SPORTSCIENCE sportsci.org
Perspectives / Performance

How to Interpret Changes in an Athletic Performance Test

Will G Hopkins

Sportscience 8, 1-7, 2004 (sportsci.org/jour/04/wghtests.htm)
Sport and Recreation, Auckland University of Technology, Auckland 1020, New Zealand. Email.
Reviewers: Christopher J Gore, Australian Institute of Sport, PO Box 219, Brooklyn Park, South Australia 5032; David B Pyne, Australian Institute of Sport, PO Box 176, Belconnen, ACT 2616, Australia

When monitoring progression of an athlete with performance or other fitness tests, it is important to take into account the magnitude of the smallest worthwhile enhancement in performance and the uncertainty or noise in the test result. For elite athletes competing in sports as individuals, the smallest worthwhile enhancement would give the athlete an extra medal per 10 competitions; the required change in performance is 0.3 of the typical variation in an athlete's performance from competition to competition, or ~0.3-1% when expressed as a change in power output, depending on the sport. In team sports, where there is no direct relationship between team and test performance, an appropriate default for the smallest change in test performance is one-fifth of the between-athlete standard deviation (a standardized or Cohen effect size of 0.20). Noise in a test result is best expressed as the typical or standard error of measurement derived from a reliability study. The noise in most performance tests is greater than the smallest worthwhile difference, so assessments of changes in performance can be problematic. An exact but somewhat impractical solution is to present chances that the true change is beneficial, trivial, and harmful. A simpler approach is to apply systematic rules to decide whether the true change is beneficial, trivial, harmful, or unclear. Unrealistically large changes can also be partially discounted when tests are noisy.

KEYWORDS: Bayes, correlation, error of the estimate, error of measurement, limits of agreement, reliability, time to exhaustion, time trial, validity.

Reprint pdf · Reprint doc · Slideshow · Commentaries by Gore and Pyne

Updated Sept 2011: the smallest worthwhile change is now stated as 0.3 of the variation in an athlete's performance, not 0.5 as previously.

The basis for this article is an updated version of a slideshow accompanying a talk entitled "making sense of performance tests", which I presented earlier this year at the Scottish Institute of Sport and more recently at a local conference. The talk was based mainly on previous research by my colleagues and me, along with some new and previously unpublished insights. The title now better reflects the emphasis on monitoring an athlete's performance from test to test.

Monitoring the progression of athletes with regular performance and other fitness-related tests is a widespread and apparently useful practice in upper competitive levels of most if not all sports in wealthy countries, but in my experience lack of understanding about the interpretation of changes in test scores is also widespread. Perhaps the most important issue is that of magnitude: to interpret the change in an athlete's performance since a previous test, you need some concept of the magnitude of change that matters to the athlete in his or her sport. The first section of the talk is therefore concerned with identifying the smallest worthwhile change in performance. Your ability to track such changes with a performance test depends on the validity and reliability of the test, which I explain in the second section. The final section is devoted to several ways of interpreting

How to Interpret Changes in an Athletic Performance Test

Will G Hopkins
Sports and Recreation
Auckland University of Technology

- What's a Worthwhile Performance Enhancement?
  - Solo sports; test performance vs competition time trial
  - Team sports and fitness tests
- What's a Good Test for Assessing an Athlete?
  - Validity; reliability; "signal" vs "noise"
- How Do You Interpret Changes for the Coach and Athlete?
  - Chances; likely limits; simple rules; Bayes

What's a Worthwhile Enhancement for a Solo Athlete?

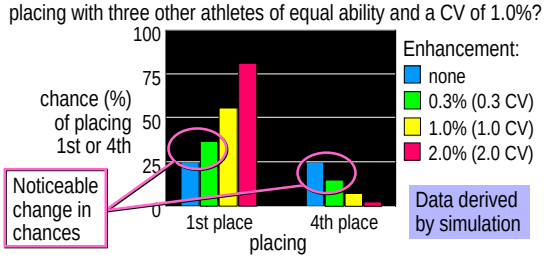
- You need the smallest worthwhile enhancement in this situation of evenly-matched elite athletes:

- If the race is run again, each athlete has a good chance of winning, because of race-to-race variability:

- You need an enhancement that overcomes the variability to give your athlete a bigger chance of a medal.

- Therefore you need a measure of the variability.
  - Best expressed as a coefficient of variation (CV).
    e.g. CV = 1% means an athlete varies from race to race typically by 1 m per 100 m, 0.1 s per 10 s, 1 sec per 100 sec...

- Now, what's the effect of performance enhancement on an athlete's placing with three other athletes of equal ability and a CV of 1.0%?



