Original Article

A Feasible Formula for Criterion-Referenced Test on Men's Health-Related Physical Fitness in Taiwan

Po-yang Hsu[1], Yu-te Tom Kuo[2] & Chin-hsung Kao[3]

[1] Teaching Center of Physical Education, Nan-Hua University, Taiwan
[2] Department of Foreign Languages and Literature, Nan-Hua University, Taiwan
[3] Department of Recreation and Leisure Industry Management, National Taiwan Sport University, Taiwan

The purpose of this study was to work out a formula for criterion-referenced testing on men's health-related physical fitness in Taiwan. 1626 Taiwanese male adults, aged between 25 and 49, took part in this study. We used one's heart rate under resting conditions, ability to perform sit-ups per 30 seconds, per minute, flexibility in doing a V-shape sit-and-reach, and cardiorespiratory endurance index while taking a 3-minute step test to predict healthy or non-healthy. Multiple logistic regression analysis revealed not only cardiorespiratory endurance index differed significantly in those two groups, also work out a criterion-referenced testing. The best cutoff value was 0.575. The test validity was phi(Φ) coefficient 0.40 and cross validity 0.82. The test reliability was proportion of agreement 0.88, Cohen Kappa 0.34 and modified Kappa 0.76. The formula proved valid and reliable. It is also of good quality in setting up a touchstone for these citizens to assess their health-related physical fitness.
Asian Journal of Exercise & Sports Science, 2009, 6(1): 23-27

Keywords: cardiorespiratory endurance index, multiple logistic regression

Introduction

Of all the test-based assessments for physical fitness, the norm-referenced evaluation and the criterion-referenced test serve as the most common and efficacious approaches (Popham, 1971). The former interprets its readings, i. e., the examinees' scores, by reference to a norm; it locates a site on the graph and measures up the distance between the collective and the individual. While the latter pre-sets a cutoff score before the test and treats it as a divide for two mutually exclusive domains (sometimes even more than three) (Freeman & Miller, 2001); meaning often arises from the opposition. Rutherford & Corbin (1994) have said that the health-related physical fitness is the acceptable nadir in one's ability to adapt to the surroundings. That is also the least physical fitness needed in grappling with the lifestyle-related diseases or meeting the basic demands for survival. Furthermore, Kang (1994) also believes the norm-referenced scale is not suitable for testing one's health-related physical fitness, for it means to compare the individual and the group, rather than to indicate one's minimum fitness for the sake of health. The same percent grade means different in the distinct groups.

In America, activities designed for fitness assessment prospered in the late 70's, and were usually evaluated in a norm-referenced fashion for their efficacy. In the wake of these related researches on physical fitness and health, there has not been much room left for a science of multi-sample analysis out of sheer percentage assessment. Moreover, the "norms" deduced from such observations vary as time goes by, and likewise the target groups shift annually, failing to provide an ongoing panorama. This "percentage" methodology may lose its validity and reliability, on the condition that the background of its target participants fluctuates almost annually. The loss clears the ground for criterion-referenced tests, in which several typical defects in a norm-referenced manner may well meet their cures. Before the investigation, a criterion-referenced test offers a definite criterion, deemed unchangeable no matter how the examinee units or the exercise hobbies are likely to fluctuate. The test is applicable to all target groups, on the condition that the cutoff score is set in reason. So since the early 80's, the criterion-referenced mode enjoyed a vogue among the health-related-fitness associations, e.g., the AAHPERD (1980), Fitnessgram (CIAR, 1987), Physical Best (1988), Fit Youth Today and YMCA Youth Fitness Test (1989) etc. But it is really hard to lay down a criterion on which all judgments are based. In 1978, the criterion was set up as 50% (the middle notch on a linear measurement) by the Physical Fitness Test Development Program in South Carolina, based on its seasoned examiners' expertise advice and their available literature on norm moduli. However, the criteria laid down by Fitnessgram (CIAR, 1987), Physical Best (1988) and YMCA Youth Fitness Test (1989) did not achieve a remarkable development due to their frequent mean errors and subjectivity that debarred any systematic classification. Since the early 90's, a number of institutes have devoted themselves to the midwifery of a workable criterion. Cureton and Warren (1990) studied the criteria of the criterion-referenced tests conducted by Fitnessgram via

