MICHIGAN COMMISSION ON LAW ENFORCEMENT STANDARDS

THE DEVELOPMENT OF A FOUR-EVENT PHYSICAL FITNESS TEST

Career Development Section

September 2004
Reviewed and Updated 2017

Michigan Commission on Law Enforcement Standards
The Development of a Four-Event Physical Fitness Test

Introduction

The Michigan Commission on Law Enforcement Standards (MCOLES), formerly known as the Michigan Law Enforcement Officers' Training Council (MLEOTC), is the state agency that has the responsibility to promulgate standards for the selection, training, and licensure of law enforcement officers in the state of Michigan (MCL 28.601, et. seq., as amended). MCOLES sets "minimum standards of physical, educational, mental, and moral fitness [that] govern the recruitment, selection, and appointment of law enforcement officers" statewide (MCL 28.609, Sec. 9., (a)). As specified in administrative law, an important component of the MCOLES legislative mandate includes physical fitness determinations (R 28.14204).

A six-event physical fitness test for law enforcement was established by MLEOTC and had been in place in Michigan since the mid-1980s. Passing this test was a requirement to enroll in a basic law enforcement training academy and to be licensed as a law enforcement officer in the state of Michigan. The primary purposes of the test were to measure the fitness levels of those in the law enforcement applicant pool and to separate the physically fit from the physically unfit.

Since its inception, the test has been composed of six events: pushups, grip strength, obstacle course, 165-lb drag, 95-lb carry and a one-half mile shuttle run. The events were intended to measure speed, strength, agility, and endurance as identified by a physical skills job analysis (Wollack & Associates, 1979). All were timed events, with the exception of the grip strength measure. The cut score, or passing performance level, was based upon average fitness for the Michigan applicant pool. The test had been administered through fourteen regional test centers to those applicants considering employment, or pre-employment training, in anticipation of becoming law enforcement officers in Michigan. Detailed information regarding the design and development of the physical fitness events and the adherence of these efforts to professionally

accepted test development guidelines can be found in the respective project reports (Wollack & Associates, 1979, Stanard, 1981).

In 2000, MCOLES embarked on a major project to evaluate and review the six-event test in an effort to determine its current utility. The test had worked well over the past twenty years. But now, practitioners in the law enforcement profession have a much more sophisticated understanding of the design, implementation, and administration of such tests. Moreover, although passing the six-event test was required for entry into the academy and for the activation of licensure, there existed no statewide physical fitness or health/wellness training standards at the state level. Many of the twenty-three academies had some form of fitness program in place, but no standardized training was being delivered.

Accordingly, the MCOLES physical fitness evaluation project had two major objectives: first, to re-evaluate the utility of the discrete events in the six-event test, along with their accompanying cut scores, and secondly, to create a physical fitness standard (a training curriculum) that would be mandated for basic academy training.

Evaluating the Six-Event Test

In evaluating the design of the six-event test, several issues immediately surfaced. It was determined that the following components needed to be thoroughly explored and re-evaluated:

- the cut score;
- age and gender norming;
- test equipment; and
- measures of physical *fitness* or job-related physical *tasks.*

Each of the above issues is discussed in detail below.

On the six-event test, the minimum qualifying score for females is 28 converted points and the minimum qualifying score for males is 29 converted points. The scoring methodology is such that raw test scores are converted into “stanine scores,” which are based on the bell curve distributions of each event. The range for stanine score is 0 through 9, so a stanine score of “5” represents the mid-range of performances on the bell curve distribution.

In scoring the applicant performance a cumulative number of points was totaled, rather than a pass/fail point for each specific event. Therefore, poor performance on one event could be compensated for by an outstanding performance on another event. For example, an applicant who could run quite fast could also do quite poorly on the push-up event but still pass the test. Should that same applicant be working as a law enforcement officer in the future and need to rescue an injured person, he or she could not make up for a lack of strength by running fast. Clearly, past job task analyses have established that all underlying constructs, or dimensions, of physical fitness are important for successful performance as a law enforcement officer (Personnel Research Consultants, 1979; Stanard & Associates, 1996). A true fitness test, then, should measure total body fitness, not just fitness in specific areas. The scoring scheme used in the six-event test did not measure overall fitness. Therefore, a requirement to pass each event independently (conjunctive scoring) needed to be seriously considered during the development of a new physical fitness test.

The six events are also scored according to gender. This bifurcation produces a situation where males are in competition with males and females are in competition with females. Years ago, it was decided to gender-norm the test to control for the innate physiological differences between males and females. Although females are required to perform at a lesser absolute standard, the cut score essentially divides their performances into categories of “fit” and “unfit.” The same methodology is then used to derive the cut score for men. The test is not age normed.

Further, expensive equipment was needed to administer the test. When the test was originally designed, specifications detailed by MCOLES (MLEOTC) were mandated for each official test site, and each was required to make the appropriate purchases. For example, the sites were required to purchase a life-form, weighted dummy for the 165-lb drag and a dynamometer in order to measure grip strength. In considering a new fitness test, it was hoped that it would be easily administered, have a high degree of portability, and be free of expensive or elaborate equipment.

A close analysis of the six events reveals that three of the events measure "pure" fitness (push-ups, grip, run) and three of the events are rather loose approximations of job task simulations (obstacle, drag, carry). Recent court rulings have addressed the issue of measuring physical *fitness* as opposed to measuring physical *requirements* for the position of law enforcement officer (*Alspaugh, Lanning*). Gender and age norming seem to be allowed by the courts if the test is intended to measure fitness rather than job related requirements. Although the six-event test is intended to be an overall fitness assessment, and is supported as such by a recent state court ruling (*Alspaugh, 2001)*, moving exclusively to either fitness or job requirements needed to be seriously considered.

To determine if there were a significant connection between pure fitness and job task simulations in terms of performance, Dr. John Berner, a professional psychometrician, conducted statistical analyses of the existing six-event data. Dr. Berner correlated performance on the three "fitness" events with the performance on the total test. The analyses demonstrated that there was a statistically significant correlation between fitness and overall performance (R=.847; P=.000; N=3350). The results of this analysis suggest that measuring pure fitness is a viable method for determining fitness for performing job tasks.

