Pertanika J. Soc. Sci. & Hum. 3(1): 87-96 (1993)

ISSN: 0128-7702

Physical Fitness Test Battery for Malaysian Schoolchildren Aged 13 - 15 Years

JABAR HAJI JOHARI and TAN KOK SIANG
National Sports Council of Malaysia
Stadium Negara, Jalan Hang Jebat
P.O. Box 10440, 50714 Kuala Lumpur, Malaysia

Keywords: reliability, test battery, physical fitness, test and re-test, interclass coefficlent, test for independence

ABSTRAK

Tujuan kajian ini adalah untuk menentukan anggaran kebolehpercayaan tujuh sub-ujian yang terdapat di dalam Ujian Kecergasan Fizilak Majlis Sukan Negara untuk menguji subjek bagi Projek Sarawak dan Projek Tunas Cemerlang. Bateri ujian tersebut telah diuji dan ulang-uji dengan menggunakan sampel yang terdiri daripada 25 pelajar (min. umur 14.2 tahun, s.d. 0.51). Ulang uji dilaksanakan selepas tiga hari, dan anggaran intrakelas R dikira menggunakan kaedah ANOVA. Ujian F (tahap keertian 0.05) digunakan untuk menguji kebebasan ulangan mendapati kesan peergantungan bagi ujian-ujian 1500 meter dan duduk dan jangkau. Pada keseluruhannya nilai koefisien kebolehpercayaan intrakelas yang tinggi telah diperolehi, yakin 0.97 (lari 1500 meter); 0.96 (duduk - jangkau); 0.96 (pecut 50 m); 0.95 (larian ketangkasan); 0.94 (gayut siku bengkok); 0.94 (lompat sarjan) dan 0.86 (bangkit tubi).

ABSTRACT

The purpose of this study was to estimate the reliability of 7 subtests in the National Sports Council physical fitness test battery as used in their Sports Science Division's 'Sarawak Project' and Talent Development Division's 'Budding Talent' (Tunas Cemerlang) programmes. The test battery was administered to a sample class of 25 students (mean age 14.2 years, s.d. 0.51) on 2 occasions, three days apart, to estimate the intraclass R using ANOVA procedures. The F-test for independence of trials (significance level of 0.05) detected a dependency for the 1500m run and the sit and reach test. High values of intraclass reliability coefficient were obtained: 0.97 (1500 m run); 0.96 (sit and reach, and 50 m run); 0.95 (agility run); 0.94 (flexed-arm hang, and sargent jump) and 0.86 (sit-up).

INTRODUCTION

The Sports Science Division of the National Sports Council of Malaysia (NSC) has developed the National Sports Council Physical Fitness Test Battery. It was the aim of this study to establish the validity and reliability of this national fitness test battery, which could contribute significantly towards the development of a truly Malaysian physical fitness test battery. The test was used by the Sports Science Division to assess the physical fitness of Sarawak schoolchildren aged 13-15 in the 'Sarawak Project'. The test was again used to assess the physical fitness of schoolchildren aged 13-15 throughout Malaysia who were selected for the 'Budding Talent' programme by the Talent Development Division of the National Sports Council. With the availability of a valid and reliable instrument for testing physical fitness parameters, the process of talent identification in Malaysia could be enhanced and rendered reliable and scientific. It is also perceived that the findings of this study could contribute significantly towards the development of a fitness test battery for both the schools and the general population of Malaysia.

The NSC physical fitness test battery comprises seven subtests. The testing procedures were carried out by trained schoolteachers and NSC officials so that the validity and reliability of the test battery, including the subtests, could be thoroughly investigated. A review of Western literature yielded only reported values of the validity and reliability for each of the tests. It was deemed pertinent to re-examine the validity and reliability for each of the tests as applied in Malaysia. Since face validity for each of the tests was accepted (Klesius 1968; Nelson and Johnson 1986; Dinucci *et al.* 1990), this paper was designed to look into a reliability estimation of the tests as applied in Malaysia.

