
Physical education course is compulsory in higher education system of Iran. The curriculum focuses on physical fitness tests and student marks are apparently based on them. Students are dissatisfied with marking system and think it is unfair and biased. This destroys incentive and reduces the probability of a student's engaging in physical activity. This study aimed to modify the evaluation system of the course. During three subsequent semesters all participants were tested and marked by a new psychomotor test battery accompanied by new standards. The study included the whole population of 2651 female and 992 male subjects (age 18-24 years). Tests were administered by the same standards, but different administrators in every semester. Test Battery included Stork stand test, Sit-ups, 30 meter sprint, Sargent vertical jump, Sit and reach, Pull-ups, T-test, Trunk lift, Phantom chair, and PACER. Demographic data included age, weight, stature, and BMI. A new set of norms and standards were set for evaluation of the course. Distribution line-charts of new marks were compared to older ones. For females, distribution of marks became closer to ideals, whilst for boys it did not take place. Standardization of fitness testing and scoring neither resolved abnormal distribution of final marks nor eradicated dissatisfaction of students with their marks. It is likely that instructors had alternated final marks based on their subjective understanding of achievements of students in affective and cognitive domains. It seems likely if standardization of scores for affective and cognitive achievements is added to standardization of fitness scores, the marking system might improve.

physical fitness, marking, evaluation, and norm.

PE is a compulsory course in Iranian Universities. This imposes millions of dollars cost on higher education system each year. According to statutory outlines set by Ministry of Science, Research, and Technology (MSRT, 1982-2007) which should be legally implemented in Iranian universities, the objective of university physical education courses are multi-dimensional and cover physical, mental, social, cultural, moral, and emotional aspects of young people development. Non-statutory recommendations of physical education experts also insist that a physical education course not only should deal with physical performance but also should care for mental performance, attitudes and psychological traits (FPEESIU, 2002; p. 21). In theory, it is the same all over the world; physical education courses are highly valued and are supposed to assist the educational system for the purpose of multi-dimensional growth and development of children and young people. As an example, let's pay attention to more details from the National Curriculum for Physical Education in England (NCPEE). The guidance of the NCPEE states that:

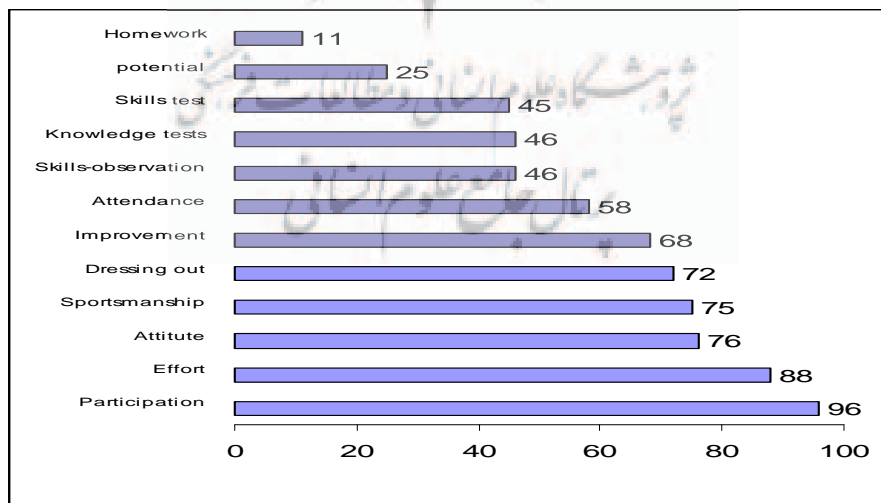
Later on, the Physical Education Association of the United Kingdom (1998) published a mission statement reinforcing the multiple ways in which physical education might contribute to education. It was stated that:



Here, physical education is clearly being associated with a diverse range of learning. In more recent reviews, it is suggested that NCPEE clarifies its stand and concentrates on more 'distinctive contribution' of Physical education to the school curriculum (QCA 1999a, and 1999b). These proposals incorporate the statement that:

Nevertheless, there is a big gap between theoretical knowledge and practices in the real world. For example, despite so much national and international documents insisting on multi-dimensional nature of physical education, in practice it does not fulfill its objectives in Iranian Universities (Kashef and Banian, 1999). In 2002, chairmen of PE faculties, heads of Academic Physical Education Bureaus, and prominent Iranian physical education lecturers gathered in a symposium to find out the ways to improve the quality of physical education courses. They all believed that university physical education is in a perilous position in all provincial regions of Iran (FPEESIU, 2002). Hardman and Marshall (2000) showed that the state of physical education in the UK and other countries do not comply with statutory guidelines mentioned above. They addressed many issues including legal status and actual implementation, inadequacies in human resources and teacher preparation, curriculum trends, as well as scepticism about the subject's future. In this paper, we are specifically going to have a close study about implications of multi-dimensional nature of physical education on its evaluation and marking system.