Fitness and Training

Based upon these considerations, the staff began their initial research in mid-2000. A decision was made at the outset to establish and mandate a fitness-training program, or curriculum, at the recruit level in Michigan. To that end, MCOLES contracted the services of the Cooper Institute for Aerobic Research (CIAR), of Dallas, Texas, to assist in the creation of a physical fitness and health/wellness curriculum for the academies. The intent was to integrate the test and the curriculum such that the test would be used both to assess the fitness of the incoming recruit as well as be a standard for successful completion of the physical fitness training program.

Using the test as part of a training standard, in addition to being a pre-enrollment requirement, serve two purposes. First, administering the test allows it to function as an assessment instrument

during the physical fitness training curriculum. The test can enhance the training by ensuring the candidates who enter are of sufficient fitness to benefit from the fitness conditioning and development done during the training. Secondly, the same test can be used to assess fitness upon exit to ensure that the recruit has attained the level of fitness needed to pass the course and enter the law enforcement profession in Michigan.

The staff believes that the new test is a simple but effective method for assessing a recruit's level of physical fitness. As a baseline test (diagnostic test) it can assess individual fitness levels prior to the commencement of the physical fitness training in the basic academy sessions. This allows the physical fitness instructor an opportunity to establish a "fitness profile" for the trainee and to set individualized goals for improvement. Failure to pass this initial diagnostic test could serve as a warning that the trainee is unlikely to attain the higher fitness levels necessary to pass the test at the end of training. Then, as an exit test it can be used to determine whether a recruit passes or fails the MCOLES physical fitness curriculum. The exit test should be administered during the final weeks of the academy session, just prior to graduation.

Moreover, placing the test and the fitness training in proximity to the time the recruit completes training better prepares the recruit when entering the job market. Under the old testing program the physical test could easily be one year or more ahead of the time of employment or academy enrollment. Once the candidate passed the six-event test there were no further fitness requirements that had to be met. In summary, the new four-event test can be used in three ways: 1) as a pre-enrollment requirement, 2) as a baseline diagnostic assessment for the physical fitness program, and 3) as a pass/fail level of performance.

It should be pointed out that although the fitness curriculum has been written and is now mandated in the academies, the remainder of this report focuses on the development of the physical fitness *test* and the determination of a reasonable pass/fail level of performance.

Test Validity and Reliability

As part of its contractual agreement with MCOLES, Cooper Institute representatives provided assistance in selecting the appropriate events for the physical fitness testing process. As a first step, the staff identified the tasks required of entry-level law enforcement officers and the physical skills necessary to carry out such tasks by examining Michigan's job task analyses. An initial job task analysis was conducted for MCOLES in 1979 and then updated in 1996. Detailed information regarding the job task analyses, including a listing of core and non-core tasks for the position of law enforcement officer can be found in the project reports (Personnel Research Consultants, 1979; Stanard & Associates, 1996).

Appendix D of the *Stanard* (1996) job task analysis identifies sixteen unique physical ability core tasks performed by law enforcement officers in Michigan. In addition, a review of the professional literature supports the need for physical fitness in the law enforcement profession (Cooper, 2001; Gebhardt, 2000, 1998; Hoffman & Collingwood, 1995; Hogan, 1991; Reintzell, 1990; Safrit, 1989; LaDou, 1982). Clearly, physical fitness is a core underlying construct.

The identification of the underlying fitness "constructs," as supported by the job task analyses and the research, lead directly to the concept of "test validity." Test validity is defined as ensuring that the test measures what it purports to measure (Mehrens & Lehmann, 1984). According to Mehrens & Lehmann, (1984), leading expert in the field of testing,

> When a test score is used to make an inference about a property or behavioral domain of the person measured, we can think of the test score as *representing* the property of that person. This is a reasonable inference to the extent that the test items do actually represent the behavioral domain. (Mehrens, p.289).

Moreover (Standards, 1999),

> Construct validity (is) a term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include…internal test structure…as well as the content of the test (p. 174).

Essentially, a construct is a characteristic that a test is designed to measure. In providing criteria for the validity and reliability of tests, nationwide standards for educational and psychological testing have been established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (American Educational Research Association, 1999). These *Standards for Educational and Psychological Testing* call for evidence of test validity (standards 1.1-1.24). This includes evidence of construct, content, or criterion validity. In the development of the physical fitness test by MCOLES, the nexus between the underlying fitness constructs identified in the job task analysis and the events designed to measure fitness in the applicant pool demonstrates construct validity. The connection between the pure fitness events and the job task events of the six-event test, as identified by Berner, supports content validity. The fundamental purpose of both the physical fitness program and the test is to shape behavior in a positive way and to create a permanent life-style of fitness.

Therefore, to increase the fidelity of the test, and to permit a valid interpretation of test scores, the staff explored a variety of physical performance events in an effort to identify four or five pure fitness assessments that could be used as pre-enrollment standards and as measures of fitness in the academy setting.

Test "reliability" is also an important consideration, as addressed in the *Standards for Educational and Psychological Testing.* Test reliability is defined as "the degree to which test scores for a group of test takers are consistent over repeated applications" (*Standards,* 1999, p. 180). Any test would need to function in the same manner whether it were administered in an Upper Peninsula location or in downtown Detroit. To ensure test reliability, individual events would need to be administered uniformly, be standard across the state, and work in a variety of physical and environmental settings.

Identifying Test Events

As the project progressed, representatives from Cooper defined "fitness" as upper, middle, and lower body strength in conjunction with aerobic and anaerobic capacity (Cooper, 2001). The challenge for the staff was to identify appropriate physical fitness events that would be consistent with the definition provided by Cooper and, at the same time, remain reliable, valid, and easily administered by law enforcement academies and law enforcement agencies statewide.

Initially, Cooper proposed a five-event test to measure fitness, which consisted of a 300-meter sprint, a vertical jump, sit-ups, push-ups, and a 1.5-mile run. A review of this initial proposal revealed that control over the conditions under which the test would be administered would be difficult to standardize. For example, Cooper recommended that the 300-meter sprint be run in a straight line and that the 1.5-mile run be administered on a track no smaller than 220 yards. Given Michigan's wide seasonal swings, this meant that these events would have to be done on an indoor track. The staff discovered that only three or four indoor tracks would meet such criteria and be available for use by the 23 academies.

