THE ESTIMATION OF MAXIMUM PHYSICAL PERFORMANCE

by

Roger Johnson
M.S. Florida State University

and

Duane Meeter
Ph.D. Florida State University
FSU Technical Report M386

Department of Statistics
Florida State University
Tallahassee, Florida 32306

Some key words: Criterion score; Maximum performance estimation; "Best" versus "average" scores.

*Research of the senior author supported by the National Institute of General Medical Sciences under Training Grant 5-T01-GM00913.
THE ESTIMATION OF MAXIMUM PHYSICAL PERFORMANCE

by

Roger Johnson
and
Duane Meeter

Abstract

Research Quarterly articles dealing with whether to use the best score or average score as the criterion score for multiple trial physical performance tests are briefly reviewed. The main purpose of this study is to compare by accepted statistical criteria the best and average scores as estimators of a subject's maximum performance. The results suggest that unless the variance of the measurement error is much larger than the within-subject variance, a rare occurrence, the best score will usually be the better estimator of maximum performance. A consistent estimator of maximum performance is also constructed.

*Research supported by the National Institute of General Medical Sciences under Training Grant 5-T01-GM00913.
In the field of physical education considerable controversy has persisted over the selection of a statistic which measures or characterizes physical performance when several trials, repeated tests under identical conditions, are available. Several articles in the *Research Quarterly* have discussed this question and no conclusive answer has been reached. The problem of how to estimate a subject's maximum physical performance also remains unanswered. This article is an attempt to resolve these questions.

**Review of the Literature**

The problem of selecting a criterion score for physical performance when several trials have been observed has been considered by many different authors (1, 2, 3, 6, 8, 10). However, it seems that a more precise definition of the statistical aspects of this problem is required before firm conclusions can be reached. The criterion score as considered by these authors is not necessarily an estimator of any physical quantity or parameter. It is simply a score characterizing the several trials in some unspecified way. In most cases the size of a test-retest correlation coefficient, as a measure of reliability, was used as the basis for selection of the criterion score.

Whitley and Smith (10) considered the criterion score problem and concluded that in strength tests the average of the several trials should be used as an individual criterion score, rather than the best score. This conclusion was based on correlation coefficients of 0.66 and 0.57, respectively, between the average of twenty scores on a test of speed of arm movement of a heavy object and (1) the average of four arm strength
scores and (2) the best of four arm strength scores. The coefficient of 0.66 was significantly greater than the coefficient of 0.57. Implicitly, Whitley and Smith have assumed that the two tests measure quantities that are closely related and that the sample correlations should reflect this fact. Correlation, however, is a measure of the linear association between two random variables. A bivariate normal distribution guarantees a linear association. The joint distribution of the average will undoubtedly more closely resemble a bivariate normal distribution than the joint distribution of the best and the average. It is possible that the best and the average score were more highly associated than the two average scores, but the association was nonlinear, allowing a smaller value for the correlation coefficient. This, rather than a greater degree of association, may be the reason for the results found by Whitley and Smith.

Berger and Sweeney (3) questioned the conclusion of Whitley and Smith maintaining that the variability of the data should influence the selection of the criterion score. Employing a test-retest correlation procedure they, in effect, reached the conclusion that the decision to choose an individual's best score or average score should be based on the ratio of between-subject to within-subject variance. As this ratio increases so does the desirability of selecting the best score rather than the average score, and vice versa as it decreases. Because Berger and Sweeney's data are not presented in their paper and because their procedure was somewhat unclear to the authors, further evaluation of this paper will not be attempted.

Kroll (8) attempted to present a theoretically defensible procedure for the selection of a criterion score based on the principles of reliability analysis. Kroll concluded that if the necessary assumptions of
reliability analysis are satisfied, i.e., uncorrelated, random error with zero mean, then the proper procedure is to use the mean of all available trials as the criterion score.

Alderman and Banfield (1) calculated reliability coefficients for eight strength measures by the test-retest method and found no significant advantage for the use of best scores over average scores.