Corresponding Author
Po-yang Hsu
Teaching Center of Physical Education,
Nan-Hua University, Taiwan
E-mail: pyhsu@mail.nhu.edu.tw

a standardized measurement. Looney and Plowman (1990) amassed a large quantity of test data, out from which the best criterion could be sifted by categorizing the activities. They meant to find out a method of acceptable validity and reliability. At the turn of the century, Yau (2001) contended that health-related physical fitness usually put aside the results from battery tests, and came in want of completeness and integrity.

Therefore, the major purpose of this study is to find out the cutoff point that best defines the distinction between the healthy and the non-healthy, and to work out the criterion-referenced test of health- related physical fitness through the participant's heart rate under resting conditions, ability to perform sit-ups per 30 seconds, per minute, flexibility in doing a V-shape sit-and-reach, and cardiorespiratory endurance index while taking a 3-minute step test.

Methods

Participants

All participants of the "National Health-related Physical Fitness Test Program," conducted by Bureau of Health Promotion, Department of Health, are citizens of Taiwan. The test subjects' corpus in the survey includes the random samples of male citizens aged between 6 and 65. The participant sieving is made through a multistage stratified sampling procedure and according to Probability Proportional to Size (PPS). In this study, only a part of the repertoire—1626 adult males aged 25~49— are sifted and analyzed.

Test Items

The test items include measurement of blood pressure (systolic and diastolic), one-minute heart rate in resting conditions, body mass index (BMI) test for body composition, 30 seconds and per minute sit ups as measures of muscular strength and muscular endurance, V shape sit and reach test for flexibility, and 3 minute step test as the core measurement for cardiorespiratory endurance fitness.

Test Design

First, 20% database would be selected randomly for extra validity of the inference, and the rest 80% database could be designed for an assessment test. According to the 2007 standards set by Department of Health in Taiwan, men with normal blood pressure (systolic pressure < 140mmHg, diastolic pressure < 90mmHg) and BMI between 20 kg/m2 to 23.9 kg/m2, are deemed healthy, and those with a higher blood pressure (systolic pressure > 140mmHg, diastolic pressure > 90mmHg) and BMI more then 27 kg/m2, are regarded non-healthy. The data of these two categories serve well as dependent variables, predicted by the five independent variables as mentioned in the "Test Items" section. The analysis resorts to a multiple logistic regression for the best cutoff score to construct a tentative criterion-referenced test on men's health-related physical fitness in Taiwan.

Finally, we use the 20% data to measure cross validation of the 80% data. The ϕ(phi) correlation coefficient help us prove out the validity of this criterion-referenced test, and ultimately the reliability of the criterion-referenced test could be accessible by an intake of proportion of agreement (PA), Cohen Kappa coefficient (κ) and modified Kappa coefficient (κq). The data are analyzed in terms of the statistical software SAS 8e.

Results

Description of Healthy and Non-Healthy

In this nationwide test, the male participants aged between 25 and 49 amount to 1626. Among them, 1300 participants—circa 80 percent of the corpus—are chosen to be exemplars, including 351 labeled healthy, and 61 non-healthy. The other 326 samples, 20 percent of the database, undergo an extra verification, in which 107 participants are regarded healthy and 28 non-healthy.

In table 1, we can see the means, standard deviation (S.D.), minimal, maximal values, of two groups are showcased respectively under the binaries of healthy and non-healthy. The healthy group demonstrates a much lower pulsation rate and lower dispersion tendency than the non-healthy one. Besides, it performs better by reaching several notches above the non-healthy group in 30s, 60s sit-ups, Sit and reach, 3-minute step.