Alternative events were also examined (e.g., air dyne bikes) but ultimately rejected because of expense and, in some instances, a lack of supporting normative data. The staff decided to investigate if the existing shuttle run could be substituted for the 300-meter sprint and the 1.5-mile run. Cooper ultimately decided that the shuttle run was a measure of both anaerobic and aerobic capacities and would support its use by MCOLES as a fitness measure. They speculated that the aerobic/anaerobic mix ran from 40/60 to 60/40 depending on the fitness of the individual taking the test. From the original work done by Wollack & Associates in 1979, the shuttle run was determined to be a measure of cardio-respiratory fitness.

Ultimately the staff determined that four events could validly and reliably measure fitness:

- vertical jump;
- sit-ups;
- push-ups; and
- one-half mile shuttle run.

The vertical jump measures lower body strength, the sit-ups measure core body strength, the push-ups measure upper body strength, and the shuttle run measures both aerobic and anaerobic capacity. The sit-ups and the push-ups are timed events, with the number performed in one minute being the recorded score. The events are easily administered, minimal equipment is needed, and the events can be performed in the facilities available to all academies.

The score on the vertical jump is determined by the distance between the examinee's reach height and the best of three recorded jumps in inches. A vertical jump test board is required for this event at a minimal cost to the test sites. To measure the examinee's vertical jump height, one Velcro spool is placed by the examinee on the jump board at the highest vertical reach by one arm to measure a baseline reach height. Then, the examinee is required to jump as high as possible with the outstretched arm and place a second spool on the board. The proctor then measures the difference to the nearest ½ inch to obtain the score.

The sit-ups are performed with the examinees flat on their backs with the knees bent at 90 degrees and the hands overlapped behind the head. For one sit-up, the examinees are required to touch their knees with their elbows and then return their shoulders back down flat on the mat. The number performed correctly in one minute is the score.

The push-up event is a standard, full-body push-up and is performed by both males and females. The examinees are required to touch their chests to a 3" indicator affixed to the floor and then back up to the elbow-locked position. The number performed correctly in one minute is the recorded score.

The one-half mile shuttle run is a timed event that requires the examinees to complete 15 round trips between two pylons placed 88 feet apart. The time it takes to complete the event is the recorded score.

Because of the test's relative ease of administration and portability, use of the test outside the academy setting is encouraged. In so doing, candidates can easily determine whether they are of sufficient fitness to enter the police academy and would be likely to pass the fitness requirements.

Such pre-testing can identify areas in need of improvement. Similarly, law enforcement agencies can pre-screen candidates to increase the likelihood that the candidates will successfully complete the academy program.

To enhance test reliability, the staff wrote a proctor manual for use during the instructor training sessions and for test administration. It is important that the test be administered uniformly at each test location; it is equally important that the test administration be consistent from one test to the next. The proctor manual specifies the facilities and equipment requirements, and the precise definitions and instructions to be used for each event. The proctor manual also addresses the reporting and recording of the test scores.

The Cut Score

Once a determination was made as to which events would compose the four-event test, and how the test was to be used, the next step was to construct a sound, structured methodology by which the appropriate normative data could be generated, analyzed and evaluated. Individual and group performances needed to be examined in order to determine an appropriate cut score, to observe how perfomance could be bifurcated along age and gender lines, and to identify various fitness levels of the law enforcement applicants in Michigan. Initially, MCOLES explored the possibility of simply adopting the so-called "Cooper norms," frequency tables of fitness performances provided by the Cooper Institute categorized by gender and age. Law enforcement agencies and state standards boards across the nation quite often use these norms. For example, New York, Idaho, North Dakota and New Hampshire use Cooper percentages for entry level law enforcement testing. The Cooper Institute provided MCOLES with the frequency distributions (norms) for the sit-up, push-up, modified push-up, and vertical jump. Although the frequencies were categorized according to age and sex, Cooper was unclear as to which candidate pool was used to generate their normative data. In fact, the frequency table for each particular event seemed to be based upon a different candidate baseline pool, none of which were law enforcement applicants per se, with the possible exception of the modified push-up event.

Certainly, Cooper did not use the Michigan applicant pool to generate physical performance levels. Any test score is more meaningful when compared to a relevant and well-defined norm group. The most meaningful norms are those developed specifically for an organization, using their applicant pool. For these reasons, the staff eventually rejected the use of national, or generic, norms for its statistical analyses and thought it feasible to develop local norms from the Michigan applicant pool.

The staff then set about establishing a methodology to collect and analyze Michigan data. The staff determined that a three-pronged approach to the collection and analyses would be the best method. First, the plan called for data to be generated by administering the test to a sample pool of test takers located in Michigan, a sample pool that would statistically and characteristically represent the existing applicant pool. Secondly, MCOLES decided to hire the services of a professional psychometrician to assist in the pilot-testing and analyses.

Through a state open bid process, it was ultimately decided that Dr. Susan Stang, Performance Based Selection, Ltd., of Westlake, Ohio, (PBS) would act in that capacity. The third "prong" of the methodology called for the staff to assemble a group of fitness subject-matter-experts in order to evaluate MCOLES' statistical analyses and determine, in a practical sense and with a practitioner's eye, what a reasonable and acceptable level of performance should look like in Michigan. It was believed that the final work product should include direction and input from those in the law enforcement field in Michigan who possessed the requisite experience, expertise, and insight into physical fitness and fitness testing. The essential purpose of any group process is to generate ideas that eventually lead to creative solutions and a qualitative understanding of the priorities of the group (Novak and Gowan, 1984; Huff, 1990). It was believed that such an interactive group process would work well in Michigan.

