Research article

The 8th Australasian Conference on Mathematics and Computers in Sport, 3-5 July 2006, Queensland, Australia

PREDICTION VERSUS REALITY: THE USE OF MATHEMATICAL MODELS TO PREDICT ELITE PERFORMANCE IN SWIMMING AND ATHLETICS AT THE OLYMPIC GAMES

Timothy Heazlewood

School of Exercise Science, Australian Catholic University, Strathfield, NSW, Australia

ABSTRACT

A number of studies have attempted to predict future Olympic performances in athletics and swimming based on trends displayed in previous Olympic Games. Some have utilised linear models to plot and predict change, whereas others have utilised multiple curve estimation methods based on inverse, sigmoidal, quadratic, cubic, compound, logistic, growth and exponential functions. The non linear models displayed closer fits to the actual data and were used to predict performance changes 10's, 100's and 1000's of years into the future. Some models predicted that in some events male and female times and distances would crossover and females would eventually display superior performance to males. Predictions using mathematical models based on pre-1996 athletics and pre-1998 swimming performances were evaluated based on how closely they predicted sprints and jumps, and freestyle swimming performances for both male and females at the 2000 and 2004 Olympic Games. The analyses revealed predictions were closer for the shorter swimming events where men's 50m and women's 50m and 100m actual times were almost identical to predicted times. For both men and women, as the swim distances increased the accuracy of the predictive model decreased, where predicted times were 4.5-7% faster than actual times achieved. The real trends in some events currently displaying performance declines were not foreseen by the mathematical models, which predicted consistent improvements across all athletic and swimming events selected for in this study.

KEY WORDS: Swimming, athletics, olympic games, mathematical functions, extrapolation.

INTRODUCTION

The prediction of future athletic performance by humans is a recurring theme during the Olympiad year, as well as forming the basis for some stimulating 'crystal ball gazing' in some of the learned sports science journals and in the mass media. Mathematics and science are based on the principles of description and more importantly prediction. The ability to make substantive and accurate predictions of future elite level sports performance indicates that such approaches reflect "good" science. Often these predictions are purely speculative and are not based upon any substantial

evidence, rather they are based on the belief that records are made to be broken and that performances must continue to improve over time. The accessibility of data in the form of results from Olympic Games, world records and world best performances in a specific year allows the analysis of performances in any number of events. From these analyses, changes in performance over time can be observed and predictions of future performance can be made utilising the process of mathematical extrapolation.

A number of researchers have attempted to predict future performances by deriving and applying a number of mathematical statistical models based on past performances in athletics. Prendergast (1990) applied the average speeds of world record times to determine a mathematical model for world records. The records or data used in the analysis spanned a 10 year period. Following his analysis, Prendergast (1990) raised the question of whether any further improvements can be expected or if the limits of human performance have been reached. The sports of athletics (Heazlewood and Lackey, 1996) and swimming (Lackey and Heazlewood, 1998) have been addressed in this manner and the knowledge of future levels of sporting performance has been identified by Banister and Calvert (1980) as beneficial in the areas of talent identification, both long and short term goal setting, and training program development. In addition, expected levels of future performance are often used in the selection of national representative teams where performance criteria are explicitly stated in terms of times and distances (Athletics Australia, 2004).

Some researchers such as Péronnet and Thibault (1989) postulate that some performances such as human male 100m sprinting is limited to the low 9 seconds, whereas Seiler (referred to by Hopkins, 2000) envisages no limits on improvements based on data reflecting progression of records over the last 50 years. According to Seiler improvements per decade have been approximately 1% for sprinting, 1.5% for distance running, 2-3% for jumping, 5% for pole vault, 5% for swimming and 10% for skiing for male athletes, whereas female sprint times may have already peaked. The differences for males and females it is thought to reflect the impact of successful drugs in sport testing on females.