Table 1. Distinction between the healthy and non-healthy test subjects

Variables	Group	Num	Mean	S.D.	Min	Max	t value	p value
resting heart rate(bpm)	Healthy	351	76.61	11.24	42	119	-3.13	.01*
	Non-healthy	61	82.41	14.73	42	118		
Sit-ups in 30s(times per 30s)	Healthy	351	14.19	4.28	0	28	1.14	.26
	Non-healthy	61	13.53	5.13	0	23		
Sit-ups in 60s(times per 60s)	Healthy	351	25.42	7.70	0	50	1.27	.20
	Non-healthy	61	24.10	8.95	0	46		
Sit and reach(cm)	Healthy	351	24.18	11.24	0	55	1.87	.06
	Non-healthy	61	21.45	10.73	0	47		
3-minute step(CEI)	Healthy	351	59.50	10.95	30.41	98	5.88	.01*
	Non-healthy	61	51.45	11.55	16.55	85.71		

*$p<.05$

Difference Analysis of Predicted Variables

In table 1, t-value and p-value show the difference comparison of the aforesaid predicted variables. A marked difference arises, so far as the item "resting heart rate" is concerned. The average resting heart rate of those participants pigeonholed under the category of "non-healthy" soars higher than that of the healthy ones (82.41 bpm (beat per minute) vs. 76.61 bpm). In the 3-minute step test, we can see that the healthy participants demonstrate a much better cardiorespiratory endurance than the non-healthy group (59.50 vs. 51.45), which also marks a noteworthy difference in our study. In the other three tests, the two groups make no significant difference.

Multiple Logistic Regression

The predicted variable of resting heart rate is a negative quantity. That is to say, as one's resting heart rate rises, it signifies a less healthy and agile subject. So we use the index of resting heart rate in a reversed fashion. The other four predicted variables are all positive quantities. With multiple logistic regression, we choose the best predicted variables stepwise, and the SLENTRY in SAS regression equation is set as 0.01. The logistic regression is processing with the sequence of importance of the predicted variables being put into the model until the predicted variables show no significant difference (SLENTRY<0.01). At the same time, we eliminate the predicted variables with significant difference when put into the model (SLSTAY<0.01).

Intercept

Maximum Likelihood Estimation (MLE) could yield estimated value, and we are in need of intercept method to test such estimated value whether it has statistical meaning or not. The estimated value is 1.75 by MLE, standard error is 0.14, χ^2 is 159.14, p=.01(p<.05). This result shows significant difference, on the ground that the intercept is highly meaningful in statistics.

The selection of the predicted variables is based on the statistical values according to the SCORE. The five predicted variables are the participants' heart rate under resting conditions, ability to perform sit-ups per 30 seconds, per minute, flexibility in doing a V-shape sit-and-reach, and cardio -respiratory endurance index while taking a 3-minute step test, of which the value of SCORE χ^2 is 56.64, 10.32, 4.21, 2.29, 1.83 respectively. 3-minute step test, namely the index of one's cardiorespiratory endurance, is the only variable meaningful in their significant differences (p<.05). It is chosen to be initiated into the regression equation.

Regression Equation

Under the analysis of Maximum Likelihood Estimation (MLE), the predicted variables educed from the aforesaid stepwise elimination give us the estimated values of parameters as follows: intercept -6.78, cardiorespiratory endurance 0.16. The χ^2 of intercept is 28.82, cardiorespiratory endurance is 42.48. Both have significant difference (p<.05), a premise that verifies a hypothesis to exclude the intercept and cardiorespiratory endurance from being zero.

Moreover, the odds ratio of cardiorespiratory endurance is estimated at 1.17, a value indicative of a higher chance of "healthy" as one uplifts his cardiorespiratory endurance for one notch. The more one enhances his cardiorespiratory capacity, the healthier he must be. They are positively correlated. The regression-based equation of health physical fitness criterion-referenced test for adult males from aged 25 to 49 is as followed.

