.85 | 0.85 | .79 | | |
| SU | −2.03 | .06 | 0.99 | 1.01 | .78 | | |
| 2006/1 | | | | | | | |
| R9 | 1.56 | .09 | 0.82 | 0.83 | .86 | .76/1.77 | 1.00/18.89 |
| DH | 0.93 | .06 | 1.11 | 1.15 | .74 | | |
| R6 | −0.46 | .08 | 0.99 | 1.04 | .69 | | |
| SU | −2.03 | .06 | 0.92 | 1.01 | .74 | | |
| 2006/2 | | | | | | | |
| R9 | 2.04 | .09 | 0.88 | 0.89 | .83 | .77/1.85 | 1.00/16.82 |
| DH | 0.11 | .07 | 1.10 | 1.12 | .79 | | |
| R6 | −0.74 | .08 | 0.88 | 0.92 | .78 | | |
| SU | −1.42 | .07 | 0.97 | 1.00 | .78 | | |

*Note:* R6: 6-min run, R9: 9-min run, SU: 1-min sit-ups, DH: dominant handgrip.
*Indicator out of order.

It can be seen from Table 4 that the properties of the RMPFS with each of the six independent subsamples are quite good. The infit and outfit MNSQs range from .74 to 1.19. The point-measure correlations range from .69 to .87. The Rasch item reliability is 1.00, and the Rasch person reliability approximates .8. The standard errors of item estimates are, although still quite small, slightly larger than those derived from the overall data due to the decrease of the sample sizes. The ordering of the item difficulty is consistent with that for the overall sample with only one exception (2003-1: the subsample from the first semester of academic year 2003–2004) for which dominant handgrip appeared to be easier than the 6-min run, whereas that is not the

case for other subsamples. Given that the item ordering remained invariant except in this one instance, the concern about potential lack of person independence in the results of the analysis for the whole dataset can be put to one side.

### Age Dependent or Age Related?

At this point, it could be easy to conclude that the RMPFS is merely reflecting changes in children's body and fitness that are determined by their age. Figure 4(A) reflects the differences in fitness levels, on average, between boys and girls at each age levels from 6 to 11 years of age. While the differences and trends are relatively consistent, the overlap across sexes and age groups remains quite substantial. Close inspection of Figures 4(B) and 4(C) reveals the subtle changes that are not easily discerned in the summary tables. The inference to be drawn from these graphs is, rather, that increases in children's fitness is merely related to, but not actually determined by or dependent on, children's age. If it were the case that physical fitness in children was age dependent rather than age related, it should be possible to determine any child's age by his/her location on the fitness scale or his/her fitness score simply referring to his/her age. However, figures showing the full distribution of fitness scores by ages for boys and girls reveal that this is clearly not the case. A boy with an RMPFS measure of 3.17 logits might be an 11 year old of above average fitness for his age, a 9 year old who is a little fitter than average, or, indeed, the fittest boy in the first year of primary school, aged 6 years. Further, from Figure 4(C), a 9-year-old girl (except for the very fittest of that age group) could have a fitness level that appears in the plots at any age group of girls from 6–11 years old.

It is obvious that students of grades 1 to 3 (arguably less fit) will not be administered the more difficult 9-min run test, while students of grades 4 to 6 (apparently more fit) will not be administered the less-demanding 6-min run test. But because the scale category calibrations of the R6 and R9 are quite close to each other (as shown in Figure 1), it might give the impression that by holding SU and DH performances constant, those who scored a 4 on the R6 scale will have the same (or more or less the same) abilities as those who scored a 4 on the R9 scale. Is it then reasonable to conclude that a second grader who took the 6-min run test will perform similarly (i.e., have a similar level of physical fitness) to a fifth grader who took the 9-min run test if that more difficult test was administered? Conversely, would a fifth grader perform similarly to a second grader in the 6-min run test if given the easier test? The prima facie evidence to support the category equivalence conclusion for R6 and R9 is displayed in Table 5: although the *number* of meters covered varies by category according to whether R6 or R9 was administered, the *speed* (meters per min) varies by category but is independent of actual test. Given that the two subsamples, grades 1 to 3 and 4 to 6, were not independently calibrated and then equated, it remains open to future investigation concerning the extent to which their person measures will be useful for tracking changes in fitness levels over time.