- 0.3 of a CV gives a top athlete one extra medal every 10 races.
  - This is the smallest important change in performance to aim for in research on, or intended for, elite athletes.
  - 0.9, 1.6, 2.5, 4.0 of a CV gives an extra 3, 5, 7, 9 medals per 10 races (thresholds for moderate, large, very large, extremely large effects).
- References: Hopkins et al. MSSE 31, 472-485, 1999 and MSSE 41, 3-12, 2009.

- An athlete who is usually further back in the field needs more than 0.3 of a CV to increase chances of a medal:

- For such athletes, work out the enhancement that matters on a case-by-case basis. Examples:
  - Need ~4% to equal best competitors in next event.
  - Need ~2% per year for 3 years to qualify for Olympics.
  - Or use the standardized (Cohen) effect size. See later.

- What's the value of the CV for elite athletes?
  - We want to be confident about measuring 0.3 of this value when we test an elite athlete or study factors affecting performance with sub-elite athletes.
- Values of the CV from published and unpublished studies of series of competitions:
  - running and hurdling events up to 1500 m: 0.8%
  - runs up to 10 km and steeplechase: 1.1%
  - cross country: 1.5% (subelite)
  - half marathons: 2.5% (subelite)
  - marathons: 3.0% (subelite)
  - high jump: 1.7%
  - pole vault, long jump: 2.3%
  - discus, javelin, shot put: 2.5%
  - mountain biking, 2.4%
  - swimming: 0.8%
  - cycling 1-40 km: 1.3%

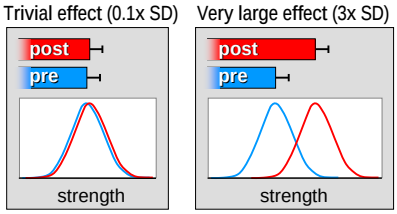
  } CV for time. Multiply by ~2-3 to get CV for mean power.

- Beware: changes in performance in lab tests are often in different units from those for changes in competitive performance.
  - Example: a 1% change in endurance power output measured on an ergometer is equivalent to the following changes…
    - 1% in running time-trial speed or time;
    - ~0.4% in road-cycling time-trial time;
    - 0.3% in rowing and swimming time-trial time.
- Beware: change in performance in some lab tests needs converting into equivalent change in power output in a time trial.
  - Example: 1% change in power output in a time trial is equivalent to:
    - ~15% change in time to exhaustion in a constant-power test
    - ~2% change in time to exhaustion in an incremental test starting at 50% of peak power.
    - ?% change in performance following a fatiguing pre-load.
- So always think about and use percent change in power output for the smallest worthwhile change in performance.
- Reference: Hopkins et al. Sports Medicine 31, 211-234, 2001.

What's a Worthwhile Enhancement for a Team Athlete?

- We assess team athletes with fitness tests, but…
- There is no clear relationship between fitness-test performance and team performance, so…
- Problem: how can we decide on the smallest worthwhile change or difference in fitness-test performance?
- Solution: use the standardized change or difference.
  - Also known as Cohen's effect size or Cohen's *d* statistic.
  - Useful in meta-analysis to assess magnitude of differences or changes in the mean in different studies.
  - You express the difference or change in the mean as a fraction of the between-subject standard deviation (Δmean/SD).
  - It's like a z score or a t statistic.
  - The smallest worthwhile difference or change is ~0.20.
  - 0.20 is equivalent to moving from the 50th to the 58th percentile.

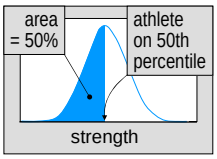
- Example: the effect of a treatment on strength

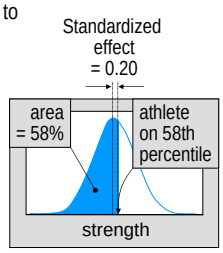


- Interpretation of standardized difference or change in means:

	Cohen	Hopkins
trivial	<0.2	<0.2
small	0.2-0.5	0.2-0.6
moderate	0.5-0.8	0.6-1.2
large	>0.8	1.2-2.0
very large	?	2.0-4.0
extremely large	?	>4.0

- Relationship of standardized effect to difference or change in percentile:





Standardized effect	Percentile change
0.20	50 → 58
0.20	80 → 85
0.20	95 → 97
0.25	50 → 60
1.00	50 → 84
2.00	50 → 98

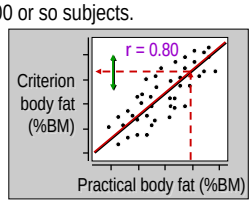
- Can't define smallest effect for percentiles, because it depends what percentile you are on.
- But it's a good practical measure.
- And easy to generate with Excel, if the data are approx. normal.