OBJECTIVE

The objective of this paper was to field-test and obtain:

a. an estimate of the intraclass reliability R, of the 1500 metre run,
b. an estimate of the intraclass reliability R, of the 50 metre run,
c. an estimate of the intraclass reliability R, of the agility run,
d. an estimate of the intraclass reliability R, of the flexed-arm hang,
e. an estimate of the intraclass reliability R, of the sargent jump,
f. an estimate of the intraclass reliability R, of the sit and reach,
g. an estimate of the intraclass reliability R, of the sit-up.

The above comprise the subtests containe in the NSC physical fitness test battery.

Significance

Reliability estimates reported in the various sources of literature were obtained from samples which vary a great deal in age, socio-cultural and physical education background from those of Malaysian schoolchildren. These samples were drawn from American populations of 6-12 and 15-17 year-old children and adolescents. Furthermore, any modification of the testing procedures, as applied in Malaysia, make such reported values inapplicable. No studies on the reliability estimation of similar physical fitness tests in Malaysia were found.

METHOD

The pilot study was undertaken by the NSC Sports Science Division in Miri, Sarawak in July-August 1991. It involved 25 subjects (9 males and 16 females). The NSC research officer was the research co-ordinator and was assisted by Md. Saadon Abd. Shukor and four local research assistants or testers employed on a contract basis.

The sampling procedure was carried out in May 1991. A sample class was obtained from Form 2A of Tun Dato' Tuanku Haji Bujang College, Miri, Sarawak. The sample class was randomly selected from a list of all secondary schools in Sarawak by a computer using the Lotus 123 version 2.3 software.

The training for the testers was conducted on 22-23 July 1991. It was comprised of the test procedures, protocols, measurement techniques and safety procedures to be used in the tests. The training culminated with a mock trial in the administration of the test battery on a selected class of 30 schoolchildren from Miri, Sarawak.

On 29th July 1991, the NSC physical fitness test battery was administered to the subjects in Trial 1. Three days later, on 1st August 1991, the same test battery was again administered to the same subjects in Trial 2. Guidelines for the safety, protection and privacy of the human subjects were strictly observed.

Description of Tests

The NSC physical fitness test battery consists of 2 sections. Section 1 contains the subjects' biodata (name, date of birth, sex, weight, standing height, sitting height and reaching height}. Section 2 comprises the subtests, namely the 1500 metre run, the 50 metre run, the agility run, the flexed-arm hang, the sargent jump, the sit and reach, and the sit-up.

The 1500 metre run test was included to gauge cardiorespiratory fitness. Research has indicated that distance runs over one kilometre adequately assess cardiorespiratory capacity (Dinucci *et al.* 1990). The subjects ran as fast as they could on a marked 400-metre grass track. The score was the time taken (to the nearest second) to complete the run.

The 50 metre run test measures speed. Face validity was accepted for this test. Subjects ran as fast as they could on a marked 50-metre grass track. A standing start was used. Timing was done manually (to 2 decimal places) on a digital stop-watch.

The agility run test was adapted from the AAHPER Shuttle Run (Safrit 1976). It measures the agility of the subjects in running and changing position. Face validity was accepted for this test (Nelson and Johnson 1986). Subjects stood behind the starting line, and on the signal "go" they ran to a baton placed 10 metres away, picked it up, returned to the starting line, placed the baton behind the line and repeated the process to pick up the second baton. Two attempts were allowed, with a 5-minute rest between them. The score was the faster time taken to complete the course. Timing was done manually (to 2 decimal places) on a digital stopwatch.

The flexed-arm hang test is used to measure upper body muscular strength and endurance. Face validity was accepted for this test (Nelson and Johnson 1986). A horizontal bar 3.5 centimetres in diameter was raised to a height such that the tallest subject could not touch the ground from the flexed-arm hang position. With the assistance of a spotter, the subject raised his/her body off the floor so that the chin was above the bars and the elbows flexed. An overhand grip was used. The score was the length of time (seconds, to 1 decimal place) this position was held (i.e., until the subject's eyes dropped below the horizontal bar).

