Reliability of Field Based Cardiovascular Fitness Running Tests for Individuals With Mental Retardation

Ted Baumgartner and Michael Horvat
The University of Georgia

The primary purpose of this investigation was to determine the reliability of cardiovascular running tests for individuals with mild and moderate mental retardation. A secondary purpose was to determine whether cardiovascular running tests had some degree of validity in that running times for these tests tended to be in excess of 6 minutes. Data for eight groups were analyzed. Times for the distance runs ranged from 51 to 914 seconds while reliability ranged from .25 to .94, depending on the group. Based on the findings for the groups and distances used in this study, the following conclusions seem justified: (a) Distance run tests of 300 to 880 yds may be reasonably reliable for rank ordering subjects in terms of scores on these tests; (b) distance run tests of 300 or 600 yds are not valid measures of cardiovascular endurance for individuals with mental retardation; and (c) a distance run test of 880 yds may be valid as a measure of cardiovascular endurance.

Individuals with mental retardation are generally lower in measures of cardiovascular fitness than their intellectually normal peers (Fernall, Tymeson, & Webster, 1988). This has been substantiated in several research studies investigating the effect of training on severely and profoundly mentally retarded populations (Jansma, Ersing, & McCubbin, 1988; Mulholland & McNeill, 1985; Tomporowski & Jameson, 1985). The mixed results for improving functional performance in these investigations is in part due to variations in testing methodology as well as to heterogeneity in this population, even with high functioning individuals (Seidl, Reid, & Montgomery, 1987). In addition, there are problems in motivating individuals with mental retardation and in knowing whether they understand test requirements (Baumgartner & Horvat, 1988).

Recent attempts to validate cardiovascular field tests have met with varying results. Fernall and Tymeson (1988) reported a correlation of -.88 between the 1.5-mile run and VO$_2$max in mildly mentally retarded adults and a correlation...
of −.71 between the 300-yd run and \( \dot{V}O_2 \text{max} \). The researchers concluded that the 1.5-mile run was a valid indicator of cardiovascular fitness for mildly mentally retarded adults while the 300-yd run was not sufficiently valid.

In another study, Cressler, Lavay, and Giese (1988) investigated the test-retest reliability of four measures of cardiovascular fitness on 17 mildly to moderately retarded adults: a cycle ergometer, treadmill, step test, and 12 minutes run/walk. Results indicated a test-retest reliability of .93 for the treadmill test, .95 for the step test, .64 for the cycle ergometer, and .81 for the 12-min run/walk.

Other cardiovascular field and laboratory tests have been used with varying results and have tended to be population specific (Krahenbuhl, Pangrazi, Burkett, Schneider, & Peterson, 1977; Reid, Montgomery, & Seidl, 1985; Reid, Seidl, & Montgomery, 1989; Schurrer, Weltman, & Bammell, 1985). In all cases the need for standardization is emphasized as well as determining the validity for field based measures with this population.

Although running tests of 300, 600, 800 yds or farther can be administered to justify physiological adaptations to aerobic activity, a field based test must be validated on a specific population in order to ensure generalizations to the population (Fernall & Tymeson, 1988; Seidl et al., 1987). Too often tests of cardiovascular fitness are validated on nonhandicapped populations and used indiscriminately on individuals with mental retardation. Short runs are more a measure of speed than of cardiovascular fitness.

Authorities in physical fitness seem willing to accept a 6-min run as more a measure of cardiovascular fitness than speed. If subjects run for most of the time and the test duration is such that they run for at least 6 minutes, the test can be accepted as a measure of cardiovascular fitness. Clarke (1976) indicated that the 6-min run has been used for some time to evaluate the cardiovascular fitness of nonhandicapped youth. In addition, Disch, Frankiewicz, and Jackson (1975) using factor analytical procedures reported that the 6-min run appeared to be more of a cardiovascular test than a speed measure.

From a measurement perspective, the initial steps for testing validity begin with establishing reliability coefficients for the testing procedures employed (Baumgartner & Jackson, 1987). For a test to be valid, it must be reliable. In addition, questions concerning desirable distances that are the most reliable for field tests as well as for meeting the physiological basis for assumptions of cardiovascular fitness must be ascertained.