Field Testing and Data Collection

As a starting point for the data analyses, MCOLES used previous performances for two of the events: the push-ups and the shuttle run, since these events had been previously administered in Michigan (six-event test). What was needed was a sample that included performances for the vertical jump and the sit-ups. It was ultimately decided to administer the entire four-event test to a sample of recruits statewide to generate the performance levels needed for statistical analyses and to reasonably determine an appropriate cut score for the test. Those entering the academies, it was believed, would statistically represent the applicant pool.

A total of 695 examinees participated in the field-testing (Table 1). Test proctors were trained to administer the test in a standardized manner and each was provided with a proctor manual. Recruits entering the academies during the fall sessions of 2001 and the spring sessions of 2002 took the test and were directed by the test proctors to perform to their maximum levels. Performances were recorded on forms provided by MCOLES. Raw scores were recorded, entered into a database, and analyzed using the *Statistical Package for Social Science* (SPSS).

Table 1
Field Test Sites

Site	N	Percent
Oakland Community College	45	6.5
Lansing Community College	56	8.1
Kalamazoo Academy	74	10.7
Washtenaw Community College	24	3.5
Ferris State University	49	7.1
Lake Superior State Univ	21	3.0
Flint Academy	19	2.7
Northern Michigan University	21	3.0
Delta Community College	49	7.1
Kellogg Community College	30	4.3
Kirtland Community College	22	3.2
Macomb Community College	79	11.4
West Shore Community College	17	2.4
Wayne County Sheriff Department	34	4.9
Detroit Police Dept	120	17.3
Schoolcraft College	35	5.0
Total	695	100

The following tables display the demographic and performance characteristics of the sample examinees. The demographics and performances of the academy sample pool were then compared to the existing applicant pool and found to be significantly similar. The staff was therefore confident that the sample statistically represents Michigan's law enforcement applicant population.

Table 2
Gender of Examinees in Sample Population

Gender	N	Percent
Male	560	81
Female	135	19
Total	695	100

Table 3
Age of Examinees in Sample Population

Age	N	Percent
18-29	500	72
30-40	153	22
40+	42	6
Total	695	100

Table 4
Average Performances and Standard Deviation for Males (Field Test)

Event	N	Mean	S.D.
Jump	555	20	3.6
Sit-ups	560	39	9.0
Push-ups	560	42	14.5
Run	557	4:12.3	29.3
Valid N	552		

Table 5
Average Performances and Standard Deviation
for Females (Field Test)

Event	N	Mean	S.D.
Jump	134	13	2.5
Sit-ups	135	33	9.6
Push-ups	135	19	11.0
Run	134	5:04	53.0
Valid N	133		

Next, frequency distributions were generated for each of the events, bifurcated according to gender. A frequency distribution displays statistics that are useful for describing performances for specific events. The values can be arranged in ascending or descending order and the corresponding cumulative percentages can be examined. Such an analysis is essential in gauging the performances of the sample population when setting a pass/fail level. An examination of Tables 4 and 5 reveals a clear distinction between the performances of males and females.

The percentile, which is based on the cumulative frequency distribution, indicates the percentage of people in the norm group who received lower scores. It represents a person's relative performance in a group. For example, a person with a percentile rank of 70 has achieved a higher score than 69 percent of the participants in the reference group.

As an example, the statistical output for the male push-ups from the experimental pool is displayed in Table 6 on the following page. The data are arrayed in categories of five percentage points. By examining these normative data, one can see that the 50^{th} percentile requires a performance of 41 pushups. This is the average performance level for this one event. Similarly, to be in the 90^{th} percentile of performance, an examinee must perform 62 push-ups.

The staff also examined the more detailed frequency distributions of all the events. Valid performance levels are associated with the corresponding percentiles. For example, in the full distribution, 41 push-ups correspond to the cumulative percentage of 51.6, or an approximate

average performance level for males. Tracing the cumulative percentages from lowest to highest categorizes performances from lowest to highest in the sample population.

Table 6

Pushup Frequency Distribution
Percentiles

Percentile	Pushups
5	22
10	26
15	29
20	31
25	33
30	35
35	37
40	38
45	40
50	41
55	43
60	44
65	46
70	48
75	51
80	53
85	57
90	62
95	70

Frequency distributions can be used to compartmentalize a particular event into selected categories of performance. For example, those below average could be considered unacceptable for law enforcement tasks, or perhaps poor, average, and superior categories could be established from the distributions. The staff generated the frequency distributions for all events from the sample population and subsequently prepared the data for distribution to the physical fitness subject-matter-experts.

Dr. Stang and her staff reviewed and evaluated the output in preparing to consider a reasonable cut score. A state standard-setter such as MCOLES has the prerogative to set a cut score, as long as it is done in a reasonable and rational fashion. Dr. Stang emphasized that setting

a cut score is a value judgement and an administrative criterion. The primary requirement is that it be set using a sound, well documented, rationale.

Validity and Reliability Revisited

Although great care was taken to identify those events that would contribute to one's understanding and knowledge *fitness* as a construct, the MCOLES staff also set about assuring that the measurements themselves were of such a quality that accurate and meaningful inferences could be made when interpreting the test results. Often, researchers use rather sophisticated statistical techniques to analyze their data, once collected, but often do not take care in assuring that the individual items used for measurement actually contribute to the construct being measured. In other words, the staff wanted to know if the four events selected – jump, sit-ups, push-ups and run – reliably and validly measure the underlying construct of fitness.

During the design phase, the staff asked several fundamental questions. First, how confident can one be in making inferences about those taking the test, that is to say, how well does the test discriminate among test takers (examinee separation)? Secondly, do the four events create a well-defined construct called "fitness" (item reliability)? Third, to what extent and in what manner do the four events contribute to an understanding of the underlying construct? Do the events contribute in an equal way or are there varying degrees in their contributions? Finally, where along a common hierarchical continuum would the events be distributed?

In order to address these questions, MCOLES used the Rasch item response model for polytomous data (Wright & Masters, 1982: Bond & Fox, 2001). The staff selected a convenience sample of 81 examinees from Grand Rapids Community College and the Flint Police Academy and subjected their raw physical performances to Rasch statistical procedures. The staff then examined person and item reliability estimates, standard errors in measurement, difficulty levels, person abilities, and model fit statistics based on the Rasch estimations.