The sargent jump test was adapted from the sargent chalk jump, 1921 (Nelson and Johnson 1986). It is used to measure the power of the legs in jumping vertically upwards. A validity of 0.78 has been reported for this test (Nelson and Johnson 1986). Subjects stood with one side against a wall, heels together, and held a one inch piece of chalk in the hand nearest to the wall. With the heels on the floor he/she then reached as high as possible to make a mark on the wall. The subject was then required to jump as high as possible and make another mark on the wall. Three attempts were given, with a one-minute rest between each one. The score was the distance in centimetres (to the nearest half centimetre) between the reach (lower) mark and the highest jump mark attained by the subject.

The sit and reach test is used to evaluate hip and back flexion as well as extension of the hamstring muscles of the legs. It was administered according to instructions as contained in the modified sit and reach test (Johnson 1977). Face validity was accepted for this test. Subjects were given 3 attempts. No rest was given between attempts. The score was the best of the 3, measured in centimetres (to 1 decimal place).

Testing Protocol

Subjects reported to the research co-ordinator at 0730 hours. Division into four groups (A, B, C and D) was followed by ten minutes of standard warming-up conducted by the research co-ordinator.

The first test administered was the sit-up test. Subjects were divided into pairs to utilize the partner system. One subject performed sit-ups and the other played the role of supporter and counted the number of sit-ups completed. After the sit-up test, subjects were given instructions to move to one of the following testing stations:

i. Group A to Station A, agility run,

ii. Group B to Station B, flexed-arm hang,

iii. Group C to Station C, sargent jump,

iv. Group D to Station D, sit and reach.

The rotating group system was employed. At the end of each test, Group A, moved to Station B, Group B to Station C, Group C to Station D and Group D to Station A subsequently.

On arrival at a station, each group was briefed on the testing procedures by the tester, who also acted as the scorer. Between tests, the subjects were given a ten-minute rest.

Upon completion of the four stations, all groups moved to the grass track for the 50 metre run. For each 50 metre run, there were 4 subjects (one from each group). The research co-ordinator was the starter who followed standard starting procedures. The testers were the timekeepers and scorers.

The last test item was the 1500 metre run, which was conducted after a one-hour rest interval. The partner system was utilized with one subject being the testee and the partner as the supporter, recording the number of laps and the final times as called out by the timekeeper. The starter was the research co-ordinator who followed standard starting procedures. The timekeeper was a tester. The research assistant and the remaining testers were responsible for supervising the recording of laps and finishing times.

Three days later, Trial 2 was conducted, with the same test battery repeated on the same subjects, at the same place, and utilizing the same testing protocol in the exact order as Trial 1. A new score sheet was used to ensure that the testers did not have access to the previous test scores.

Data Analysis

All the data collected were centrally processed at the NSC Sports Science Division by the research co-ordinator. The data were analysed using Lotus 123 version 2.3 and dBase IV computer programs.

Reliability estimates for interclass R were computed using Pearson Product-Moment. This method is most appropriate in looking at correlation between Trial 1 and Trial 2.

TABLE 1
Age, height and weight of subjects

No.	Name	Sex	Date of Birth	Age	Height (cm)	Weight (kg)
1	Adeline	P	26-04-78	13.3	150.0	42.0
2	Alison	P	28-10-77	13.8	143.0	38.0
3	Aloysius Siran	L	09-07-77	14.1	164.0	55.0
4	Annie Chua	P	21-11-77	13.7	160.3	63.0
5	Collin Tang	L	09-10-77	13.8	168.2	51.0
6	Devrin Jack	L	23-10-76	14.8	166.0	45.0
7	Eii Meng Yin	P	24-09-77	13.9	168.7	50.5
8	Florence	P	17-07-77	14.0	154.6	48.5
9	Freda Wan	P	30-10-77	13.0	155.2	60.0
10	Irene Tulsa	P	12-05-77	14.2	153.5	45.0
11	Jacey	P	03-01-77	14.6	151.5	43.0
12	Jerey Awang	L	27-01-77	14.5	158.4	43.5
13	Kelvin	L	01-07-77	14.1	160.2	46.0
14	Lahat Wan	P	06-12-75	15.7	144.0	41.5
15	Marina Tan	P	22-02-77	14.4	154.6	44.0
16	Nellie	P	26-01-77	14.5	158.3	46.0
17	Philomena	P	11-09-77	14.9	150.0	48.5
18	Polly	P	04-01-77	14.6	150.0	48.5
19	Raymond	L	19-10-77	13.8	164.5	49.0
20	Ruth	P	03-08-77	13.8	164.5	49.0
21	Salvia	P	19-11-77	13.5	144.2	33.0
22	Simbah	P	15-12-77	13.6	164.5	65.0
23	Susan	P	04-07-77	14.1	154.5	42.0
24	Terrance	L	23-03-77	14.4	150.0	42.0
25	Ting Chek Chang	L	23-04-77	14.3	158.0	42.5
	Mean			14.18	155.96	47.06
	S.D.			0.51	7.31	7.22
	Highest			15.7	168.7	65.0
	Lowest			13.3	143.0	33.0