The primary purpose of this investigation was to determine the reliability of cardiovascular running tests. A secondary purpose was to determine whether cardiovascular running tests had some degree of validity in that running times for these tests tended to be in excess of 6 minutes. The population selected were individuals with mild and moderate mental retardation in public school classes.

**Methods**

**Procedures**

Teachers who had students with mental retardation in their physical education classes were requested to administer at least one distance run (300, 600, or 880 yds) to their students on several occasions so that reliability of cardiovascular running tests could be estimated. Classifications were determined based on the
Reliability of Fitness Tests

Guidelines of the Division for Exceptional Children, Georgia Department of Education (1988). IQ scores were based on the WISC-R, Stanford Binet, and adaptive behavior by the AAMR Vineland Adaptive Behavior Scales. Subjects with an IQ score of 55–69 were classified as mild, and moderate if IQ scores were 40–54. Directions concerning how to administer the distance run tests and record the data were sent to each teacher. The distance selected was based on the teacher's judgment and observation as to the subjects' ability to maneuver the running course and motivation to complete the test. One day was used for introduction and to practice running the distance selected.

Data for the selected distance were collected on two to four successive occasions to determine reliability of measurements over time. This was more data than are usually collected in a typical testing program, but it was necessary because it enabled the researchers to determine whether the mean performance of a group changed from day to day and whether one day of testing was sufficient to obtain a reliable test score for a subject. Rest intervals between days were included to ensure that tests were administered at least 1 and no more than 7 days apart. Generally the test was administered on the same day each week with the same distance being completed during each testing session.

Test procedures involved either running on a circular track or running from a starting line to a mark and back to the starting line. Times to run the test distance were recorded to the nearest second and observations were recorded on motivation of students, percent of time walking, and administrative concerns. The procedure was repeated on each testing day.

Subjects

Data were obtained on 13 classes of subjects. Usually these classes were composed of 4 to 13 subjects of both sexes and various ages. Sometimes these classes were composed of subjects with different classifications of mental retardation but with similar functional ability and adaptive behavior. Thus the researchers combined similar subjects from two to three similar classes together to form larger groups. Criteria used to form larger groups were that (a) all subjects ran the same distance (300, 600, or 880 yds), (b) subjects were similar in age, and (c) classes combined were similar in composition in regard to gender, classification, and age. All subjects in classes combined were not always used in forming larger groups. For example, classes of 10- to 15-year-olds and 9- to 11-year-olds were combined to form a larger group of 9- to 12-year-olds. Subjects in these larger groups tended to be a mix of males and females, several ages, and two classifications of mental retardation, but the groups were more homogeneous than a class.

It was possible that the larger groups formed were abnormally heterogeneous in performance due to the subjects' differing in gender, age, and classification which influenced the results of the data analysis. Thus other groups were formed with all subjects in a group being the same sex, classification, and approximate age.

Group 1 was formed by combining three classes of students. This group included 23 mild and moderately mentally retarded boys (n = 15) and girls (n = 8) 12 to 14 years of age. Group 2 was formed by combining two classes of students but only using the data of subjects ages 9 to 12. There were 9 boys and 5 girls in this group of mildly and moderately mentally retarded students. Group 3...
consisted of 26 moderately mentally retarded subjects 10 to 14 years of age in an adapted physical education program. There were 16 boys and 10 girls in the group.

Group 4 was formed by combining two classes of male \((n=6)\) and female \((n=6)\) moderately mentally retarded students ages 15 to 18 years. These subjects tended to be uncooperative and walked most of the 880-yd distance. Group 5 consisted of six mildly mentally retarded boys ages 12 to 14 in two classes. Group 6 comprised 7 moderately mentally retarded boys ages 12 to 14 in two classes. Groups 5 and 6 were tested on the 300-yd run. Group 7 included 6 moderately mentally retarded boys ages 12 to 14 in one class. They were tested on the 600-yd run. Group 8 had 15 moderately mentally retarded boys ages 12 to 14 in two classes who were tested on the 880-yd run.