On what bases should physical education students be marked? In physical education, there are three domains of objectives, namely psychomotor, cognitive, and affective (Morrow et al, 1995: p. 132; Krathwohl et al, 1964). Therefore, assessment strategies and instruments should also assess some factors in every category. Here, the next question arises: what factors should be included in the mark and how should they be weighted? Further to physical performance tests, physical education instructors, intuitively or based on declared objectives of the course, include traditional objectives such as attendance, correct uniform, shower taking, leadership, attitude, sportsmanship, participation, discipline, and so forth. Figure 1 summarizes the reported attributes traditionally used for marking.



However, as there is still no standard approach to assess most of affective attributes, some measurement texts consider them as dubious objectives and recommend 'not to mark on them' (Morrow et al, 1995; p. 127; Hensley and East, 1989). It is suggested (Tritschler, 2001: p. 181) that these attributes should not be part of the mark, but they should be addressed by department policies. It seems such recommendation is unlikely

to be easily approved by instructors. It simply causes them to lose control over their classes. On the other hand, it contradicts the multi-dimensional nature of physical education courses and leaves the affective domain out of effect. If developing proper attitudes is one of the main objectives of physical education, should not it be reflected in the marks? Good marks act like a mirror reflecting students' achievement in terms of stated objectives.

This study focuses on the present state of physical education assessment system of a typical Iranian university. It searches for requirements needed for modification of its assessment and marking system. For any program, accomplishment of the objectives requires ongoing program evaluation. From students' perspective, one of the major considerations for evaluating the success of a course is 'fairness of its assessment system'. Subjective concept of fairness is not measurable. However, measurement experts believe that the quality of assessment would be guaranteed, if validity, reliability, and objectivity of measurements are obtained (Morrow et al, 1995; Tritschler, 2001). Assessment should have content validity. That is, if a course covers all 3 cognitive, affective, and psychomotor domains, the final marks of the students should also reflect their achievements in those 3 domains (FPEESIU, 2002; p. 21). Nevertheless, at the time of the study, instructors claimed that about 80-90% of the final mark of the students was based on their final fitness tests and only 10-20% of it was based on student's attitudes and knowledge. So far so good, apparently the system abides with guidelines stated in assessment textbooks;

. We could consider it as a proof for validity. But, for such an approach of marking, there was no unique standard and every instructor took his/her own subjective method. Consequently, qualities such as test-retest stability, inter-rater consistency, and even within-rater consistency might be at risk. In measurement language, it is said that the marks lacked reliability and objectivity. That is, with the same records, students could obtain low or high marks, depending on by whom, when, and how the test were administered and the records were interpreted. It seemed that the marking system of the university could be modified by standardizing the applied physical fitness tests, producing their norms, and teaching instructors to standardize their marking approach. Produced physical fitness norms could be helpful in a variety of ways. It would provide standards for a logical, fair, and realistic evaluation and therefore it could motivate students for more efforts. In order to investigate the usefulness of the modification, the physical education quality and distribution of marks will be compared before and after provision of the new standards and marking system.

: The study took place in Isfahan University, Iran. It included every student who enrolled for the PE course from 2001 to 2002 and completed it. 2651 girls and 992 boys participated in the study in 3 consecutive semesters. The norms were produced in the form of percentile ranks. The data were analyzed separately for each semester. They all were comparable and produced identical norms. It is considered as an evidence for reliability of test scores. The detailed report of reliability analyses comes in another paper. In this report, norms are based on data of 1179 females and 445 males who took part in the first semester of the study.

General measurements included age, height, and weight. Body mass index (BMI) was calculated by dividing weight by height square (Tritschler, 2000; p. 229). Other factors were measured as follows: Balance by Stork stand test (Beashel and Taylor, 2001: p. 83), Trunk endurance by YMCA 1-minute timed sit-up test (Morrow et al, 1995: p. 234), Agility and body control by T-Test (Baechel, 1994: p. 268), Anaerobic power by Sargent vertical jump test (Baechel, 1994: p. 265), Thigh endurance by Phantom chair test (Safrit, 1995: p. 60), Flexibility by Trunk lift (Safrit,