Reliability estimates for intraclass R were computed using ANOVA procedures. If the trials lacked independence, trials causing significant effect were deleted, and a new ANOVA summary table was developed to test the trials for randomness again. This procedure was advocated by Nelson and Johnson (1986), and Safrit (1981).

RESULTS

Data collected in the study are presented in the following tables. In Table 1 it can be seen that the mean age of the subjects was 14.18 + 0.51 years, mean height was 155.96 + 7.31 cm and mean weight was 47.06 + 7.22 kg.

In the 1500 metre run (Table 2), the mean score for Trial 1 was 489.62 + 65.21 seconds. In Trial 2 it improved to 469.90 + 55.84 seconds. This improvement in performance was significant (df15, F = 4.27) at the 0.05 level.

In the 50 metre run (Table 3) the mean score for Trial 1 was 8.61 + 1.03 seconds. For Trial 2 it was 8.53 + 1.03 seconds. There was no significant difference (df24, F = 0.97) at the 0.05 level between the two trials.

In the agility run (Table 4), the mean score for Trial 1 was 11.22 + 0.85 seconds. For Trial 2 it was 11.15 + 0.83 seconds. There was no significant difference (df24, F = 0.97) at the 0.05 level between the two trials.

In the flexed-arm hang (Table 5), the mean score for Trial 1 was 12.9 + 16.9 seconds. For Trial 2 it was 11.3 + 13.5 seconds. There was no significant difference (df24, F= 1.30) at the 0.05 level between the two trials.

TABLE 2
1500 metre run test results

Name	1500m run Trial 1	 Trial 2
	(sec.)	
Adeline	474	478
Alison		
Aloysius Siran	437	418
Annie Chua	553	528
Collin Tang	362	355
Devrin Jack	395	406
Eii Meng Yin	538	491
Florence	505	507
Freda Wam	473	490
Irene Tulsa	562	489
Jacey	580	562
Jeerey Awang	440	431
Kelvin	463	468
Lahat Wan	534	509
Marina Tan	562	511
Nellie		
Philomena	549	548
Polly		
Raymond	492	427
Ruth	522	500
Salvia	470	433
Simbah		
Susan	578	528
Terrance	408	420
Ting Chek Chang	385	369
Mean	489.62	469.90
s.d	65.21	55.84
max. score	580.0	562.0
min. score	362.0	355.0

TABLE 3
50 metre run rest results

Name	50m run Trial 1	 Trial 2
	(sec.)	
Adeline	7.98	8.17
Alison	9.91	9.15
Aloysius Siran	7.29	7.28
Annie Chua	9.43	8.87
Collin Tang	7.10	6.78
Devrin Jack	7.47	7.60
Eii Meng Yin	8.72	6.64
Florence	8.94	8.96
Freda Wam	7.82	8.16
Irene Tulsa	8.97	9.00
Jacey	8.50	8.82
Jeerey Awang	7.76	7.84
Kelvin	7.87	7.71
Lahat Wan	10.04	9.83
Marina Tan	8.98	9.19
Nellie	7.20	7.48
Philomena	10.69	10.43
Polly	10.12	10.53
Raymond	8.58	8.06
Ruth	9.28	8.08
Salvia	8.40	8.20
Simbah	9.96	10.65
Susan	8.67	8.60
Terrance	7.41	7.62
Ting Chek Chang	7.92	7.34
Mean	8.61	8.53
s.d	1.03	1.03
max. score	10.69	10.65
min. score	7.10	6.78

In the sargent jump (Table 6), the mean score for Trial 1 was 42.71 + 7.59. For Trial 2 it was 42.53 + 9.57 cm. There was no significant difference (df23, F = 0.04) at the 0.05 level between the two trials.