The predictions of Heazlewood and Lackey (1996) paradoxically predicted the men's 100m to improve to zero by year 5038 and the women's 100m to reach zero by year 2429, which indicates a more rapid improvement over time for women sprinters. In their model (Heazlewood and Lackey,

1996), the women's times would be faster than men by 2060 where it was predicted the finalist at the Olympic Games would average 9.58s for men and 9.57s for women respectively. A similar crossover effect, where predicted female performances would exceed male performances, was noted for the 400m and high jump. The crossover effect was based on trends in athletic performances obtained prior to 1996; where in some events female improvements were more rapid than males.

In the sport of swimming (Lackey and Heazlewood, 1998), a similar crossover effect was observed for the 50m freestyle where predicted zero time was the year 2994 for men and 2700 for women. The concept that athletes will complete 100m sprints on land and 50m sprints in water in zero seconds appears unrealistic, however mathematical model based on actual data do derive these interesting predictions.

The curves that fit the data have also displayed interesting findings as no one curve fits all the data sets. Different events displayed different curves or mathematical functions (Lackey and Heazlewood, 1998) of best fit. In swimming the men's 50m freestyle was inverse, 100m freestyle compound, 200m sigmoidal, and the 400m and 1500m freestyle cubic. For the women's freestyle events the 50m was inverse, 100m cubic, 200m sigmoidal, 400m cubic and 800m sigmoidal.

In athletics for the men's events the mathematical functions (Heazlewood and Lackey, 1996) were 100m inverse, 400m sigmoidal, long jump cubic and the high jump displayed four functions (compound, logistic, exponential and growth). In the women's events the mathematical functions were 100m cubic, 400m sigmoidal, long jump inverse and high jump displayed four functions (compound, logistic, exponential and growth). This may indicate that different events are dependent upon different factors that are being trained differently or factors underpinning performance evolving in slightly different ways. This has resulted in different curves or mathematical functions that reflect these improvements in training or phylogenetic changes over time.

However, at some point in time how accurately the predictive models reflect reality can be assessed. Since the models of Heazlewood and Lackey (1996) for athletics and Lackey and Heazlewood (1998) for swimming were derived, the 2000 and 2004 Olympic Games have occurred. Hindsight or real data can now enable the assessment of these models over a short timeframe, that is, 8-10 years. Assessing the accuracy of the models predicting performances hundreds or

thousands of years into the future will be based on the research interests of future mathematicians, sports scientist and computer scientists.

The current research problem is how well the actual times and distances achieved by athletes at the 2000 and 2004 Olympic Games fit the predicted model for athletics and swimming based on the Heazlewood and Lackey (1996) and Lackey and Heazlewood (1998) prediction equations?

METHODS

The previous models were based on following model fit criteria used by Heazlewood and Lackey (1996) for athletics and Lackey and Heazlewood (1998) for swimming. The average time and distances for the finalist in each event were utilised to generate the data for the statistical analysis for curve estimation. Potentially both linear and non linear functions can be derived. The mean score of the actual performances from the 2000 and 2004 Olympic Games finalists were compared with the predicted values in the athletic events selected in this study (Wikipedia, 2006). Times for the 100m and 400m were in seconds and distances for the long and high jump were in metres.

The results for the finalists in current Olympic freestyle swimming events (50m, 100m, 200m, 400m, 800m women and 1500m men) were collected from internet based results (Wikipedia, 2006). Times were recorded to one hundredth of a second which is the recording method used by Federation Internationale de Natation Amateur (FINA, 1997). These times were then converted from a minutes and seconds format to a seconds only format to facilitate calculations when applying the regression methods. The mean of the finalists in each event for each year in the study was then calculated. The mean was used as it is a measure that is representative of all scores in each group (Rothstein, 1985). The use of the mean of the finalists in this study may be more representative of the changes in human performance that world records as used by Jokl and Jokl (1976a; 1976b; 1977) and Edwards and Hopkins (1979). A world record holder's performance may be far in advance of that of any other competitor and not be representative of overall performance in an event. For example, the women's 400m freestyle world record as set by Tracey Wickham in 1978 was not bettered until 1988 at the Seoul Olympic Games (Wallechinsky, 1996).