Henry (6) stated that the purpose of his paper was "to explain the logical basis for the findings by Whitley and Smith and to demonstrate through the use of random numbers the mechanisms involved when such numbers are classified on a variability basis to cause the emergence of correlations among them." Employing randomly generated data Henry formed two groups of artificial subjects. One group contained subjects with the greatest within-subject variability and the other group the smallest. Henry found that $r^A_H < r^B_H < r^B_L < r^A_L$, where, for example, $r^A_H$ denotes the sample test-retest correlation coefficient between average scores in the high within-subject variability group. He also found $r^A_L < r^B_L < r^B_H < r^A_H$ when the two groups were formed so that one group had the greatest possible among-subject variability and the other the smallest possible. This procedure corresponded to the procedure used by Berger and Sweney.

Henry's results conflicted with theirs, where $r^B_L < r^A_L < r^A_H < r^B_H$, leading one to question the results of Berger and Sweney. In the second section of his paper, Henry employed a normal distribution to set up a statistical model which he claimed to be realistic in relation to actual experimental data. He was able to prove that the "true score", the quantity to be estimated, is more highly correlated with the average score than with
the best score by assuming that an individual's obtained score
normally distributed. This assumption was incorrect, however.
ture scores were normally distributed, but an individual's
not, since they were the sum of the individual's
erm term selected equally likely from the integers zero to 99.99 and
other results led Henry to the same conclusion as
"that the practice of using the best rather than
average individ Parser does not rest on a sound foundation."

In contrast with the aims of the other authors, Hetherington was
mainly interested in obtaining an estimate of an
individual's pt maximum performance. He argued that

because there is no possibility for a person to perform
better than his maximum...if there were no measurement
error, trial-to-trial variation would be either equal
or below this maximum. In this hypothetical situation
where measurement error was nonexistent the best score
would most closely approximate this maximum.

Hetherington then stated the following rule for determining what to use
the average score or the best score when estimating an individual maximum
score.

Hetherington's Rule: "...the experimenter should use average scores when
the major source of unreliability is measurement error, and use...best
scores when within-subject variation is high compared to measurement error."

Hetherington illustrated the above rule with data obtained from a
strength test of 105 young boys. By means of analysis of variance the various
variance components were estimated. Hetherington found that the instrument
error variance was small compared to the within-subject variance, which is
in agreement with the results of experiments by Henry (4,5). On the basis
of the rule just stated, Hetherington concluded that in this case the appropriate estimate of his maximum score.

Hetherington questioned the assumption made by both Henry (6) and Krol (8) that trial-to-trial variation is distributed randomly (with mean zero) about the true score, suggesting that when estimating an individual's maximum score only measurement error would be so distributed. Henry and Kroll, however, were not interested in estimating the maximum score. Hetherington also questioned the practice of selecting a criterion score on the basis of its correlation coefficient and suggested instead that the experimenter should select his criterion score in order to "most closely approximate" the true score, but this phrase was not defined.

Baumgartner (2) presented a brief review and critique of the earlier Research Quarterly articles. This critique, as well as the conclusions of his article, were based on the assumption, also made by Henry and Kroll, that the mean score is the "true score." This assumption is valid, however, only when one is interested in estimating typical or mean performance and is not true when one is estimating maximum performance, as Hetherington was. Baumgartner also administered four multiple trial tests to approximately 260 subjects on each of two days and tested the hypothesis of normality of a subject's data for each test on each day. Based on the individual tests of each subject (tests of low power because the number of trials ranged from only four to twelve) it was not possible to reject this hypothesis. Baumgartner then concluded that (1) "the multiple trial data of a subject taking a physical performance test is usually normally distributed" and that (2) "in almost all situations the use of a criterion score based on the mean
of the trial scores is more defensible than the

However, the second conclusion is based on the f

conclusion should not be accepted simply because

reject it (see Realistic distributions for physi-

cal perfo. 

ment error, below).

A common criticism of all the articles revi

explicit statistical models. By defining a stat

and author will be able to better understand the
dure employed.