In the sit and reach (Table 7), the mean score for Trial 1 was 55.04 + 8.27 cm. For Trial 2, it improved to 57.06 + 8.70 cm. This improvement in performance was significant (df21, F = 3.88) at the 0.05 level.

In the sit-up (Table 8), the mean score for Trial 1 was 27.16 + 8.08 times. For Trial 2 it was 26.32 + 7.51 times. There was no significant difference (df24, F = 0.57) at the 0.05 level between the two trials.

Table 9 gives a summary of the analysis of variance for reliability estimate.

Interclass R is 0.99 (sit and reach); 0.92 (1500 metre run and 50 metre run); 0.91 (agility run, flexed-arm hang, and sargent jump); and 0.76 (sit-up).

Intraclass R is 0.97 (1500 metre run); 0.96 (sit and reach, and 50 metre run); 0.95 (agility run); 0.94 (flexed-arm hang, and sargent jump); and 0.86 (sit-up).

DISCUSSION

The reliability of a test refers to the dependability of scores and their relative freedom from error. It is usually thought of as 'repeatability of the test', whereby an individual's score should not differ markedly on repeated administrations of the same test. This paper discusses the estimation of reliability within a norm-referenced framework with the underlying rationale of detecting individual differences. Hence, a better definition would then be:

TABLE 4
Agility run test results

Name	Agility Run Trial 1	(sec.) Trial 2
Adeline	10.93	10.89
Alison	12.30	12.20
Aloysius Siran	10.37	10.02
Annie Chua	11.17	11.62
Collin Tang	9.66	9.51
Devrin Jack	10.57	10.07
Eii Meng Yin	11.23	11.67
Florence	11.70	11.59
Freda Wam	10.59	10.58
Irene Tulsa	11.13	11.29
Jacey	11.26	11.14
Jeerey Awang	10.55	10.38
Kelvin	11.28	10.97
Lahat Wan	11.74	11.40
Marina Tan	11.59	11.56
Nellie	9.79	9.87
Philomena	12.52	12.30
Polly	12.38	12.23
Raymond	11.20	11.06
Ruth	11.72	11.32
Salvia	11.05	10.96
Simbah	13.40	12.50
Susan	11.53	12.65
Terrance	10.21	10.08
Ting Chek Chang	10.62	10.29
Mean	11.22	11.15
s.d	0.85	0.83
max. score	13.40	12.65
min. score	9.66	9.51

TABLE 5
Flex arm hang test results

Name	Flexed-arm Hang Trial 1	(sec.) Trial 2
Adeline	4.4	2.7
Alison	0.0	0.0
Aloysius Siran	42.3	40.3
Annie Chua	0.0	0.0
Collin Tang	32.3	21.1
Devrin Jack	36.2	46.1
Eii Meng Yin	1.2	1.4
Florence	0.0	0.0
Freda Wam	4.3	2.6
Irene Tulsa	0.0	2.7
Jacey	8.9	10.9
Jeerey Awang	16.9	19.3
Kelvin	28.1	18.2
Lahat Wan	5.6	9.5
Marina Tan	1.0	0.0
Nellie	36.6	13.1
Philomena	0.0	0.0
Polly	0.0	0.0
Raymond	5.2	9.5
Ruth	1.6	4.2
Salvia	9.4	14.6
Simbah	0.0	0.0
Susan	2.4	2.6
Terrance	53.1	38.1
Ting Chek Chang	32.2	24.9
Mean	12.9	11.3
s.d	16.2	13.5
max. score	53.1	46.1
min. score	0.0	0.0

'...... The reliability of a test refers to the proportion of variance in test scores due to true differences within a population of individuals on the attribute being measured by the test .....' (Safrit 1981).