## CONCLUSION AND RECOMMENDATIONS

Four physical fitness indicators: 6-min run, 9-min run, 1-min sit-ups, and dominant handgrip, were calibrated successfully to form the RMPFS. The other six routinely used fitness indicators—BMI,

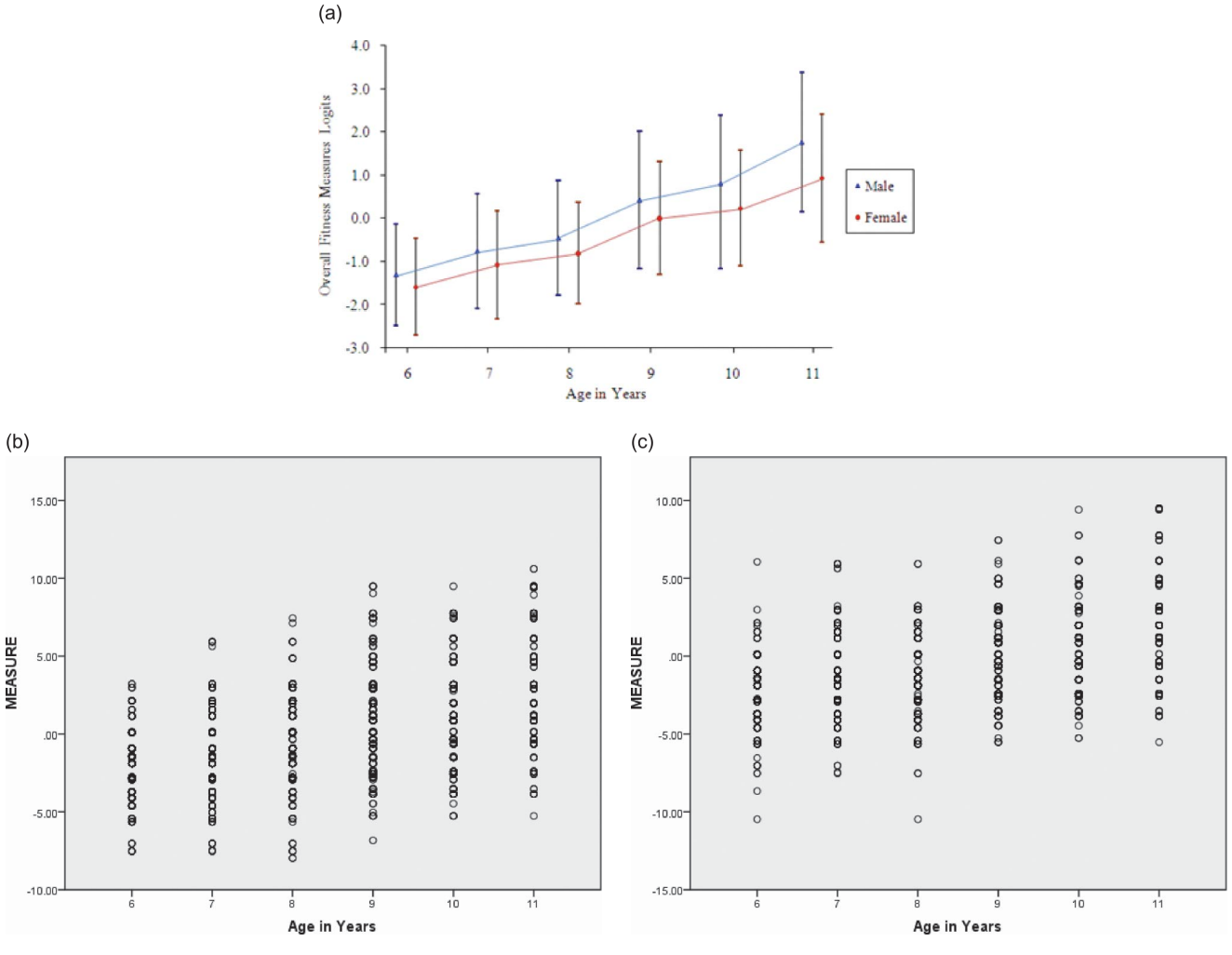


FIGURE 4 (A) Fitness development by age and sex ($M \pm 1\ SD$), (B) distribution of fitness levels for boys, and (C) distribution of fitness levels for girls (color figure available online).