What's a Good Test for Assessing an Athlete?

- Needs to be valid and reliable.
- Validity of a (practical) measure is some measure of its one-off association with other (criterion) measures.
  - "How well does the practical measure measure what it's supposed to measure?"
  - Important for distinguishing between athletes.
- Reliability of a measure is some measure of its association with itself in repeated trials.
  - "How reproducible is the practical measure?"
  - Important for tracking changes within athletes.

Validity

- We usually assume a sport-specific test is valid in itself…
  - …especially when there is no obvious criterion measure.
  - Examples: tests of agility, repeated sprints, flexibility.
  - Researchers usually devise such tests from an analysis of competitions or games.
- If relationship with a criterion *is* an issue, usual approach is to assay practical and criterion measures in 100 or so subjects.
  - Fitting a line or curve provides a calibration equation, the error of the estimate, and a correlation coefficient.
  - Preferable to Bland-Altman analysis of difference vs mean scores.
    - B-A analysis can indicate a systematic offset error (bias) when there is none.



- Beware of units of measurement that lead to spurious high correlations.
  - Example: a practical measure of body fat in kg might have a high correlation with the criterion, but…
  - Express fat as % of body mass and correlation = 0!
  - So the measure provides no useful information.
- For many measures, use log transformation to get uniformity of error of estimate over the range of subjects.
  - Check for non-uniformity in a plot of residuals vs predicteds.
  - Use the appropriate back-transformation to express the error as a coefficient of variation (percent of predicted value).
- The error of the estimate is the "noise" in the prediction.
- The smallest worthwhile difference between athletes is the "signal".
- Ideally, noise < signal (more on this shortly).
- If signal = Cohen's 0.20, we can work out the validity correlation…
  - r^2 = "variance explained" = $(SD^2\text{-error}^2)/SD^2$.
  - But want noise < signal; that is, error < 0.20*SD.
  - So ideally r > 0.98! Much higher than people realize.

- Some researchers dispute the validity of constant-power and incremental time-to-exhaustion tests of endurance for athletes.
  - They argue that such tests don't simulate the pacing of endurance races, whereas constant-work or constant-duration time trials do.
  - True, if you want to study pacing.
  - But if you want to study power output, pacing can only add noise.
  - Besides, peak power in incremental tests and time to exhaustion in constant-power tests have strong relationships with the criterion measure of race performance.
  - But a definitive longitudinal validity study and/or comparison of reliability for time to exhaustion vs time trials is needed.
- Longitudinal validity
  - How well does the practical measure track changes in the criterion?
    - Example: skinfolds may be mediocre for differences between individuals but good for changes within an individual.
  - There are few such studies in the literature.

Reliability

- Reliability is reproducibility of a measurement if or when you repeat the measurement.
  - It's the same sort of thing as reproducibility in an athlete's performance between competitions.
  - For performance tests, it's usually more important than validity.
- It's crucial for practitioners…
  - because you need good reproducibility to monitor small but practically important changes in an individual athlete.
- It's crucial for researchers…
  - because you need good reproducibility to quantify such changes in controlled trials with samples of reasonable size.

- How do we quantify reliability?
  Easy to understand for one subject tested many times:

Subject	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Mean ± SD
Chris	72	76	74	79	79	77	76.2 ± 2.8

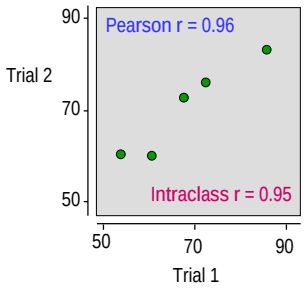
  - The 2.8 is the standard error of measurement.
  - I call it the typical error, because it's the typical difference between the subject's true value and the observed values.
  - It's the random error or "noise" in our assessment of clients and in our experimental studies.
  - Strictly, this standard deviation of a subject's values is the total error of measurement rather than the standard or typical error.
    - It's inflated by any "systematic" changes, for example a learning effect between Trial 1 and Trial 2.
    - Avoid this way of calculating the typical error.

- We usually measure reliability with many subjects tested a few times:

Subject	Trial 1	Trial 2	Trial 2-1
Chris	72	76	4
Jo	53	58	5
Kelly	60	60	0
Pat	84	82	-2
Sam	67	73	6
	Mean ± SD:		2.6 ± 3.4

  - The 3.4 divided by √2 is the typical error.
  - The 3.4 multiplied by ±1.96 are the limits of agreement.
  - The 2.6 is the change in the mean.
  - This way of calculating the typical error keeps it separate from the change in the mean between trials.