The values of the reliability coefficient obtained in this study were high. Table 11 compares the obtained values with those of some reported values. The obtained values of R compare favourably with the reported values.

The high values of the reliability coefficient can probably be attributed to the following reasons:

a. Systematic variation, that is variation in behaviour of a biological nature, is reduced. Testees were untrained and the rest period of three days between the test and retest was adequate for them to recover. They were probably equally motivated on both occasions.

b. Error variance, that is variability due to measurement error, is greatly reduced because:

i. Equipment remained the same throughout, and was calibrated in units appropriate to the scale being measured.

ii. Scorers/testers were retained for the same test during the retest.

iii. Test protocol was easily understood by testers and testees. Furthermore, the skill demands of the tests were low. If skill or accuracy component of the tests was high, then a high coefficient would not necessarily be obtained.

The reliability in this study was estimated using interclass (Pearson Product-Moment) and Intraclass (ANOVA procedures). Even though the values obtained from both methods did not differ greatly, it was found that the results of the Pearson Product-Moment method were not

TABLE 6
Sargent jump test results

Name	Sargent Jump Trial 1	 Trial 2
	(sec.)	
Adeline	43.0	41.0
Alison		
Aloysius Siran	51.5	55.2
Annie Chua	35.5	44.0
Collin Tang	58.0	64.5
Devrin Jack	50.0	52.0
Eii Meng Yin	48.0	45.0
Florence	39.0	38.0
Freda Wam	43.0	41.0
Irene Tulsa	39.0	40.0
Jacey	38.0	40.0
Jeerey Awang	48.0	50.0
Kelvin	42.0	49.5
Lahat Wan	34.0	32.0
Marina Tan	38.0	33.5
Nellie	47.0	43.0
Philomena	31.0	28.0
Polly	31.0	27.0
Raymond	41.0	43.5
Ruth	38.0	26.6
Salvia	45.0	45.0
Simbah	37.0	36.0
Susan	36.0	35.5
Terrance	58.0	55.5
Ting Chek Chang	54.0	55.0
Mean	42.71	42.53
s.d	7.59	9.57
max. score	58.0	64.5
min. score	31.0	26.6

TABLE 7
Sit and reach test results

Name	Sit and Reach Trial 1	 Trial 2
	(sec.)	
Adeline	63.8	62.2
Alison	61.8	63.2
Aloysius Siran	62.8	66.6
Annie Chua	54.8	58.0
Collin Tang	51.6	52.4
Devrin Jack	46.0	42.5
Eii Meng Yin	60.2	57.7
Florence	61.9	61.5
Freda Wam	71.2	67.0
Irene Tulsa	57.5	56.1
Jacey	58.0	58.9
Jeerey Awang	38.7	41.9
Kelvin	35.4	33.8
Lahat Wan	50.3	56.5
Marina Tan	55.0	56.7
Nellie	51.7	59.1
Philomena	49.3	51.7
Polly	49.4	48.0
Raymond	48.6	53.7
Ruth	48.3	56.8
Salvia	53.9	56.3
Simbah	68.9	76.1
Susan	60.7	64.7
Terrance	57.9	62.4
Ting Chek Chang	58.3	62.6
Mean	55.04	57.06
s.d	8.27	8.70
max. score	71.2	76.1
min. score	35.4	33.8

conclusive. This was because, in computation where correlation between two trials of the same test for the same individual was employed, the correlation coefficient merely reflects the relative position of paired scores in each of two Z-score distribution. Such conversion of raw scores to Z scores would mark any systematic increases or decreases from trial to trial (Safrit 1981). This was best demonstrated in the case of the 1500 metre run and the sit and reach test, where significant differences were only detected between Trial 1 and Trial 2 by using ANOVA procedures. Mathematically, the Pearson Product-Moment is a bivariate statistic whereas reliability estimates are univariate.