Problem

In a typical testing situation there exist

trial-to-trial variation in an individual's scor

measurement error, which can be assumed to occur

mean. The second source is within-subject varia

trials a subject cannot be expected to perform i

one would expect a sprinter to run 100 yards in a

series of track meets. Here we assume that the

a small number n of trials and wishes to estimate

The Concept of Maximum Performance

In many physical fitness tests exercise physi

olists hyz

"maximum score," which the individual is incapable

either achieve or else come arbitrarily close to

tests this limit is determined by the structure a

and prevailingio
state of the performing muscles. The maximum score will vary over time due to changes in the individual's physical condition. Psychological factors have been shown to be a very important factor in tests of muscular performance. Because such factors are readily modified this will obviously make estimation of the physiological limit very difficult in tests of muscular performance.

Wilmore (11) has demonstrated that motivation does not result in a statistically significant change in certain maximal physiological responses — e.g., maximum heart rate, ventilation, oxygen consumption and oxygen pulse. Because these organic function tests seem to be unaffected by psychological factors, they will be more suitable than muscular performance tests for the estimation of maximum performance.

Procedures and Results

The statistical model. The following statistical model for the physical performance of a single subject under constant experimental conditions has been constructed. Let \( X_1, X_2, \ldots, X_n \) denote a random sample of \( n \) trials (usually \( n \leq 10 \)) from the probability distribution \( f_X \), and assume that the \( X \)'s have an upper limit \( \theta \), which it is not possible to exceed.* Suppose also that each \( X_i \) is measured with error by \( k \) different measuring instruments, so that the actual observations are

\[
y_{ij} = X_i + e_{ij}; \quad i = 1, \ldots, n, \quad j = 1, \ldots, k,
\]

*Failure to appreciate this point may have caused some of the controversy. The parameter \( \theta \) is not the population mean value, often called expected value or \( E(X) \). It is the upper limit of the true performance scores.
where the $e$'s are a random sample from the probability distribution of the $e$'s. Also assume that the $X$'s are independent observations. Over short periods of time and under constant experimental conditions $\theta$ can be considered as a variable which assumes a specific value, $\theta(t)$, at any point of time $t$.

Let $Y_i$ and $e_i$ be the respective averages of observed scores and the (unobserved) errors on trial $i$ of $n$ performance trials. Let $Y(n)$, $X(n)$, and $e(n)$ be the maxima of $Y_i$, $X_i$, and $e_i$, respectively, and let $\overline{Y}_n$ be the average over the entire sample of observations.

A counterexample to Hetherington's Rule. Having defined the statistic model it is now possible to present a counterexample such that with probability .95, $X = 50$; with probability .05, $X$ is equal to any value between 50 and 60. This population has a mean $E(X) = 50.25$ and a variance $\sigma^2_x = 1.6$. Let the error $e$ have a normal distribution with zero mean and variance $\sigma^2_e = 4$, and let $Y = X + e$.

According to Hetherington's Rule $\overline{Y}_n$ should be selected as the estimator of $\theta$, which here is 60. But for $n \leq 10$, $E(Y(n)) \leq E(Y(10))$, and $E(Y(10)) < E(X(10)) + E(e(10))$. It is possible to compute the means for this example; we get $52.16 + 3.08 = 55.24$. That is, for $n \leq 10$, in a sample of size $n$ the sampling distribution of the largest observation has a population mean less than 55.24, which is less than the largest observation, and hence further from $\theta$. It is always true that the sample mean cannot exceed the population mean of $\overline{Y}_n$ is less than that of $Y(n)$,
So the largest observation is, for \( n \leq 10 \) and this particular distribution, better than the sample average in the sense of absolute mean deviation. This is true despite the fact that the within-subject variance is small compared to the error variance.

The above example is not as farfetched as it might seem; in strength tests where psychological factors often inhibit a person from displaying his true strength, a long-tailed distribution seems very possible.