As can be seen in Table 7, the reliability statistics for both persons and events is 0.75 and 0.82 respectively. The person reliability of 0.75 indicated that the measurement scale discriminated

relatively well among the test takers. Thus, the staff was confident in making inferences about the examinees' abilities from their performances. Similarly, the item reliability of 0.82 demonstrated that the events created a well-defined construct, that is, the items are unique and distinct, but measure one competency.

The outfit mean square statistic of .97 indicated that the events fit the Rasch estimations well, as do the z-std statistics of -0.3 and -0.2. Outfit measurements are unweighted estimates of the degree of fit to the model estimations and are expressed in terms standardized *z* or *t* scores. Outfit statistics are sensitive to unexpected extremes, whereas infit statistics are weighted estimates that give more value to on-target observations. For model fit, the analyst looks for outfit and infit statistics near 1 and z-std statistics near 0. The logit measures indicate how close the events are to one another in terms of their contribution to the construct. The outfit statistics in Table 7 indicate that all four events fit the Rasch expectations and are therefore suitable as measures of fitness.

Table 7

SUMMARY OF 81 MEASURED PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	11.5	4.0	-.02	.71	.97	-.3	.97	-.3
S.D.	3.2	.0	1.61	.15	.76	1.0	.77	1.0

REAL RMSE	.81	ADJ.SD	1.40	SEPARATION	1.73	PERSON RELIABILITY	.75
MODEL RMSE	.73	ADJ.SD	1.44	SEPARATION	1.99	PERSON RELIABILITY	.80
S.E. OF PERSON MEAN	.18						

SUMMARY OF 4 MEASURED ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	233.0	81.0	.00	.15	.99	-.1	.97	-.2
S.D.	15.8	.0	.36	.00	.03	.2	.01	.1

REAL RMSE	.15	ADJ.SD	.33	SEPARATION	2.16	ITEM RELIABILITY	.82
MODEL RMSE	.15	ADJ.SD	.33	SEPARATION	2.17	ITEM RELIABILITY	.83
S.E. OF ITEM MEAN	.21						

Figure 1 displays the comparison of the examinees and the events on a common hierarchical scale represented in logits. The examinees are plotted to the left of center and the events are located to the right of center. The better performing examinees appear toward the higher end of the continuum and the lesser performing examinees appear toward the lower end of the continuum.

Each "X" symbol represents two examinees and the "#" symbol represents one examinee, where N = 81. Similarly, events that contribute less to an understanding of fitness are located toward the higher end of the continuum. That is, it is more difficult to determine its contribution to the underlying construct and therefore less likely to be endorsed by the staff. The Rasch model establishes the midpoint logit values at zero (M+M) along the measurement scale. S and Q are one and two standard deviations from the mean respectively.

In examining Figure 1 one can see that all events are located in close proximity in terms of their contributions to an understanding of the construct. Push-ups and sit-ups contribute slightly less and the shuttle run and jump contribute slightly more. It can also be seen that the push-ups differentiate among those of higher ability slightly better than the sit-ups, jump, or shuttle run.

The run and jump perhaps differentiate among the lesser performing examinees better than the push-ups. The sit-ups work best in the midrange of student ability. The examinee abilities are spread rather evenly across all levels but resemble the characteristic bell curve. The four events are located near the mid-range of the examinees' abilities so the test is not beyond the ability of the examinees. The four events have slightly different degrees of contribution to the construct, but all are quite close together on the scale. Table 8 displays the individual logit measurements for each event.

Figure 1

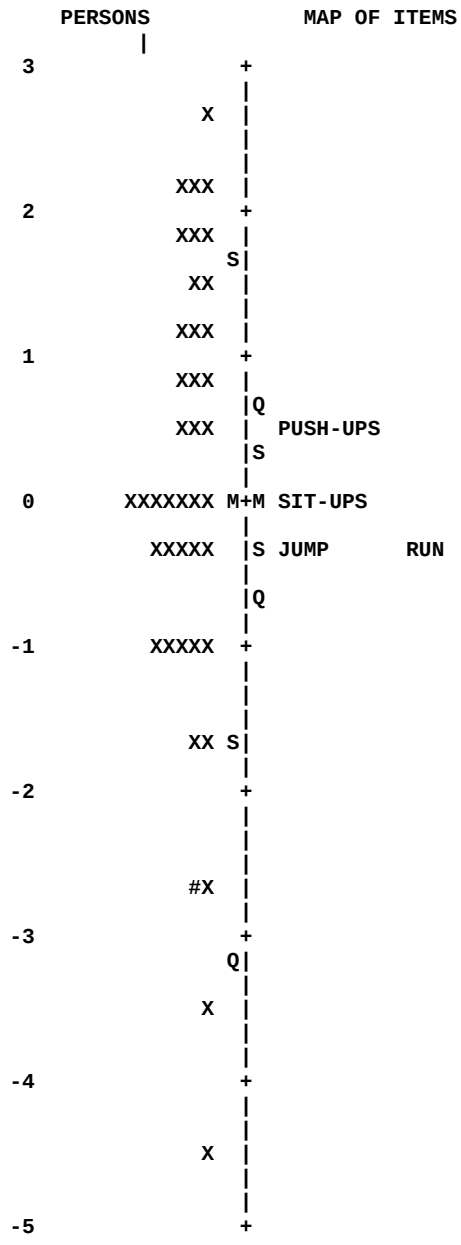

Table 8

ITEMS STATISTICS: ENTRY ORDER

ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	ITEMS
1	244	81	-.25	.15	.98	-.1	.97	-.2	.56	JUMP
2	229	81	.08	.15	.95	-.3	.99	-.1	.50	SIT-UPS
3	209	81	.55	.16	1.01	.1	.95	-.3	.49	PUSH-UPS
4	250	81	-.38	.15	1.03	.2	.96	-.2	.51	RUN
MEAN	233.	81.	.00	.15	.99	-.1	.97	-.2		
S.D.	16.	0.	.36	.00	.03	.2	.01	.1		

Subject-Matter-Expert Meetings

In July 2002, the staff met with the subject-matter-experts (SMEs). A list of the participants appears as Appendix I of this report. The participants were selected based upon their expertise in physical fitness training as well as their knowledge of the position of law enforcement officer in Michigan. Some are academy directors, some are physical fitness trainers, and some are experts in the area of fitness; all are current practitioners in Michigan. The staff distributed the data analyses from the field tests and solicited input from the SMEs. The primary purposes of the meeting were to discuss the viability of the selected four events as measurements of performance and to discuss ways in which a reasonable cut score for the pre-enrollment test could be established. The meeting agenda included the following items:

- selecting events for testing;
- setting a cut score;
- requiring a total score v. passing each event separately;
- norming for age and gender; and
- setting an exit score.