In the analysis of variance for intraclass correlation, the F test for trials was computed to determine whether there were significant differences between the trials. If there were no differences, the variance for the trials was simply included as part of the measurement error in the calculation of R. If there were differences, the procedures advoceehed by Safrit 1981 were used. The researcher discarded any dissimilar scores and then performed a new analysis using the same procedures with the remaining scores until no significant trial difference was found.

In this pilot study, significant difference was found between Trial 1 and Trial 2 in the 1500 metre run and the sit and reach test. Further examination of the scores showed an improvement in Trial 2 and that the differences do not appear to be random. Hence, the trial variability should be removed and not considered as a measurement error. There was no necessity for a trend analysis because our primary interest lay only in detecting dependency among trials and identifying

TABLE 8
Sit-Up test results

Name	Sit-Up Trial 1 (sec.)	Sit-Up Trial 2 (sec.)
Adeline	28	34
Alison	18	19
Aloysius Siran	33	36
Annie Chua	18	19
Collin Tang	32	31
Devrin Jack	23	32
Eii Meng Yin	16	21
Florence	17	25
Freda Wam	22	15
Irene Tulsa	33	31
Jacey	30	34
Jeerey Awang	37	37
Kelvin	35	35
Lahat Wan	24	24
Marina Tan	32	17
Nellie	41	27
Philomena	22	17
Polly	34	33
Raymond	26	23
Ruth	9	12
Salvia	33	33
Simbah	30	27
Susan	20	18
Terrance	36	36
Ting Chek Chang	39	33
Mean	27.16	26.32
s.d	8.08	7.15
max. score	41.0	37.0
min. score	9.0	12.0

whether trial fluctuations were random. The exact classification of the trend was not important in this case.

The possible sources of systematic variability detected in this study could be:

a. increased familiarity with the equipment in the sit and reach test. This test used equipment which subjects had never seen before.

b. increased motivation in performing the tests. Subjects tried their best to outdo each other and better their scores obtained during Trial 1. This was quite apparent in the 1500 m run where the mean of Trial 2 was 20 seconds faster than that of Trial 1.

A factor that influences the reliability of a test is the range of ability within the group taking the test. The greater the range, the higher the reliability. This effect was minimized because the test reliability reported in this pilot study was based on a particular age group (see Table 1). Hence, a wide range of ability due to significant age differences is greatly reduced.

From this pilot study it is suggested that:

a. comprehensive research be carried out using the NSC physical fitness test battery. Data from this research can be used to further enhance the reliability of the test battery,

b. research be undertaken to estimate the reliability of the health-related physical fitness test known as the UKJK (ujian kecergasan jasmani kebangsaan) currently in use by the Ministry of Youth and Sports, Malaysia. This is to ensure the reliability of the scores obtained by those undergoing this test,

c. a reliability study be carried out on the physical fitness test advocated by the Ministry of Education, Malaysia. This test is compulsory for Malaysian secondary schoolchildren ranging from pre-pubertal to post-pubertal. The tests are carried out by both trained and untrained testers or teachers. Furthermore, facilities, equipment and testing time differ greatly throughout the country. Thus it is essential that the reliability tests and the scores obtained are closely monitored,

d. an estimate of the reliability of a 'whole battery of tests', instead of the 'reliability of each test' as is done in this study, be undertaken. A suggested model would be the canonical correlation model as proposed by Wood and Safrit (1984), and further tested out by Dinucci *et al.* (1990).

This study detected an improvement in performance during Trial 2 for the 1500 metre run and the sit and reach test. Further investigations need to be undertaken to determine the causal relationship for the increase. There is a need to even re-examine the test and testing protocol for sources of error because this improvement might be absent in larger samples.