Restatement of Hetherington's Rule. Although the preceding example shows that Hetherington's Rule is incorrect, the rule seems to have some merit and probably should be restated in the following more precise form. Because "better" has not been defined, this restatement is not susceptible to proof.

Restatement of Hetherington's Rule: Let \( R \) be the ratio of the standard deviations of the true scores and the errors; \( R = \sigma_x/\sigma_e \). Then if \( Y(n) \) is better than \( \bar{Y}_n \) when \( R = R_0 \), it is better for all \( R \geq R_0 \). If \( R \) is fixed, then if \( \bar{Y}_n \) is better than \( Y(n) \) for some sample size \( n_0 \), it is better for all \( n \geq n_0 \).

The effect of \( n \) is motivated by the following two facts:

a) \( E(Y(n)) \) is an increasing function of \( n \).

b) \( \bar{Y}_n \leq Y(n) \) for all \( n \).

When absolute mean deviation is the criterion for comparison of estimators of \( \theta \), the effect of \( n \) is a direct consequence of the above two facts.

Criteria for selection and evaluation of estimators of \( \theta \). Because \( Y \) is the sum of two random variables, with \( X \) having an upper bound, the assumption of specific realistic distributions for \( X \) and \( e \) usually results in a complicated distribution for \( Y \) which cannot be expressed explicitly. For this reason it is ordinarily very difficult to employ such estimation criteria as uniform
minimum variance unbiasedness, maximum likelihood, sufficiency or minimum mean squared error in selecting an estimator of the parameter \( \theta \). However, we have found it possible to compare the two estimators of specific families of distributions, on the basis of mean squared error (MSE), which may be defined as variance + (bias)\(^2\), and absolute mean deviation, respectively.

It should be noted that the properties of an estimator of \( \theta \) will ordinarily vary when the distributional assumptions are changed. For example, the median is a more efficient estimator of the population mean than is the sample mean under some non-normal distributions.

Comparisons of \( \bar{Y}_n \) and \( Y(n) \) as estimators of maximum performance for specific types of distributions. In this section some specific types of distributions will be assumed for \( X \) and \( e \) and a comparison of estimators of \( \theta \) will be made. The examples were chosen for their tractability rather than their realism.

Suppose that the true scores \( X \) are equally likely to be any point in the interval \((\theta_1, \theta_2)\), the errors \( e \) are equally likely to be any point in the interval \((-a, a)\), and the number of measuring instruments is one (\( k = 1 \)). The distribution of \( Y = X + e \) has a trapezoidal shape and is illustrated in Figure 1 for the case \( \theta_2 - \theta_1 > 2a \).

**Theorem 1:** If \( X_i \) is Uniform \((\theta_1, \theta_2)\), \( e_i \) is an independent error which is Uniform \((-a, a)\), and \( Y_i = X_i + e_i \), \( i = 1, \ldots, n \), then the ratio of mean squared errors of \( Y(n) \) and \( \bar{Y}_n \) is...
Figure 1 Probability distributions for within-subject and observed scores-uniform case.
\[
L_n = \frac{\text{MSE}(\bar{Y}(n))}{\text{MSE}(Y_n)} = \begin{cases} 
\left\{ \frac{3n}{(n+1)[R^2(3n+1)+1]} \right\} & \text{if } R \geq 1 \\
\left\{ \frac{3n}{(n+1)[R^2(3n+1)+1]} \right\} & \text{if } 0 \leq R \leq 1 
\end{cases}
\]

where \( R = \frac{\sigma_x}{\sigma_e} = (\theta_2 - \theta_1)/2a. \)

Theorem 1 shows that \( L_n \) is a function of \( R \) alone, as desired. Thus, the question as to which estimator is the better one depends only on the ratio of the within-subject standard deviation to the standard deviation of the measurement error and not on the individual standard deviations.

Although \( L_n \) is not a strictly decreasing function of \( R \) (see Figure 2, \( n=2 \)), the Restatement of Hetherington's Rule is supported by this example. In most cases we would expect within-subject variance to be larger than error variance; for these cases \((R>1)\) note that \( Y(n) \) is better than \( \bar{Y}_n \) for all sample sizes considered.