- And we can define retest correlations:
  Pearson (for two trials) and intraclass (two or more trials).



- The typical error is much more useful than the correlation coefficient for assessing changes in an athlete.

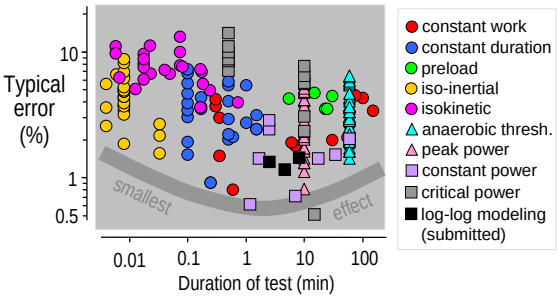
- Noise (typical error) vs signal with change scores
  - Think about ± the typical error as the noise or uncertainty in the change you have just measured.
  - You want to be confident about measuring the signal (smallest worthwhile change), say 0.5%.
  - Example: you observe a change of 1%, and the typical error is 2%.
    - So your uncertainty in the change is 1 ± 2%, or -1% to 3%.
    - So the change could be harmful through quite beneficial.
    - So you can't be confident about the observed beneficial change.
    - But if you observe a change of 1%, and the typical error is only 0.5%, your uncertainty in the change is 1 ± 0.5%, or 0.5% to 1.5%.
    - So you can be reasonably confident you have a small but worthwhile change.
  - Conclusion: ideally, you want typical error < smallest change.
    - If typical error > smallest change, try to find a **better test**.
    - Or **repeat the test** with the athlete several times and **average the scores** to reduce the noise. (Four tests halves the noise.)

- More on noise...
  - When testing individuals, you need to know the noise of the test determined in a reliability study with a time between trials short enough for the subjects not to have changed substantially.
    - Exception: to assess change due **specifically** to, say, a 4-week intervention, use 4-week noise.
  - For estimating sample sizes for research, you need to know the noise of the test with the same time between trials as in your intended study.
    - Beware: noise may be higher in the study (and therefore sample size will need to be larger) because of **individual responses** to the intervention.
    - (Individual responses can be estimated from the difference in noise between the intervention and control groups.)
    - Beware: noise **between base and competition phases** can be much greater than noise **within a phase**, because some athletes improve more than others between phases.

- Even more on noise.
  - If published reliability studies aren't relevant, measure the noise yourself with the kind of athletes you deal with.
  - As with validity, use log transformation to get uniformity of error over the range of subjects for some measures.
    - Check for **non-uniformity** in a plot of residuals vs predicteds or **change scores vs means**.
    - Use the appropriate back-transformation to express the error as a **coefficient of variation** (percent of subject's mean value).
  - Ideally, noise < signal, and if signal = Cohen's 0.20, we can work out the reliability correlation:
    - Intraclass r = (SD^2-$error^2$)/SD^2.
    - But want noise < signal; that is, error < 0.20*SD.
    - So **ideally r > 0.96**! Again, much higher than people realize.

- How bad is the noise in performance tests?
  - Quite bad! Many have a lot more noise than the smallest important change for competitive athletes.
  - So, when monitoring an individual athlete, you won't be able to make a firm conclusion about a small or trivial change.
  - And when doing research, you will need possibly 100s of athletes to get acceptable accuracy for an estimate of a small or trivial change or "effect".
    - "No effect" or "a small effect" is not the right conclusion in a study of 10 athletes with a noisy performance measure.
    - Authors should publish **confidence limits** of the true effect, to avoid confusion.
  - A few performance tests have noise approaching the smallest worthwhile change in performance.
    - Use these tests!

Typical error of mean power in various types of performance test:



- Best explosive tests are iso-inertial (jumping, throwing).
- Best sprint tests are constant work or constant duration.
- Best endurance tests are constant power or peak incremental power.

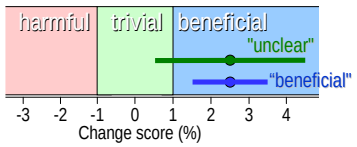
- General reference for this section:
  - Hopkins WG (1997-2004). A New View of Statistics, newstats.org
- Validity:
  - Paton CD, Hopkins WG (2001). Tests of cycling performance. Sports Medicine 31, 489-496
- Reliability:
  - Hopkins WG (2000). Measures of reliability in sports medicine and science. Sports Medicine 30, 1-15
  - Hopkins WG, Schabort EJ, Hawley JA (2001). Reliability of power in physical performance tests. Sports Medicine 31, 211-234

How Do You Interpret Changes for the Coach and Athlete?