In the swimming pool the factor of wind resistance is not considered significant and as such wind readings are not required for swim records. In athletics assistive and resistive winds are thought to

influence performance in events such as the 100m and long jump and the wind variable can be corrected to assess performance in still air conditions. The wind correction calculations are not presented in this paper just the times and distances reported for the athletic events, however correcting for the influence of wind may result in slightly different values for the original data.

The means were then included as a data set for each event for each Olympic Games for analysis using the Statistical Package for the Social Sciences (SPSS) program version 6.1 (Norušis, 1993) to derive a number of possible regression equations. A number of criteria were used to evaluate the goodness of fit of each derived function for each individual event.

General method of determining the appropriate regressions models

To investigate the hypotheses of model fit and prediction, the eleven regression models were individually applied to each of the athletic and swimming events. The regression equation that produced the best fit for each event, that is, produced the highest coefficient of determination (abbreviated as R^2), was then determined from these eleven equations. The specific criteria to select the regression equation of best were the magnitude of R^2, the significance of the analysis of variance alpha or p-value and the residuals.

The coefficient of determination

The coefficient of determination (R^2) is a measure of accuracy of the model used. A coefficient of determination of 1.00 indicates a perfectly fitting model where the predicted values match the actual values for each independent variable (Norušis, 1993). Where more than one model was able to be selected due to an equal R^2, the simplest model was used under the principle of parsimony, that is, the avoidance of waste and following the simplest explanatory model.

Residuals

The residuals are the difference between the actual value and the predicted value for each case, using the regression equation (Norušis, 1993) and the smaller the residual, the better the fit of the model. For each model the residuals were generated by the SPSS program. A large number of positive residuals indicate that the prediction is an over estimation (faster than the actual performance) and a large number of negative residuals indicates an underestimation (slower time than the actual performance).

Table 1. Athletic events, type of mathematical function, equations and R^2 values.

Event	Regression Type	Equation **	R^2*
Men's 100m	Inverse	$Y = b_0 + (b_1/t)$	.659
Women's 100m	Cubic	$Y = b_0 + b_3t^3$	.902
Men's 400m	S	$Y = e^{(bo + b1/t)}$	.907
Women's 400m	S	$Y = e^{(bo + b1/t)}$	.843
Men's Long Jump	Cubic	$Y = b_0 + b_3t^3$	.777
Women's Long Jump	Inverse	$Y = b_0 + (b_1/t)$	.894
Men's High Jump	Compound	$Y = b_0(b_1)^t$	.944
	Logistic		.944
	Exponential	$Y = b_0e^{b1t}$	.944
	Growth	$Y = e^{bob1t}$	.944
Women's High Jump	Compound	$Y = b_0(b_1)^t$	.947
	Logistic		.947
	Exponential	$Y = b_0e^{b1t}$	.947
	Growth	$Y = e^{bob1t}$	.947

* All R^2 values in table 1 significant at p<0.05.

** Where b_0 = a constant; b_1, b_3 = regression coefficients; t = year, y = mean result for each event.

Level of significance
The level of significance, or p value, is a representation of the relationship between the model and the data. The smaller the p value, the higher the level of significance and the greater the relationship where a small p value indicates a small possibility that the closeness of the predicted values to the actual values due to chance is small.

Logical acceptance based on extrapolations
The ability of the model to generate extrapolations that appear to be reasonable when compared to previous means was also taken into consideration. When a model generated extrapolations that appear to be inconsistent with the actual results this model was discarded and the model with the next highest coefficient of determination was selected.

Applying the model of best fit
After selection of the model to be used, according to the criteria previously stated, the equation of best fit was determined by applying the derived constants and coefficients to the generic formula for that model. Using this equation, a prediction of the mean result for the event at each Olympiad was calculated. At this stage, graphs representing the means of past and future performances for each event in each Olympiad were also generated in addition to predicted means using the appropriate regression equation.