The staff solicited input from the participants at the meeting. At the outset, the group agreed that the four events identified by MCOLES were appropriate for physical fitness testing and evaluation. Next, the staff distributed preliminary statistical output from the sample group of examinees to the participants, which displayed the performance levels. The preliminary data

were based on the frequency distributions for each event and the raw performances of the examinees. After evaluating the data, the group agreed that setting a cut score at the "average" performance level would be a reasonable determination of fitness, based upon passing each event *separately*. Such a cut score would effectively divide the examinees into a group of "fit" people and a group of "unfit" people, for purposes of accessing Michigan's law enforcement training program. In other words, the staff would look at the performances of the sample examinees and set a cut score where 50 percent of the sample would "fail" the test and 50 percent of the sample would "pass" the test. The participants believed that one retest should be allowed for each event failed by the candidate. The SMEs also discussed the feasibility of gender and age norming and initiated preliminary discussions regarding a methodology for setting an appropriate exit standard. At the conclusion of the meeting it was decided that the staff would continue to analyze the data from the field test, set a reasonable cut score at the average fitness level, and consider whether to gender and age norm the performances.

Age and Gender Norming

In setting an appropriate cut score, age and gender norming became an important consideration. Should the passing mark be set at the same level for both men and women and should all age groups perform at the same level? Some experts and practitioners argue that both men and women should perform at the same level since they will be performing the same job tasks once hired by an agency (Cooper, 2001). Others believe that there are clear physiological differences between the genders that must be taken into consideration and that appearing to promote gender bias in law enforcement is unacceptable (*Alspaugh, 2001)*.

The courts have been somewhat conflicted, but three rulings have brought some clarity to the issue of gender norming. In *Alspaugh v. Michigan Law Enforcement Officers' Training Council*, 246 Mich App 547 (2001), the court upheld MCOLES' gender norming of the six-event test. In their claim, the plaintiffs argued "that the performance skills test is not designed to assess general physical fitness, but rather, designed to measure the minimum physical skills necessary to be a

police officer and that gender norming the scores gives preferential treatment to female candidates thus constituting unlawful affirmative action" (p. 4). In defending the gender norming, MCOLES maintained "that ignoring the immutable physiological differences between males and females as regards the performance skills test would disproportionately exclude female candidates from that pool of individuals eligible for certification as police officers" (p. 4). The court agreed with MCOLES. Here, the judges ruled that if the test were designed to measure general physical fitness rather than establishing minimum job-related fitness standards for employment, gender norming would not violate any civil rights principles in Michigan.

In *Lanning v. Southeastern Pennsylvania Transportation Authority (SEPTA)*, 528 U.S. 1131 (2000), the court had to decide if the physical requirements set by SEPTA, requirements that were the same for men and women, were the minimum necessary to perform successfully as a SEPTA transit officer. The court ruled that the SEPTA physical fitness requirements met this burden.

Essentially, the court ruled that using the same scores for men and women on "employment screening examinations is impermissible unless shown to measure minimum qualifications necessary for successful performance of a job in question" (p. 1). In other words, using identical performance levels for men and women is permissible, but a hiring agency must demonstrate job-relatedness. It should be noted that *Lanning* applies specifically to SEPTA, which has unique job tasks and responsibilities because of its work environment and is even distinct from other transit authority police departments. Moreover, through its travels through the appeals process, the Third Circuit suggested in a note that SEPTA could achieve their goals by using separate cutoff scores without violating Title VII of the Civil Rights Act. Title VII prohibits sex discrimination (42 U.S.C. sec. 2000(e)-16(a)) and the use of different cutoff scores on employment tests (42 U.S.C. sec. 2000e-2*(1)*).

In 2016, in *Bauer v. Lynch* (Docket No. 14-2323), the United States Court of Appeals for the Fourth Circuit ruled against a Federal Bureau of Investigation trainee who claimed the FBI discriminated against him on the basis of sex, as prohibited by Title VII, in that female trainees

are required to do fewer push-ups than male trainees. The FBI test battery consists of sit-ups, push-ups, a 300-meter sprint, and a 1.5-mile run. It is a gender-normed evaluation where minimum standards were required for each gender. The judges stated, "Put succinctly, an employer does not contravene Title VII when it utilizes physical fitness standards that distinguish between the sexes on the basis of their physiological differences but impose an equal burden of compliance on both men and women, requiring the same level of physical fitness in each" (p. 25-26). The judges remanded the case back to the lower court.

The MCOLES four-event test is not intended to identify minimum requirements to become a law enforcement officer. Instead, the intent is to assess general fitness, thereby creating a pool of applicants from which agencies can choose. In other words, the mission of MCOLES involves *inclusion* in order to widen the applicant pool. The mission of agencies is, to a large extent, to *exclude* those not qualified for specific job tasks.

In considering the issues about gender norming, and in light of the court rulings, the Commission directed the staff to continue to gender norm the new four-event test. Accordingly, the staff set about examining the performance distributions of male and females separately in determining an appropriate cut score.

Since the six-event test was not age normed, this issue became a consideration as well. On a rational basis, and in consultations with PBS, the staff felt that age norming the test would be necessary to control for the "concomitant decreases in muscular strength, endurance, and aerobic capacity attributable to the aging process" (*Alspaugh, p. 1*). The age distribution of the sample population, and therefore the applicant pool, was skewed toward the younger age groups.