A second example was also considered. For this example \( X \) had a negative exponential distribution defined by

\[
f_X(x) = a \exp(a(x - \theta)), \quad -\infty < X < \theta, \quad a > 0, \quad \theta > 0.
\]
Values of the ratio of mean squared errors, $L_n(R)$, when within-subject and error variance are both uniformly distributed, with $R$ the ratio of standard deviations.

$\bar{Y}$ better in this region

$Y_{(n)}$ better in this region

$R = \sigma_x/\sigma_e$
This example also provided support for the Restatement of Hetherington's Rule. Realistic distributions for physical performance and measurement error. The distributions for physical performance assumed in the previous examples have not been very realistic. In order to describe a realistic distribution, data from a large experiment were obtained through the courtesy of Baumgartner (2).

Baumgartner tested two groups of junior high school, high school and college students on four different physical fitness tests: standing long jump, shuttle run, reaction time and speed of arm movement. Although Baumgartner found that it was not possible to reject the hypothesis of normality of a subject's data for each test on each day, this may be because of insufficient sample size (n=4, 6, or 12). For example, Stephens (9) estimated the power of tests for normality, including the Shapiro and Wilks test used by Baumgartner. He estimated that this test had a power of 29%, when n=20, of detecting that the underlying distribution was $\chi^2$ with ten degrees of freedom; this alternative to normality has greater skewness and kurtosis than Baumgartner's data, estimated by using the following procedure. For each subject the data from the first two tests were standardized to have zero mean and unit variance. Within each of the six classes (school x group) histograms were plotted for the two tests based on the standardized data, pooling all subjects in the class. The histograms, based on multiple trials from many subjects, are not approximately normal. See Figure 3 for a histogram of the long jump data for one group of junior high school students. Statistics describing the distributions of the standardized data were also calculated. It was found that the standing long jump data were negatively skewed and had negative kurtosis for all six classes. Because the shuttle run is an example where lower scores are favored, the positive skewness corresponds to the
negative skewness of the standing long jump data. Thus, a realistic distribution for physical performance for tests where higher scores are favored would seem to have both negative skewness and negative kurtosis.

Experiments by Henry (4,5) and Hetherington indicate that ordinarily the within-subject variance is much larger than the error variance. Thus, a realistic distribution for the measurement error would be a symmetric, unimodal distribution with mean zero and variance $\sigma_e^2$, with $\sigma_e^2$ much smaller than $\sigma_X^2$.

A consistent estimator of $\theta$. An estimator $T_n$ is said to be a consistent estimator of a parameter $\theta$ if, given any small interval containing the parameter, the probability that $T_n$ lies in the interval tends to one as $n$ increases without bound. This is obviously a desirable property of an estimator, because it ensures that the estimator will be close to the parameter's true value if a sufficiently large sample is taken.

So far the only estimators of $\theta$ that have been considered are $\bar{Y}_n$ and $Y_{(n)}$. Ordinarily neither of these estimators is unbiased or consistent under the given model. Intuitively, as $n$ gets large, $\bar{Y}_n$ will be closer to the mean value of $Y$, which is not the same as $\theta$, the upper bound of $X$. On the other hand, the largest observation $Y_{(n)}$ will, as $n$ increases, exceed $\theta$ and keep on increasing unless the errors $e$ have an upper bound.

What is needed is a method of allowing for the possibility that $Y_{(n)}$ may exceed $\theta$ because of random error. Unless $k \geq 2$, (more than one measurement per trial) it would be impossible to estimate an adjustment for the size of the error $e$. Theorem 2 below presents a consistent estimator of $\theta$, assuming only that the random errors are bounded by $\pm a$ and that these bounds are the smallest possible. Thus, the estimator could be considered non-parametric,
but note that its bias, variance, and hence its mean-square error will depend on the particular distributions assumed for X and e.