Analysis

Subjects were tested on each of 2 to 4 days so that reliability of the specified distance run could be estimated and change in the mean performance of the subjects over days could be determined. The teacher of a class decided how many days to administer a distance run test. A two-way ANOVA (Subjects \(\times\) Days) with days being the repeated measure was applied to the data of each group. From this ANOVA design it was possible to test whether the means for days were equal. An alpha level of .05 was used. The reliability of a single day score was also estimated using the intraclass reliability coefficient presented by Baumgartner and Jackson (1987).

Reliability was estimated for one day because the researchers believed that this is consistent with field based testing programs conducted in the public schools. Typically subjects are tested on one day and the scores obtained are used as indications of their abilities.

Findings

Statistics for Group 1 on the 300-yd run are reported below. There was a nonsignificant difference among the trial means \((p>.05)\). The reliability of one trial of the 300-yd run was .63, which is low. All subjects completed the 300-yd run in under 3 minutes: range = 51–155 seconds; mean Trials 1, 2, and 3 = 97.30, 97.13, and 93.43, respectively.

Statistics for Group 2 on the 600-yd run show that, again, there was a nonsignificant difference among the trial means. The reliability of one trial of the 600-yd run was .69, which is low. Seventy-nine percent of the 600-yd run scores were less than 5 minutes: range = 175–344 seconds; mean Trials 1, 2, and 3 = 234.36, 232.00, and 231.14, respectively.

Statistics for Group 3 on the 880-yd run show that there was a nonsignificant difference among the trial means. The reliability of one trial of the 880-yd run was .75, which is good. Of the 104 scores \((26\ subjects \times 4\ trials)\), only 2 were less than 6 minutes: range = 313–914 seconds; mean Trials 1, 2, and 3 = 579.88, 587.85, and 559.88, respectively.

Group 4 was tested on the 880-yd run. There was a significant difference among the trial means \((p<.05)\). These subjects ran slower the second day than the first day. The reliability of one trial of the 880-yd run was .49, which is unacceptable. Seventy-one percent of the 880-yd run scores were larger than 6
minutes: range = 184–541 seconds; mean Trials 1 and 2 = 372.50 and 438.33, respectively.

Statistics for Group 5 on the 300-yd run show that there was a nonsignificant difference among the trial means. The reliability of one trial of the 300-yd run was .47, which is not acceptable. All subjects completed the 300-yd run in less than 3 minutes: range = 51–153 seconds; mean Trials 1, 2, and 3 = 80.33, 97.17, and 86.33, respectively.

Group 6 was tested on the 300-yd run. There was a nonsignificant difference among the trial means. The reliability of one trial of the 300-yd run was .84, which is quite good. All subjects ran the 300-yd distance in less than 3 minutes: range = 51–150 seconds; mean Trials 1, 2, and 3 = 103.14, 93.71, and 90.86, respectively.

Statistics for Group 7 on the 300-yd run are reported below. There was a nonsignificant difference among the trial means. The reliability of one trial of the 600-yd run was .25, which is not acceptable. All subjects completed the 600-yd run in less than 6 minutes: range = 240–332 seconds; mean Trials 1, 2, and 3 = 303.17, 293.83, and 276.50, respectively.

Statistics for Group 8 on the 880-yd run show that there was not a significant difference among the trial means. The reliability of one trial of the 880-yd run was .94, which is excellent. Eighty-two percent of the 880-yd run scores were larger than 6 minutes: range = 237–914 seconds; mean Trials 1, 2, and 3 = 559.07, 550.13, and 565.80, respectively.

Discussion

In this investigation most of the subjects tested were 9 to 14 years old. The eight groups differed in composition and length of the distance run test, which prohibited comparisons among groups.

The range in the scores for each group is quite large. Heterogeneity of scores within a group tends to result in higher reliability coefficients than when the scores are homogeneous. This has been a constant problem in evaluating this population. The low to excellent reliability coefficients (.63, .69, .75, .84, and .94) for Groups 1, 2, 3, 6, and 8, and the findings of nonsignificant differences among the trial means, suggests that administering one trial of a distance run test after the subjects have had some experience with the test may be acceptable. The unacceptable reliability for Group 4 (.49) is probably not representative of reliability values typically obtained because the group was so uncooperative. The unacceptable reliability for Groups 5 (.47) and 7 (.25) was due to the subjects’ being very inconsistent in their performance from trial to trial and exhibiting no pattern as a group in their inconsistency. Some subjects became better while others became worse across trials.