Therefore, strictly grouping the sample examinees according to five or ten year increments was not realistic, although frequency distributions provided by the Cooper Institute and the US Army, and other normative data sets are often categorized as such. Instead, such categorizations should be identified by examining the performances of a specific pool of test takers. Perhaps even

identifying one age group as “40 years and older” would be consistent with the spirit and letter of the federal Age Discrimination Employment Act.

An analysis of the sample data supported this thinking. Statistical analysis revealed that, given the relatively young age of the sample population, age groupings of 18-29, 30-40, and 40+ would be appropriate categorizations. The data analysis consisted primarily of a factor analysis of the raw scores for each event. This statistical procedure “identifies underlying variables, or factors, that explain the pattern of correlations within a set of observed variables” (SPSS, p. 313). Here, the factor is age. The analysis organizes and categorizes the performances on the test according to age groupings and plots the data on a two-dimensional map, based upon the strength and direction of the individual correlations. Then, a hierarchical clustering procedure produces “cluster centers” around which the various age groupings gather. The program eventually settles on the final cluster centers after a specified number of iterations, or best fit determinations. Ultimately, cluster centers were identified at ages 23, 30, and 41. Accordingly, based upon both an intuitive rationale and statistical analyses, the staff determined that these three age groupings, as identified through the factor analysis, would be used in scoring the four-event test.

In summary, the staff made a number of decisions regarding the test. The decisions were based on the statistical analyses of the field test data, on the advice and consultations with the fitness subject-matter-experts, and through independent consultations with a professional psychometrician. In the end, the four-event test would have the following characteristics:

- specific events that measure pure fitness;
- a reasonable cut score that separates fit candidates from unfit candidates;
- age and gender categories; and
- a requirement that the candidates pass each event separately.

The Cut Score

Based upon the methodology outlined earlier in this report, the staff conducted one final analysis of the experimental data in an effort to set an appropriate cut score. The data were analyzed separately for men and women, and separately for the three age categories. In the

sample pool, it was believed that the participants, who would have already passed the six-event test, would be of slightly higher fitness level than the general applicant pool. But by examining the data in detail, the staff discovered that the participants were not, in fact, in any better physical shape than those in the general applicant pool. Apparently, enough time had past so the academy recruits had an opportunity to "get out of shape" after previously passing the test.

To determine the average fitness level of the sample pool, the analysis consisted of producing a 50 percent passing rate and a 50 percent failing rate for all examinees. Those performing below average would not pass the test and those performing at average or above average would be able to enter the training academies. It should be pointed out that the scoring methodology used here produces a cumulative effect when the examinees are required to pass each event separately. Every attempt was made to set the cut score at the "average" fitness level of the Michigan applicant pool, however, that does not mean that each *specific* event is set at its average performance level. For example, the average number of push-ups for a young male in the applicant pool is 42 in one minute. The new standard calls for at least 32 push-ups in one minute. But the new standard also calls for minimum performances in three other events as well. This cumulative scoring methodology affected the way the pass/fail numbers were ultimately determined. Separately scoring sub-categories on any test increases its overall difficulty.

In a practical sense, however, 50 out of every 100 examinees were not expected to fail the test once the live administrations began. Instead, because the test is much less complicated than the previous six-event test, it was expected that most candidates would practice the events to improve their performances, or perhaps "self-select" out of the process if they were unable to perform. Similarly, it was expected that many law enforcement agencies would pre-screen their candidates prior to sending them to an official test. And, allowing a retest for each failed event was expected to raise the raw passing rate during the live sessions. Table 9 displays the final standard of performance for the entry physical fitness test.

Table 9
Entry Level Standards

MALES

AGE	VERTICAL JUMP	SIT-UPS	PUSH-UPS	SHUTTLE RUN
18-29	17.5	32	30	4:29.6
30-39	16.0	30	30	4:38.2
40+	15.0	30	28	4:54.7

FEMALES

AGE	VERTICAL JUMP	SIT-UPS	PUSH-UPS	SHUTTLE RUN
18-29	11.0	28	7	5:35.4
30-39	9.0	19	7	5:59.1
40+	8.0	18	7	6:13.3

It should be pointed out that the staff conducted a number of informational meetings during the course of the fitness project. The members of the Commission on Law Enforcement Standards and the academy training directors were consistently apprised of the progress of the project. Similarly, the staff worked with their Curriculum Advisory Committee in establishing the training curriculum and worked closely with the subject-matter-experts and independent consultants to produce the four-event test. The staff conducted several train-the-trainer sessions at locations throughout the state. These sessions prepared the physical fitness cadre to deliver the physical fitness and health/wellness training in the academies. The full Commission officially approved the performance standards of the four-event pre-enrollment test at their October, 2002 meeting.

Exit Standard

At the December 12, 2002, Commission meeting, the staff was directed to begin field testing an exit standard, a standard higher than the entry-level standard. At that time, the Commission established the entry pass-fail level as an interim exit standard until an official exit standard could

be determined. Accordingly, since January 2003, students have been required to perform at the entry-level in order to pass the physical fitness curriculum at the completion of their academy training. Exit level field-testing began in the spring of 2003. Its purpose was to allow the staff an opportunity to observe and document how well the students would perform in a real-life environment, given an experimental set of higher performance requirements.

During field testing, students were asked to perform to their maximum levels on all four events. In previous consultations with subject-matter-experts, testing experts, and exercise physiologists, the staff determined that the students should perform, on average, approximately 10-20 percent higher on the exit assessment. This conclusion is based on the composition of the physical fitness curriculum, the time devoted to improving fitness during the academy session, and the nature of the event itself. Statistically, this represents an increase of one-half standard deviation above the average for each event.

The exit test consists of the same four events as the entry test. During exit testing, however, students are administered two major assessments, each consisting of two attempts. Should a student fail an event during the first attempt, he or she is allowed an immediate retest on the failed event. Should a student fail the second attempt, a 72-hour rest period is required for muscle recovery before the second major assessment, complete with two more attempts, is administered. Failure on the second assessment results in a failure of the physical fitness course and subsequent dismissal from the academy. Table 10 displays the entry and exit performance standards.