- I will deal only with change since a previous test.
  - Hard to be quantitative with trends in multiple tests.
- You have to make a call about magnitude of the change, taking into account the noise in the test. Do it in several ways...

1. Use the chances that true value is greater or less than the smallest important change.
   - Example: the athlete has changed by +1.5% since last test;
     - noise (typical error) is 1.0%;
     - smallest important change is 0.5%;
     - so chances are 76% for a beneficial change, 16% for a trivial change, and 8% for a harmful change.
   - This method is exact, but…
   - It's impractical: you need a spreadsheet for the chances.
     - Get it from newstats.org (spreadsheet for assessing an individual).

2. Use likely limits for the true value (my favorite option).
   - Easiest likely limits are the observed change ± the typical error.
   - State that the true change could be between these limits.
     - "Could" means "50% likely" or "odds of 1:1" or "possible".
     - Interpret the limits as beneficial, trivial, harmful.
     - Call the effect clear only if both limits are the same.
     - The spreadsheet shows clear calls will be right >76% of the time.
   - Example: the athlete has changed by +2.5% since the last test, smallest worthwhile change is 1.0%…
     - If the typical error is ±2.0%, the true change is unclear.
     - If the typical error is ±1.0%, the true change is beneficial.



3. Use these simple rules:
   - If the test is good (noise ≤ smallest signal), believe or interpret all changes as clearly helpful, harmful, or trivial.
     - You will be right >50% of the time (usually much more).
   - If the test is poor (noise > smallest signal), believe or interpret changes only when they are greater than the noise.
     - That is, any change > noise is beneficial (or harmful); any change < noise is unclear.
     - Calls of benefit and harm will be right >50% of the time.
   - Example: typical error (noise) is 2.0%, smallest change is 1.0%, so…
     - This is a poor-ish test, so…
     - If you observe a change of 2.5%, call it beneficial.
     - If you observe a change of 1.5%, call it unclear.
     - If you observe a change of -3.0%, call it harmful.
   - More on making correct calls…

- You can be more conservative with your assessments by changing the rules. For example…
- Believe/interpret changes only when they are greater than 2x noise.
  - Calls of benefit and harm will be right >76% of the time.
  - But for most performance, noise>smallest worthwhile change, so all trivial changes and many important changes will be "unclear".
  - So this rule is too conservative and impractical for athlete testing.
- Using limits of agreement amounts to believing changes only when they are greater than 2.8x the noise.
  - Error rates are even lower, but even more calls are "unclear".
  - Limits of agreement are therefore even more impractical.

4. Blame noise for an extreme test result.
   - An example of Bayesian thinking: you combine your belief with data.
   - Example: the athlete has changed by 5.7% since the last test.
     - But you believe that changes of more than 3-4% are unrealistic, given the athlete and the training program.
     - And you know it's a noisy test, e.g., typical error = 3.0%...
     - So you can partially discount the change and say it is "probably more like 3-4%".
     - (The spreadsheet for assessing an individual can be used to show that chance of changes <3.5% is 30%, or "possible".)
   - We could be more quantitative, and we could apply this approach to all test results, if only we knew how to quantify our beliefs.

- General reference for this section:
  Hopkins WG (1997-2004). A New View of Statistics, newstats.org

Summary

- Find out the smallest worthwhile change or difference in the test.
  - Performance tests with solo athletes: 0.3 of the event-to-event variation in a top athlete's competitive performance.
  - Fitness tests with team sports: ~0.20 of the between-athlete SD.
- Measure such changes in your athletes with a well-designed or well-chosen low-noise test that is specific to the sport.
  - Read up or measure the noise in the test for athletes similar to yours.
  - Improve the test or reduce the noise by doing multiple trials.
- Be up front about the noise when you feed the results of the test back to the athlete.
  - Use chances, likely limits, or rules.
  - Discount unlikely extreme changes with noisy tests.
- Stay on the lookout for less noisy tests.

This presentation is available from:

SPORTSCIENCE sportsci.org

A Peer-Reviewed Site for Sport Research

See Sportscience 8, 2004

the test results for the athlete or coach. See also commentaries by Christopher Gore and David Pyne, to whom I am indebted for valuable interactions and feedback on this topic.

The reprint pdf version of this article contains printer-friendly images of the PowerPoint slideshow and references. View the slideshow to see each slide build sequentially.

Published Nov 2004