TABLE 5
Category Equivalence for R6 and R9

| *Raw Scores (M) Min–Max* | *Meters/Min Min–Max* | *R6 R9 Category* | *Meters/Min Min–Max* | *Raw Scores (M) Min–Max* |
|---|---|---|---|---|
| 360–471 | 60–79 | 1 | 62–81 | 560–726 |
| 472–676 | 79–113 | 2 | 81–114 | 727–1,027 |
| 677–809 | 113–135 | 3 | 114–136 | 1,028–1,221 |
| 810–969 | 135–162 | 4 | 136–161 | 1,222–1,452 |
| 970–1,160 | 162–194 | 5 | 161–192 | 1,453–1,727 |
| 1,161–1,389 | 194–232 | 6 | 192–228 | 1,728–2,054 |
| 1,390–1,520 | 232–253 | 7 | 228–249 | 2,055–2,240 |

*Note:* The two columns on the left present the meters covered and speed for each category of R6; the two columns on the right present the meters covered and speed for each category of R9.



sit-and-reach, right handgrip, left handgrip, standard push-ups, and modified push-ups—were excluded from the RMPFS because of violation of the Rasch model's requirements or other practical considerations. The RMPFS now provides a single overall person measure of health-related physical fitness for these Hong Kong primary school-aged students. The RMPFS and its scale indicators showed fit to the Rasch model sufficient for the intended purposes of measuring overall fitness of children and monitoring changes in fitness levels over time.

This Rasch-calibrated physical fitness scale and indicators have a number of benefits over reliance on independent scores for separate fitness indicators or components. First of all, the RMPFS transforms the ordinal raw scores into equal-interval measures on a logit scale, which have consistent and stable meaning on the underlying trait continuum (i.e., physical fitness) so that it facilitates interpretation and comparison of students' physical fitness levels.

The second, the measures provided by the RMPFS, have the features (sample-distribution-free and item-distribution-free estimation) espoused by objective measurement. Estimates of both students and physical fitness indicators can be located on a common physical fitness scale and interpreted in a stable interpretive measurement framework, making it easy to compare students' performances on different physical fitness indicators as long as that indicator is calibrated on the scale.

The third, the RMPFS, calibrates students' overall fitness levels, and the indicator difficulties are calibrated on a single uni-dimensional scale. This is a simplified way of generating an overall physical fitness measure that summarizes a student's physical fitness in different components. Even if the student had not performed on all four RMPFS physical fitness indicators, that student can be given an overall RMPFS fitness measure. By constructing an objective measurement scale for overall fitness, the fitness of the students can be located on an interval-level measurement scale, which of course maintains the lower level ordinal relationships between the positions of all students. This means that any fitness level (e.g., +1.0 logit) is objective, i.e., independent of any personal characteristic (e.g., sex or age) of the child. Every child with an RMPFS measure of +1.0 logit has the same fitness level (within error), which is 2 logits higher than any child with a –1.0 logit RMPFS score and 2 logits lower than those with a +3.0 fitness level. On the basis of the overall RMPFS measures, it would be possible to construct a norm-based scoring system

for specific age groups if the interpretation of students' performance under a norm-reference framework is needed. Although this study explored the theoretical possibility of estimating an overall person measure of fitness based on very limited testing scores, it is not recommended to determine a person's fitness from only one assessment, because in practice, the information from just one indicator is too limited; i.e., the measurement error is too large for practical purposes. This overall measure combines core fitness components—cardio-respiratory fitness, muscular endurance, and muscular strength—but is not the simple average of the performances on different components. This simplified approach and reporting system provides a more efficient method and reduces teachers' workloads so that they could put more time and resources into the teaching and learning that promotes children's physical fitness and health.