CONCLUSION

The pilot study indicated high values of intraclass reliability coefficient. The values were 0.97 (1500 m run); 0.96 (sit and reach, and 50 m run); 0.95 (agility run); 0.94 (flexed-arm hang); 0.94 (sargent jump) and 0.86 (sit-ups). Consequently, the National Sports Council physical fitness test battery is a valid and reliable

TABLE 9
Summary of analysis of variance for Relliability Estimate

Test	Soure of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
1500m	Subject	130366.9	*5	8691.1	
	Trials	830.3	1	830.3	4.27
	Interaction	2914.2	15	194.3	
	Total	134111.4	31		
Sit and Reach	Subjects	2871.75	*21	136.75	
	Trials	16.94	1	16.94	3.88
	Interaction	91.70	21	4.37	
	Total	2980.39	43		
50m	Subject	60.00	24	2.12	
	Trials	0.08	1	0.08	0.97
	Interaction	2.01	24	0.08	
	Total	62.09	49		
Agility Run	Subject	33.69	24	1.40	
	Trials	0.07	1	0.07	0.97
	Interaction	1.63	24	0.07	
	Total	35.39	49		
Flexed-arm Hang	Subject	10488.00	24	437.00	
	Trials	31.87	1	31	1.30
	Interaction	589.66	24	24.57	
	Total	11109.53	49		
Sargent Jump	Subject	3370.93	23	146.56	
	Trials	0.40	1	31.87	1.30
	Interaction	211.58	23	9.20	
	Total	3582.91	47		
Sit Up	Subject	2674.17	24	111.42	
	Trials	8.82	1	8.82	0.577
	Interaction	368.68	24	15.36	
	Total	3051.67			

The significance level for the F test is 0.05
* Trials causing significant effect were deleted.

TABLE 10
Interclass and intraclass correlation coefficient

TEST	Interclass R (Pearson-Product)	Interclass R (using ANOVA procedures)
1500 m Run	0.92	0.97 *
Sit and Reach	0.99	0.96 *
50 m Run	0.92	0.96
Agility Run	0.91	0.95
Flexed-arm Hang	0.91	0.94
Sargent Jump	0.91	0.94
Sit Up	0.76	0.86

* There was a significant difference between the two trials at a level of significance of 0.05 for both the 1500 m run and the Sit and Reach test. Further examination yielded an improvement in performance.

instrument to test and collect data in Malaysia, provided that the testing procedures and protocol are adhered to fully.

REFERENCES

BAUMGARTNER, T.A. and A.S. JACKSON. 1970. Measurement for schedules for tests of motor performance. *Research Quarterly* **41**: 10-14.

DINUCCI, J. *et al.* 1990. Reliability of a modification of the health-related physical fitness test for use with physical education majors. *Research Quarterly for Exercise and Sport* **61**: 20-21.

JOHNSON, B.L. 1977. *Practical Flexibility Measurement with the Flexo-measure.* Portland: Brown and Littleman.

KLESIUS, S.E. 1968. Reliability of the AAHPER youth fitness test items and relative efficiency

TABLE 11
Comparison of reliability coefficient

Test	R – (obtained)	R – (reported)	
1500 m Run	0.97	0.98	1 Mile Run, Dinucci, et. aal., (1990)
Sit and Reach		0.96	0.94 Nelson and Johnson, (1986)
50 m Run	0.96	0.95	50-yd run, Baumgaartner and Jackson, (1970)
Agility Run	0.95	not reported	
Flexed-arm Hang	0.94	0.90	Nelson and Johnson, (1986)
Sargent Jump	0.94	0.93	Nelson and Johnson, (1986)
Sit Up		0.86	0.68 Klesius (1986)
		0.98	Nelson and Johnson (1986)
		0.91	Dinucci, et. al., (1990)

of the performance measures. *Research Quarterly* **39**: 801-811.

Nelson, J.K. and B.L. Johnson. 1986. *Practical Measurements for Evaluation in Physical Education.* 4th edn. Minnesota: Burgess.

Safrit, M.J. 1976. *Reliability Theory.* Washington DC: AAHPER.

Safrit, M.J. 1981. *Evaluation in Physical Education.* New Jersey: Prentice Hall.

Walpole, R.E. 1982. *Introduction to Statistics.* New York: Macmillan.

Wood, T.M. and M.J. Safrit. 1984. A model for estimating the reliability of psychomotor test batteries. *Research Quarterly for Exercise and Sports* **55**: 53-63.

(Received 26 July 1993)