Further, the magnitude of the reliability coefficients does not seem to be influenced by the distance of the run or whether a group is homogeneous or heterogeneous in terms of gender, classification, and age. Both acceptable and unacceptable reliability coefficients were found for each distance of the run and for both homogeneous and heterogeneous groups.

If higher reliability is desired, administering several trials and using the best score or mean score as the criterion score would be more accurate as a criterion for evaluating this population. For example, the reliability for a criterion
score based on the three trials of the 300-yd distance run test for Group 1 was .84 rather than .63 for one trial.

These results are similar to Fernall and Tymeson (1988), who reported that VO₂max and 1.5-mile run values did not even reach the criteria for poor cardiovascular fitness. In contrast, the 12-min run incorporated by Beasley (1982) seems to indicate similar results while an earlier investigation by Rarick, Widdop, and Broadhead (1970) using the 300-yd run yielded slower scores than those in this investigation.

Likewise, the VO₂max and 300-yd run results in the Fernall and Tymeson investigation yielded a correlation of .71 and lends some support to the findings in this investigation. More specifically, the work of Lavay, Giese, Bussen, and Dart (1987) and Cressler et al. (1988) on reliability of cardiovascular measures is supported by this investigation. Cressler et al. also reported a reliability of .81 for the 12-min run.

It seems, both in this investigation and others with similar populations, that variable results are difficult to eliminate in a field based test. The score of the subjects is influenced by the effort required for running and the various moods, dispositions, and motivations they have. Further, the dissimilar characteristics of the groups even within a similar classification affect results of field based assessments. For example, the difficulty of getting subjects to run the entire distance without prompting will affect the distance traveled and subsequent time.

Repeating the same situation over trials is an ongoing problem with this population. Also, the variation in mental functioning within a classification is a problematic concern that will affect test results. Although individuals in mild and moderate classifications have a greater capacity for achievement than individuals in a severe classification, performance cannot always be linked directly with their classification. In a similar vein, lack of performance cannot be entirely based on the level of retardation. Systematic evaluation is required to determine performance gains or level of achievement for a group.

Previous researchers have indicated a concern for valid and reliable tests for individuals with mental retardation. For a test to be valid, it must be reliable (Baumgartner & Jackson, 1987). The majority of the reliability coefficients found in the present study do not suggest that any of the distance run tests used were highly reliable for the groups tested. Further, the subjects in this study usually completed the required distance in under 6 minutes. Disch et al. (1975) reported that the 6-min run appears to be more a cardiovascular test than a speed test. This suggests that distance runs of less than 6 minutes are more a measure of speed than of cardiovascular endurance. Thus, for subjects like those in this investigation, the 300- and 600-yd runs seem inadequate and not sufficient distances to be considered measures of cardiovascular endurance. The 880-yd run looks favorable as a measure of cardiovascular endurance for the subjects studied. Most subjects ran the distance in over 6 minutes, although some refused to run the entire distance. For younger subjects with mental retardation and/or those with other disabilities, the 880-yd run or a shortened distance may be acceptable.

As noted by Clarke (1976) and by Disch et al. (1975), and 6-min run can be used as a measure of cardiovascular endurance. Thus a sound procedure might be to administer a 6-min run test to a group of subjects to determine the distance completed so that this distance could be utilized in the future. A distance wherein the majority of the subjects run for at least 6 minutes should be selected. It seems
that a test in which subjects run a distance for time rather than time for distance is more conducive for test administration, and subjects tend to generate a greater effort. Running a distance for time is goal oriented and may ensure more reliable results.

**Conclusions**

Based on the tests and subjects used in this study, the following conclusions seem justified:

1. One trial of distance run tests of 300 to 880 yards may be reasonably reliable for rank ordering subjects in terms of scores on the tests.
2. Distance run tests of 300 or 600 yards are not valid measures of cardiovascular endurance.
3. A distance run test of 880 yards may be valid as a measure of cardiovascular endurance.

**References**