Logistic regression equation:

$$p = \frac{1}{1 + e^{-(-6.78+0.16\ \text{cardiorespiratory})}}$$

Evaluation of the Quality of Logistic Regression Model

In general, the SAS statistical software is designed to assess the quality of a model approach and to analyze the correlation between response values and the predicted probability values. The probability for concordance is 79.5%, the ratio for discordance is 19.7%, and the percentage tie is 0.8%. These values indicate that a good sifting mechanism—the one with higher percentage in concordance—can categorize the samples with better efficacy. Our model above is a quality one with a ratio of concordance soaring to 79.5%.

In addition, SAS provides three related indices which define the quality of a model. Somers' is 0.60, Gamma is 0.61, τ-a is 0.15, the higher the index is educed, the better the effect. In our case, except for its lowness in τ-a, the criterion-referenced test of the Taiwanese men's health-related physical fitness between 25 and 49 has good and cogent quality.

Cutoff Score

In this study, we have recourse to adjusting deviation for the purposes of classification and the best cutoff score, as shown in table 2. The range of probability is from 0 to 1, with an interval 0.005, and the values from 0.56 to 0.58 are extracted. The method of classification is based on whether the predicted probabilities are higher or lower than the cutoff score. If the former, it can be classified as correct truth result or incorrect truth result; if the latter, it can be deemed

Table 2. Classification of adjusting deviations

Probability level	Truth		False			Percentage			
	Truth result	False result	Truth result	False result	Correct probability	Truth positive	Truth negative	False positive	False negative
0.560	346	11	50	5	86.7	98.6	18.0	12.6	31.3
0.565	346	14	47	5	87.4	98.6	23.0	12.0	26.3
0.570	346	14	47	5	87.4	98.6	23.0	12.0	26.3
0.575	346	15	46	5	87.6	98.6	24.6	11.7	25.0
0.580	344	15	46	7	87.1	98.0	24.6	11.8	31.8

as correct false result or incorrect false result. Observing the different initializations of cutoff scores in table 2, we find that there comes the highest correct proportion 87.6% on the probability level of 0.575. Then, we can conclude that 0.575 is the best cutoff score and the validity of evaluation is highest which 87.6% is. In other words, we can say that 0.575 is the dividing line of the healthy and the non-healthy; the ratio of correctness has risen to 87.6%.

Reliability

A criterion-reference test can supply us with three kinds of reliability indices. PA (proportion of agreement) is the correct classify proportion under the situation of observed and predicted that has never been adjusted. In this study, the PA reaches 0.88. The second index, kappa coefficient, joined marginal product in order to revise the probability influenced in PA. A revision in our case results in a kappa coefficient, 0.34. The third index ushers in a more eclectic route, through which one may revise the marginal product of the kappa coefficient, making it 0.5, so as to improve the reliability. This modified kappa coefficient lifts its value to 0.76.

Table 3. Intrinsic data classification table (80%)

		Predicted		Total
		Health	Non-health	
	Healthy	346	5	351
		83.98%	1.21%	85.19%
Observed	Non-healthy	45	16	61
		10.92%	3.88%	14.81%
	Total	391	21	412
		94.90%	5.1%	100%

(1) Proportion of agreement:
PA=(346+16)/412=0.88
(2) Kappa coefficient
Pe=0.85×0.95+0.15×0.05=0.82
κ=(0.88-0.82)/(1-0.82)=0.34
(3) Modified Kappa coefficient
κq=(0.8786-0.5)/(1-0.5)=0.76

Validity

Intrinsic Validity

In our study, the intrinsic validity is represented by the phi(ϕ) correlation coefficient, a measure of association for two dichotomous variables, cited as follows.

$$\phi \frac{(346 \times 16 - 45 \times 5)}{\sqrt{351 \times 61 \times 21 \times 391}} = 0.40$$

Extra Validity

Extra validity assesses the validity of applying a survey-based inference to the entire population, which can be notified by cross validity. In this study, we reserved 20% of the data for an extra verification. There are 325 persons, including 107 labeled healthy, and 28 non-healthy. The cutoff score is 0.575, as on display in the classification table of cross data in Table 4. We find that the cross validity reaches 0.81 ((97+13)/135=0.81).