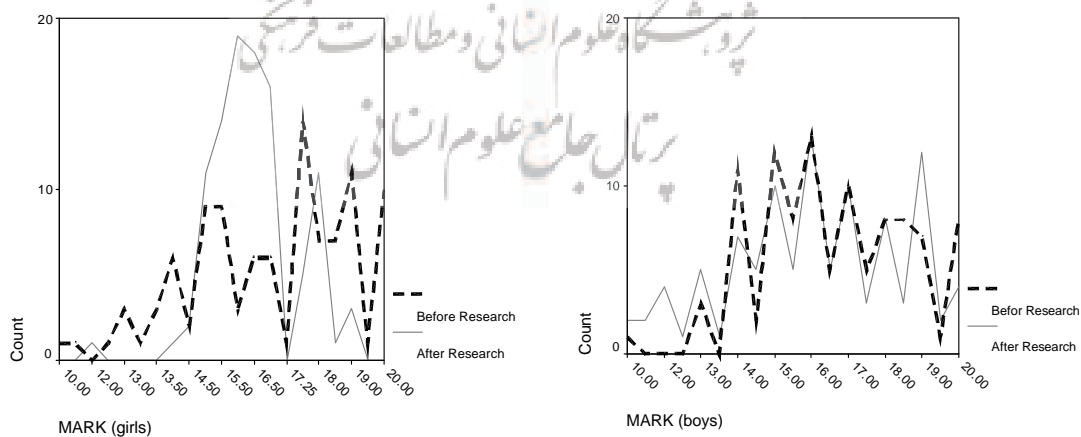
1995: p. 68) and Sit and reach tests (Baechel, 1994: p. 271). Aerobic power was measured using PACER test (Safrit, 1995: p. 59). For males, Upper extremity endurance was measured by Pull-ups test, and for females by modified Pull-ups test (Safrit, 1995: p. 55).

Table 1 shows the statistics related to measured factors for males and females, respectively. Data are compared using independent t-student test. Tables 2 and 3 show the physical fitness records and their corresponding percentiles and marks for females and males, respectively. Reliability of norms was confirmed as produced norms were almost identical for 3 consecutive semesters (relevant statistical analyses will be reported in another article).

Mean (\pm SD) of the measured items (males = 445, females = 1179).

Age (year)	19.7 (\pm 1.62)	19.6 (\pm 1.57)
Height (cm)*	175 (\pm 6.32)	159 (\pm 5.58)
Weight (kg)*	64.9 (\pm 11.30)	53.4 (\pm 8.32)
Body Mass Index (kg/m ²)	21.13 (\pm 3.32)	21.02 (\pm 3.16)
Stork (s)	25.4 (\pm 22.66)	12.9 (\pm 14.05)
Sit-ups (count)	40 (\pm 9.1)	31 (\pm 8.5)
Pull-ups (count)	5.8 (\pm 4.25)	5.5 (\pm 4.87)#
Phantom Chair (s)	114 (\pm 52.16)	87.8 (\pm 41.68)
30 Meter Sprint (s)	4.79 (\pm 0.413)	5.97 (\pm 0.436)
T-test (s)	11.67 (\pm 0.95)	15.48 (\pm 1.41)
Vertical Jump (cm)	50 (\pm 7.3)	31.27 (\pm 5.4)
Pacer (phase lap)	39 (\pm 9.04)	41 (\pm 6.7)
	33 (\pm 7.37)	36 (\pm 6.9)
	7.7 (\pm 1.48)	3.4 (\pm 1.03)

#=Modified Pull-ups is used for females. *= Males have a better achievement than females (p 0.05). = Females have a better achievement than males (p 0.05). cm= centimeter, kg= kilogram, s= second, SD= standard deviation.



Distribution of marks before and after application of the new marking system. Only distribution of 'female marks after research' was not statistically different from a normal distribution.

Figure 1 shows the distribution of PE marks before and after the study for a systematic random sample of 102 subjects in each independent group. In Iranian educational system, the marks range is 0 to 20 and a mark could be scored anywhere in this range. As shown in figure 1, the observed physical education mark range is actually between 10 and 20. The mark distribution pattern of boys remained the same

dressing out, improvement, attendance, skills observation, potential, and homework. Physical education instructors here at Isfahan University are customized to use the same kinds of attributes and between 10-20% of their marks include some attributes which belong to affective domain. They do not believe they are inappropriate attributes at all. They think if these attributes have a place in marking, they could better control their classes and it would consequently contribute to social, cultural, and lifestyle positive changes necessary for education of responsible and accountable citizens. Some instructors go beyond this and claim that physical gains from physical education classes are temporary and just means, but positive habits and attitudes as well as social skills and psychological traits which come with them are ends and what we should care for. For these instructors, more than 50% of marks might be derived from affective attributes. As we thought instructors were somehow right, we did not reinforced them to completely use the introduced approach for marking and let them to derive 10-20% of marks from their own affective evaluation.