**Theorem 2:** Suppose that the given statistical model holds that k is fixed, k ≥ 2, and that the smallest possible bounds on the error are ±a. Define $Q_{n,k}(Y)$ by

$$Q_{n,k}(Y) = \max_{1 \leq i \leq n} \max_{1 \leq j \leq k} y_{ij} - 1/2 \max_{1 \leq i \leq n} \text{range}(y_{i1}, y_{i2}, \ldots, y_{ik}).$$

Then $Q_{n,k}(Y)$ is a consistent estimator of $\theta$.

The idea behind the estimator is as follows. Take $k$ measurements on each of $n$ trials. The largest of these $n \cdot k$ observations will estimate $\theta + a$ (from below; it cannot exceed $\theta + a$). Find the trial with the largest range among its $k$ measurements. Half of this range will estimate $a$ (from below). The difference of these two estimators will estimate $\theta$.

**Examination of the consistent estimator.** The consistent estimator was compared with $\overline{Y}_n$ and $Y(n)$ by means of a simulation study. We took three cases, each with $k = 2$ measurements per trial.

I) $X$ Uniform on $\theta_1$ to $\theta_2$, $e$ Uniform on $-a$ to $a$;

II) $X$ Beta $(3,2)$, $e$ Uniform on $-a$ to $a$;

III) $X$ Beta $(3,2)$, $e = e_1 + e_2 + e_3$, each $e_i$ Uniform on $-a$ to $a$.

Case I was used previously to compare $\overline{Y}_n$ and $Y(n)$, with $k = 1$. Cases II and III use a reasonably realistic distribution for within-subject variation, the Beta distribution with $p = 3$, $q = 2$ (compare Figures 3 and 4). The uniformly distributed errors in Case II might be appropriate when rounding is the main source of error. The error in Case III reasonably approximates a normal distribution truncated at three standard deviations.

In Case I $\overline{Y}_n$ has mean-squared error (MSE) given by $(a^2/3n)((3n + 1)\sigma^2 + 1/2)$. 


Figure 3. Standing long jump data, from six standardized trial scores on 48 subjects.
Figure 4. The Beta (3,2) distribution.
In the other two cases the MSE of $\overline{Y}_n$ can be shown to be $(1 + 1.5n + 0.16) \sigma_x^2 / \sigma_e^2$, where $R = \frac{\sigma_x^2}{\sigma_e^2}$, the ratio of within-subject and standard deviation. In no case was the MSE of $\overline{Y}_n$ competitive with that of either the first observation $Y_{(n)}$ or the consistent estimator $\theta$. The consistent estimator was better than $Y_{(n)}$ only for Case I with R=1 and n=20, although Table 5 suggests, Q was not too bad in Case III when R was large.

The results suggest that, for the models and the correction for the possible upward bias caused by the errors too large, A consistent estimator that outperforms the largest observation, $Y_{(n)}$ in the small-sample situations typical of maximum physical performance trials remains to be discovered.
Figure 5 - Ratio of \( \frac{\text{MSE}(Y_n)}{\text{MSE}(Q_{n,2})} \) for different values of \( R \).

- \( R = 8.0 \)
- \( R = 1.0 \)
Summary

Previous research into the statistical estimation of maximum physical performance has suffered from a lack of precisely defined models and estimation criteria. This has inhibited the drawing of firm conclusions.

The results of this paper support the idea that the best observation is to be preferred over the average as an estimator of maximum performance unless the ratio of within-subject to error variance is small.

Examination of standardized maximum performance data pooled over many subjects suggests the approximate form for within-subject variation in a performance trial where high scores are favored will have slightly negative skewness and kurtosis.

Since neither the best nor the average scores are consistent estimators of maximum performance in the presence of experimental error, a consistent estimator was developed for the case of more than one measurement per trial. A simulation study involving reasonably realistic models showed that the best score outperformed the consistent estimator for the cases considered, suggesting the need for a consistent estimator with better small-sample properties.
References


Acknowledgements

The problem considered here was suggested by Drs. Ronald Byrd and Douglas Smith; Drs. Gerald van Belle and Naftali Langberg made several helpful suggestions.