Table 4. Extra data classification table (20%)

		Predicted	Total	
		Health	Non-health	
	Health	97	10	107
		71.85%	7.41%	79.26%
Observed	Non-health	15	13	28
		11.11%	9.63%	20.74%
	Total	112	23	135
		82.96%	17.04%	100%

Discussion

1626 Taiwanese adult males, aged from 25 to 49, are selected as samples analyzed in this survey. These participants are divided into two groups—the healthy and the non-healthy—the former referring to those who have blood pressure in the healthy range (systolic pressure <140mmHg,and diastolic pressure<90mmHg) and whose BMI rests between 20 kg/m^2 and 23.9 kg/m^2, the latter to those who are symptomatic of hypertension (systolic pressure ≥ 140mmHg, and diastolic pressure ≥ 90mmHg) or whose BMI soars higher than 27 kg/m^2. 80 percent of the database provides the raw material for constructing a test; while the rest of it is reserved for a cross-test. According to the analysis, the healthy participants amount to 351, and the non-healthy ones total up to 61. In the cross-test data, the count of the healthy reaches 107, and the non-healthy comes up to 28.

We have recourse to a multiple logistic regression model for test analysis. The predicted variables selected as health indicators in this survey and its ensuing analysis include one's heart rate under resting conditions, ability to perform sit-ups per 30 seconds, per minute, flexibility in doing a V-shape sit-and-reach, and cardiorespiratory endurance index while taking a 3-minute step test. But only the cardiorespiratory endurance indices are put into regression modeling, and, having calculated the coefficients, we put down the regression equation as follows:

$$p = \frac{1}{1 + e^{-(-6.78+0.16\ \text{cardiorespiratory})}}$$

For example, if one gets 50 in a 3-minute step, then his equals 1.08 ρ=1/(1+$e^{-1.08}$)=0.75. We can come to the conclusion that a male with 50 in his cardiorespiratory endurance index has a 0.75 chance to be "healthy." Is the health probability 0.75 really high? If the cutoff score is clinched at 0.75, any who lifts his index higher than it will be shelved above the waterline. However, does the value prove to be the best cutoff score?

We resort to the Table of Adjusting Deviations for the best cutoff score in logistic regression analysis. The maximum of "correct probability" in this table is the best cutoff score which based on the analysis in this study, rests on the probability level of 0.575. A cutoff score is the ratio of probability that signifies a divide between healthy and non-healthy based on an analysis of multiple logistic regression.

However, the core problem lies in the question whether the dividing line bespeaks a valid and reliable distinction. This is also what a criterion is for in any criterion-referenced test.

Finally, we built up a checklist for reliability and validity. The best cutoff score, 0.575, as mentioned above, serves as the critical point between the observed and predicted values. We can use the classification bar, just as table 3 indicates. In the observed classification, the healthy group is composed of 346 persons who are regarded healthy and only 5 persons non-healthy. In the non-healthy group, 16 persons were assessed as non-healthy, but 45 of them were deemed healthy. The incorrect percentage is too high, that is the primarily reason of low reliability. With this study we have access to three kinds of reliability indices and two validity ones for reference. The values of reliability indices are respectively PA 0.88, Kappa coefficient 0.34, and modified Kappa coefficient 0.76. PA has a high quality. There has a poor value of kappa, because 16 in the non-healthy group are classified as non-healthy, whereas 45 out of it still defy the label of non-healthy.

The values of validity indices are φ coefficient 0.40 in intrinsic validation, which means the intrinsic validity of our samples' criterion-referenced test has showed that the correlation between the observed and the predicted only reaches a mild positive correlation in its validity. The cross validity is 0.81 in extra data. In contrast to the value of proportion of agreement (PA=0.88), it has carved a deficit of 0.07 (0.88-0.81 =0.07). The smaller the deficit is, the better the test can show in its validity of predictive application.