This research has its own limitations, and future research that emphasizes the following will extend the contributions of this research to physical fitness assessment. First, since this study took the data-fit-the-model approach, the indicators for two components of fitness (BMI for body composition and sit-and-reach for flexibility) were excluded from the scale due to failure to fulfill the Rasch requirements or practical considerations. Clearly, body composition and flexibility are important and related both to fitness and to health. The conclusion is not that they are not important, but this research reveals that they do not behave, in the measurement sense, in the same way as the other fitness indicators. At this stage, the assertion is that they should be assessed, recorded, and utilized as indicators in their own right—but not included in the construct of general fitness measure for these children. Future research could explore this point further through two angles. One is to adhere to the data-fit-the-model approach and to make efforts to identify more appropriate indicators for the components of body composition and flexibility, which can be calibrated successfully into the Rasch measurement scale; those attempts could be made in smaller, more closely controlled fitness testing contexts. The other is to explore multi-dimensional and continuous Rasch models so as to identify a model with a better fit to the data. However, the model-fits-data approach is likely to lose the strong measurement benefits that could be derived by adherence to Rasch model principles.

Second, the RMPFS relies on the data exclusively from the partner school. That brings a limitation to immediate generalization of the RMPFS developed in this research to application in other samples. Thus, this study's core value remains in trying a new approach to physical fitness measurement and building a good model practice, rather than providing a ready-for-use instrument for general physical fitness assessment. Future research could utilize the techniques used in this research and extend them to a larger sample that might be representative for the whole Hong Kong primary or other school-aged student population so that an RMPFS could be developed for use with all Hong Kong primary or other school-aged students. On the other hand, future research could use the same technique to develop school-based databases for other similar samples to derive the same benefits. In addition to replicating the practical benefits, there are also theoretical benefits that could be derived from applying the same technique to other samples. The invariance of indicator measures (sample distribution free) required by Rasch model's means that indicator measures should be independent of any particular sample used for indicator calibration. However, this research itself did not provide direct evidence of this feature since it did not apply the RMPFS to other samples. Future investigations using already existing data from other resources could provide further evidence of the validity to the RMPFS.

Finally, this research does not deny that psychometric approaches to data analysis, other than the Rasch model, might be appropriate for producing more comprehensive descriptions of the

variability in this large longitudinal dataset of children's physical fitness indicators. At the conclusion of this research, it remains an open question as to whether other quantitative approaches might produce results that have better fit of the model to the data. The completion of such an investigation could provide an interesting complement to the results of the data-fit-the-model Rasch measurement approach explicitly adopted at the outset of this research.

## ACKNOWLEDGMENTS

This study is funded by a James Cook University Doctoral Publication Award. An earlier version of the article was presented to the Pacific Rim Objective Measurement Symposium (PROMS) at The Hong Kong Institute of Education, Hong Kong, July 28–30, 2009. The authors thank the reviewers for their constructive critique.



## REFERENCES

Andersen, E. B. (1995). What Georg Rasch would have thought about this book. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 383–390). New York: Springer.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, 1–16.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Bowles, R. P., & Ram, N. (2006). Using Rasch measurement to investigate volleyball skills and inform coaching. *Journal of Applied Measurement*, *7*, 39–54.

Büsch, D., & Strauss, B. (2005). Qualitative differences in performing coordination tracks. *Measurement in Physical Education and Exercise Science*, *9*, 161–180.

Büsch, D., Strauss, B., Seidel, I., Pabst, J., Tietjens, M., Müller, L., et al. (2009). Die Konstruktvalidität des Allgemeinen Sportmotorischen Tests für Kinder [Construct validity of the general athletic motor skill test for children]. *Sportwissenschaft* [Sports Science], *39*, 95–103.

Fleishman, F. A. (1964). *The structure and measurement of physical fitness*. Englewood Cliffs, NJ: Prentice-Hall.

Hands, B., & Larkin, D. (2001). Using the Rasch measurement model to investigate the construct of motor ability in young children. *Journal of Applied Measurement*, *2*, 101–120.

Heesch, K. C., Masse, L. C., & Dunn, A. L. (2006). Using Rasch modeling to re-evaluate three scales related to physical activity: Enjoyment, perceived benefits and perceived barriers. *Health Education Research, 21*, electronic version. Retrieved October 3, 2006, from http://her.oxfordjournals.org/cgi/reprint/cyl054v1.

Heyward, V. H. (2002). *Advanced fitness assessment and exercise prescription* (4th ed.). Champaign, IL: Human Kinetics.

Hong Kong Education and Manpower Bureau. (2005). *Hong Kong school physical fitness award schemes: Teachers' handbook*. Retrieved August 10, 2006, from http://cd1.emb.hkedcity.net/cd/pe/tc/rr/pfas/handbook.