Table 10
Entry and Exit Pilot Numbers

MALE

Age	Jump		Sit-ups		Push-ups		Run	
	Entry	Exit	Entry	Exit	Entry	Exit	Entry	Exit
18-29	17.5	19.0	32	36	30	37	4:29.6	4:11.8
30-39	16.0	17.5	30	34	30	37	4:38.2	4:18.2
40+	15.0	16.5	30	34	28	35	4:54.7	4:27.8

FEMALE

Age	Jump		Sit-ups		Push-ups		Run	
	Entry	Exit	Entry	Exit	Entry	Exit	Entry	Exit
18-29	11.0	12.0	28	32	7	12	5:35.4	5:02.6
30-39	9.0	10.0	19	23	7	12	5:59.1	5:19.0
40+	8.0	9.0	18	20	7	11	6:13.3	5:25.5

Eighteen (18) academies reported field test results to MCOLES. The data are taken from academy sessions during a 16-month period (April 2003 through July 2004). To date, 896 administrations of the exit test have been analyzed in the field test pool. As displayed in Table 11, 841 test takers passed at the higher experimental level and 55 test takers either failed or declined further testing after initial failure. The overall pass rate is 94 percent.

Table 11
Pass/Fail Rates for the Pilot Test

	Number	Percent
Pass	841	94
Fail/Decline	55	6
Total	896	100

The raw numbers seem to indicate that approximately six percent of the students will fail at the higher recommend level. However, field testing never occurs in a perfect environment. A more accurate picture would emerge if the students had taken advantage of their full range of attempts and assessments. What happened was this: although some students were unsuccessful after their first try at the exit level, they knew they had passed the course at the *interim* standard and therefore declined any further testing. For example, after the first attempt 60 students were eligible to continue testing. Thirty-nine (39) declined further testing and were marked as a "fail" for field-test purposes. The staff believes, however, that if the students knew that the higher

levels were mandated, they would undoubtedly put forth more effort to pass at the higher level and would take advantage of their full range of assessments.

The staff presented the full Commission on Law Enforcement Standards with their findins regarding the exit standard. The commissioners formally approved the exit standard at their October 27, 2004 meeting, to become effective for academy training sessions that begin on or after January 1, 2005.

APPENDIX I
Physical Fitness Subcommittee

Participant	Agency
Dan Antieau	Wayne County Regional Police Academy
Mike Bath	Northern Michigan University
Jerry Boerema	Kirtland Community College
Dave Bower	Michigan State Police
Ralph Galvin	Washtenaw Community College
Brian Johnson	Grand Valley State University
Bill Martin	Lansing Community College
Mike Metz	Macomb Community College
Jeff Munoz	Michigan State Police
Susan Stang	Performance-Based Testing, Ltd.
Kathy Vonk	Ann Arbor Police Department

References

Alspaugh and Kujawa v. Michigan Law Enforcement Officers Training Council a/k/a Commission on Law Enforcement Standards, 246 Mich.App. 547 (2001).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D.C: American Educational Research Association.

Americans with Disabilities Act (A.D.A., 1990). Pub. L. No. 101-336, 42 U.S.C. sections 12101 et. seq.

Bond, T. & Fox, C. (2001). *Applying the rasch model.* London: Lawrence Erlbaum and Associates.

Civil Rights Act of 1991, S. 1745, 102nd Congress, (1991).

The Cooper Institute for Aerobic Research. (2001). Independent consultation for the Michigan Commission on Law Enforcement Standards.

Gebhardt, D.L., Baker, T.A., and Sheppard, V.A. (1998). *Volume 2: Q-2 police officer physical performance test and medical guidelines development and validation report.* Hyattsville, MD: Human Performance Systems, Inc.

Gebhardt, D.L. (2000). Establishing performance standards. In S. Constable and B. Palmer (Eds.) (2000) Wright-Patterson AFB, OH: Human Systems Information Analysis Center (HSIAC-SOAR).

Hoffman, R. & Collingwood, T. (1995). *Fit for duty: The peace officer's guide to total fitness.* Champaign, IL: Human Kinetics.

Hogan, J.C. (1991). The structure of physical performance in occupational tasks. *Journal of Applied Psychology, 76,* 495-507.

Huff, A. (1990). *Mapping strategic thought.* New York: John Wiley and Sons.

LaDou, Joseph. (1982). Health effects of shift work. *The Western Journal of Medicine,* December, 1982.

Lanning v. Southeastern Pennsylvania Transportation Authority, 528 U.S. 1131 (2000).

Mehrens, W. and Lehmann, I. (1984). *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart and Winston.

Michigan Compiled Laws. MCL 28.609(1)(a).

Novak, J. and Gowan, D. (1984). *Learning how to learn.* Cambridge: Cambridge University Press.

Personnel Research Consultants. 1979. Statewide job analysis of the police patrol officer position. Unpublished report to the Michigan Law Enforcement Officers Training Council.

Phillips, S.E. (1993). *Legal implications of high-stakes assessment: What states should know.* Oak Brook, IL: North Central Regional Educational Laboratory.

Psychological Services, Inc. (1981). Development of job-related statewide entry-level police officer selection and training standards. Unpublished report to the Michigan Law Enforcement Officers Training Council.

Reintzell, J. (1990). *The police officer's guide to survival, health and fitness.* Springfield, IL: Charles C. Thomas.

Safrit, M.J. & Wood, T.M. (1989). *Measurement concepts in physical education and exercise science.* Champaign, IL: Human Kinetics Books.

Statistical Package for Social Science. (1999). *SPSS base 9.0 user's guide.* Chicago, IL: SPSS, Inc.

Stanard & Associates. (1996). Statewide job analysis of the patrol officer position. Unpublished report to the Michigan Commission on Law Enforcement Standards.

Wollack & Associates. (1979). A job analysis of police physical skill requirements. Unpublished report prepared for the Michigan Law Enforcement Officers Training Council.

Title VII of the Civil Rights Act of 1964.

Wollack & Associates. (1981). Validation of entry-level police officer employment tests. Unpublished report prepared for the Employment Standards Section of the Michigan Law Enforcement Officers Training Council.

Wright, B. and Masters, G. (1982). *Rating scale analysis.* Chicago: MESA Press.