Conclusion

Four factors are taken into consideration in this study as the essential indices on one's health (body composition, muscle strength and muscle endurance, flexibility and cardiorespiratory endurance). The study comes to a conclusion: what matters most for men in Taiwan is their cardiorespiratory function. We derive three suggestions from the analysis. First, we discover a more accessible procedure in health index classification. That is, the first step, as educed in this study, is to make a healthy-or-not distinction in terms of the participants' BMI values. Once his concordance rate exceeds 64 percent, and his discordance rate drops down below 34.9 percent, the subject is seen as "non-healthy." If we add to the regression with a new variable, say, one's blood pressure, the concordance rate will rise up to 79.5 percent, and the discordance rate will plummet all the way to 19.7 percent. The output will also seem more appropriate. Second, the criteria of BMI and blood pressure are always notched down to the acceptable lows by the Department of Health in Taiwan. This policy signifies people can only be hypertension-aware before reaching the threshold of illness, and it also indicates that the alarming signs, not yet in sight, are destined to be precautionary failures. The classification remains blurring. For instance, an athlete may measure a little higher than 27 kg/m^2 and 140/90 mm/Hg in his BMI and blood pressure. But he/she may also show a high cardiorespiratory endurance index (CEI). If we consider this athlete non-healthy, we may prove erroneous, undermining the validity of the aforesaid regression model. Some scholars suggest that we can link the healthy and the non-healthy via a fuzzy theory model (Xie & Beni, 1991), blurring the borderline in the distinction; only in this way can we achieve a more satisfactory result. Last, we should test out some indices with higher validity and tighter relation to health-related physical fitness. Basically, size matters in sampling. When the sampling process is done on a large scale, some easily conducted test items may be manipulated for the sake of convenience, such as 3-minute step test for cardiorespiratory endurance capacity. When the survey is conducted goes on smaller scale, we would have recourse to some tests of higher validity, say, 800m walk-run test, for cardiorespiratory endurance (Heyward, 2006). Sampling on a larger or smaller scale is all up to the researcher. Our study here is only a preliminary step in testing out a more comprehensive test for health-related physical fitness evaluation. There is still some room for improvement, especially in regards to choosing the test items for higher validity, and working out a formula for predictive application. Hopefully we aim at constructing a quality means of fitness measurement.

References

American Alliance for Health, Physical Education, Recreation and Dance, AAHPERD (1980). Health related physical fitness test manual. Reston, VA.

The Cooper Institute for Aerobics Research, CIAR. (1987). Fitnessgram user's manual. Dallas TX.

Cureton J & Warren L (1990). Criterion-referenced standards for youth health-related fitness tests: A tutorial. *Res Q Exerc Sport*, 61(1), 7-19.

Forbus R (1990). The suitability and reliability of the physical best fitness test with selected special populations. Athens, Georgia: The University of Georgia press.

Freeman L, Miller A (2001). Norm-referenced, criterion referenced, and dynamic assessment: What exactly is the point? *Educa Psych in Practice*, 17(1), 3-16.

Heyward H (2006). Advanced fitness assessment and exercise prescription. Champaign IL: Human Kinetics.

Kang J (1994). Development of criterion-referenced standards for health -related physical fitness test. Paper presented at the Asian Sports Sci Congress'94 Hiroshima, held in Hirosima, Japan, 25-27.

Looney A, Plowman A (1990). Passing rates of American children and youth on the fitness gram criterion- referenced physical fitness standards. *Res Q Exerc Sport*, 61(3), 215-223.

Popham J (1971). Criterion-referenced measurement: An introduction. Educational Technology Publications, Inc., Englewood Cliffs, New Jersey.

Rutherford J, Corbin B (1994). Validation of criterion-referenced standards for tests of arm and shoulder girdle strength and endurance. *Res Q Exerc Sport*, 65(2), 110-119.

Xie L, Beni G. (1991). A validity measure for fuzzy clustering. IEEE Trans Patt Anal Machine Intell. 13(8), 841-847.

Yau HD, Ho RG (2001). Construct criterion- referenced measurement of service long test in table tennis. *Chinese Association of Psych Testing*, 41(1), 105-123.