Hsueh, I. P., Wang, W. C., Sheu, C. F., & Hsieh, C. L. (2004). Rasch analyses of combining two indices to assess comprehensive ADL function in stroke patients. *Stroke*, *35*, 721–726.

Kang, J., & Kang. M. (2006). Rasch calibration of perceived weights of different sports games. *Measurement in Physical Education and Exercise Science*, *10*, 51–66.

Linacre, J. M. (2000). New approaches to determining reliability and validity. *Research Quarterly for Exercise and Sport*, *71*, 129–136.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

Linacre, J. M. (2006a). *A user's guide to WINSTEPS/MINISTEPS: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2006b). Data variance explained by Rasch measures. *Rasch Measurement Transactions*, *20*, 1045.

Liu, Y. (2008). Youth fitness testing: If the "horse" is not dead, what should we do? *Measurement in Physical Education and Exercise Science*, *12*, 123–125.

Mahar, M. T., & Rowe, D. A. (2008). Practical guidelines for valid and reliable youth fitness testing. *Measurement in Physical Education and Exercise Science*, *12*, 126–145.

Marsh, H. W. (1993). The multidimensional structure of physical fitness: Invariance over gender and age. *Research Quarterly for Exercise and Sport*, *64*, 256–273.

Marsh, H. W., & Redmayne, R. S. (1994). A multidimensional physical self-concept and its relations to multiple components of physical fitness. *Journal of Sport & Exercise Psychology*, *16*, 43–55.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Silverman, S., Keating, X. D., & Phillips, S. R. (2008). A lasting impression: A pedagogical perspective on youth fitness testing. *Measurement in Physical Education and Exercise Science*, *12*, 146–166.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*, electronic version. Retrieved July 7, 2009, from http://www.biomedcentral.com/1471-2288/8/33.

Smith, Jr., E. V. (2001). Evidence for the reliability of measures and the validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, *2*, 281–311.

Stratton, G., & Williams, C. A. (2007). Children and fitness testing. In E. M. Winter, A. M. Jones, R. C. R. Davison, P. D. Bromley, & T. H. Mercer (Eds.), *Sport and exercise physiology testing guidelines* (pp. 211–223). New York: Routledge.

Strauss, B., Büsch, D. & Tenenbaum, G. (2007). New developments in measurement and testing. In G. Tenenbaum & R. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 737–756). Boston, MA: Wiley.

Tenenbaum, G., Strauss, B., & Büsch, D. (2007). Applications of generalized Rasch models in sport, exercise and the motor domains. In M. V. Davier & C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 347–357). New York: Springer.

Verhelst, N. D., & Glas, C. A. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–237). New York: Springer.

Welk, G. J. (2008). The role of physical activity assessments for school-based physical activity promotion. *Measurement in Physical Educationand Exercise Science*, *12*, 184–206.

Williams, C. S., Harageones, E. G., Johnson, D. J., & Smith, C. D. (2000). *Personal fitness: Looking good/feeling good* (4th ed.). Dubuque, IA: Kendall/Hunt Publishing Company.

Wright, B. D. (1997). A history of social science measurement. *Educational measurement: Issues and Practice*, *16*(4), 33–45.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Mok, M. M. C. (2000). Rasch models overview. *Journal of Applied Measurement*, *1*, 83–106.

Zhu, W. (2001). An empirical investigation of Rasch equating of motor function tasks. *Adapted Physical Activity Quarterly*, *18*, 72–89.

Zhu, W., & Cole, E. L. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research Quarterly for Exercise and Sport*, *67*, 24–34.

Zhu, W., & Kurz, K. A. (1994). Rasch partial credit analyses of gross motor competence. *Perceptual & Motor Skills*, *79*, 947–961.

Zhu, W., & Safrit, M. J. (1993). The calibration of a sit-up task using the Rasch Poisson counts model. *Canadian Journal of Applied Physiology*, *18*, 207–219.

Zhu, W., Timm, G., & Ainsworth, B. (2001). Rasch calibration and optimal categorization of an instrument measuring women's exercise perseverance and barriers. *Research Quarterly for Exercise and Sport*, *72*, 104–116.