Figure 1 showed something was wrong with the marks. As showed, the distributions of the marks were so platykurtic and skewed. That is, too many students had poor marks of less than 14 or top marks of more than 18 (Sharon et al, 2006; pp. 35-41). Further to abnormal distribution of marks, our qualitative analysis of interviews with students showed that many of them believed marks were biased and unfair. The marks were especially disappointing for students with poor marks. They simply did not understand why they should get such poor marks, whilst classmates with more or less the same achievements could make so better marks. According to their viewpoints, student with the same qualities of behavior and fitness test scores were marked differently by different teachers and in different classes. Measurement experts believe such marks lack the qualities such as reliability and objectivity (Sharon et al, 2006; pp. 121-162). Therefore, they are not neither understandable and acceptable by students, nor could they be incentives for lifestyle and behavior improvement.

To modify the marking system we had decided: 1. to give no mark less than 14 in order to avoid reducing students' motivation, 2. to base the marks on the same physical fitness tests, 3. and to use the same testing procedure and norms for final summative evaluation. A center was established for final administration of tests and calculation of marks. Marks produced by the center were suggestive and the right to revise them was reserved for instructors. Instructors changed the proposed marks based on their semester notes and their subjective evaluation about standards of behavior of students and their knowledge and understanding about the subject of course. This process of marking revealed that for physical education instructors the affective and cognitive domains are as important as psychomotor domain.

Logically, the final marks of students were based on their achievements on all three dimensions. It seems many instructors could not abide with 20% limit which was set for affective domain. Having a close thought on this process, we believe that the root of faults with our proposed system of marking was non-standardized approach to marking affective and cognitive domains in our physical education classes. We simply followed the guidance of assessment textbooks and underestimated the importance of affective domain for our colleagues. In future studies, we should give more attention and care to these domains. On the other hand, ongoing pressure of the program for development of physical fitness alone was one of the major obstacles for students to have fun and enjoy physical education. It seems counter productive and could hinder development of social and cultural abilities of students. It is as if the physical education course is not to improve exercise habits and positive attitudes toward sport and healthy lifestyle.

In Iranian universities, according to religious codes, physical education classes are gender restricted. Also, girls are taught only by female instructors and boys by male instructors. Statistical analyses linked to figure 1 showed that before the study, for neither males nor females the distribution of marks was normal. After the study only for

females the distribution became much better. Also, the number of marks less than 14 was more for males than for females. Apparently, compared to female instructors, male instructors were more reluctant to abide to new standards for evaluation and need more time for adaptation.

Generally, it is concluded that standardization of fitness test scoring could improve the evaluation of achievements in a physical education course, but it is not enough at all. The same standardization approach should be taken for scoring affective and cognitive achievements of students during a course. Also, compared to female instructors, male instructors seem more reluctant to apply a new system of evaluation for P.E courses and for their adaptation more time and effort is needed.

1. Baechel TR, 1994: Essentials of strength training and conditioning. Champaign; IL: Human Kinetics.
2. Beashel P, Taylor J, 2001: The world of sport examined. UK: Nelson.
3. Department of Education and Science (Welsh Office), 1992: Physical education in the national Curriculum. London: DES.
4. Faculty of PE and Exercise Sciences of Isfahan University, 2002: The report of booklet of first symposium of PE authorities of Iranian universities (in Persian). 5 Aug, 2002, Isfahan; Iran.
5. Hardman K, and Marshal J, 2000: The state and status of Physical education in schools in international context. *European Physical Education Review*, 6 (3): 203-229.
6. Hensley LD, and East WB, 1989: Testing and marking in the psychomotor domain. In MJ Safrit and TM Wood (Eds.): *Measurement concepts in physical education and exercise science*. Champaign; IL: Human Kinetics.
7. Kashef M, and Banian A, 1999: Iranian academic staff viewpoints about current curriculum and evaluation system of universities physical education course (in Persian). *Sports Science Research Center*. Tehran, Iran.
8. Krathwohl DR, Bloom BS, and Masia BA, 1964: *Taxonomy of educational objectives*. New York: McKay.
9. Kurpius SE, and Stafford ME, 2006: *Testing and Measurement; a user-friendly guide*. Thousand Oaks, California; Sage.
10. Ministry of Science, Research, and Technology, 1982-2007: *Statutory outlines for Physical Education Courses 1 and 2* (in Persian). Department of Education, Higher Council of Cultural Revolution.
11. Morrow JR et al, 1995: *Measurement and evaluation in human performance*. Champaign; IL: Human Kinetics.
12. Physical Education Association of the United Kingdom, 1998: *PEAUK Mission Statements*, *British Journal of Physical Education*, 29(2): 4-7.
13. Qualifications and Curriculum Authority, 1999a: *The review of the national curriculum in England: The secretary of state's proposals*. London: QCA.
14. Qualifications and Curriculum Authority, 1999b: *The review of the national curriculum in England: The consultation materials*. London: QCA.
15. Safrit M, 1995: *Complete guide to youth fitness testing*. USA: Human Kinetics.
16. Tritschler K, 2000: *Barrow and McGee's practical measurement and assessment* (5th ed.). Philadelphia: Lippincott Williams and Willkins.