Final predictions for the year 2000 and 2004
To predict the level of performance in the year 2000 and 2004, the data set that provided the greatest accuracy was chosen and the data from 1996 re-included in the data set, where appropriate. A series of regressions were made using the best fitting model and data set for each event. Using the constants and coefficients generated by regression models the future predictions were then calculated.

It is important to note that in some events the average for a complete field of competitors was not always possible due to disqualifications or injury. In the case of injury a competitor did not finish the event. This situation only occurred in a few events. At this point in time no attempt was made to re-evaluate the 1996 and 1998 models based on inclusion of 2000 and 2004 data.

RESULTS

The data for the predicted values and the actual values are provided in the table for each event. Table 1 indicates the events, mathematical functions, equations and R^2 values derived from the Heazlewood and Lackey (1996) for athletic events of the men's and women's 100m, 400m, long jump and high jump.

The trends in the mathematical functions indicate the men's and women's 400m and the high jump show identical trends in changes in performance over time, where the 400m was sigmoidal and the high jump was compound, logistic, exponential and growth. In the majority of events the explained variance or R^2 values were statistically very significant (p < 0.01). It is interesting to note the R^2 values were highest for the men's and women's high jump (0.94) and lowest for the men's 100m (0.66) and men's long jump (0.78).

Table 2 indicates the predicted performances and the actual performances achieved from the 2000

Table 2. The predicted and actual performances for men's and women's 100m, 400m, long jump and high jump for 2000 and 20004 Olympics.

Year Predict	100M	100W	400M	400W	MLJ	WLJ	MHJ	WHJ
2000	10.07s	10.84s	43.84s	48.29s	8.36m	7.27m	2.42m	2.05m
2004	10.03s	10.76s	43.63s	47.78s	8.43m	7.39m	2.45m	2.09m
Year Actual	**100M**	**100W**	**400M**	**400W**	**MLJ**	**WLJ**	**MHJ**	**WHJ**
2000	10.05s	11.15s	44.92s	49.92s	8.26m	6.80m	2.32m	1.98m
2004	9.93s	11.04s	44.67s	50.00s	8.33m	6.92m	2.31m	1.98m

and 2004 Olympics Games. It can be observed that both 100m times for men exceeded the prediction, whereas the female 100m times were well below the predicted values. For the 400m, long jump and high jump both male and female athletes were below the predicted time and distances. However, the men's actual 400m times and long jump distances did show improvements from 2000 to 2004. In the women's 400m there was a performance decline and in the women's and men's high jump performances remained relatively static from 2000 to 2004.

The mathematical models derived for swimming from pre-1998 data (Lackey and Heazlewood, 1998) for the men's and women's 50m, 100m, 200m, 400m, women's 800m and men's 1500m freestyle events are displayed in Table 3. The functions for the 50m (inverse), 200m (sigmoidal) and 400m (cubic) are the same for both men and women, however the functions for the 100m (men compound and women cubic) and the longer distances displayed their own specific function (800m women sigmoidal and men 1500m cubic). The non significance for the men's and women's 50m freestyle equations and R^2 values is a result of the small degrees of freedom when calculating the level of significance, due to the 50m

freestyle only being included in the Olympic swimming program from 1988.

The comparison of the predicted times with the actual times for each event and displayed in Table 4, indicates congruence for the men's 50m and women's 50m and 100m. In all other events for both men and women the predicted times are faster than actual times, indicating rate of progress in these events appears to have slowed down based on data up to 1996. This indicates the prediction equations over estimated the rates of improvement.

The predictions were closer for the shorter swimming events where men's 50m and women's 50m and 100m, where actual times are almost identical to predicted times. In both men and women, as the swim distances increased, the accuracy of the predictive model decreased, where predicted times were 4.5-7% faster than actual times achieved. For example, the predicted men's 1500m of 489.03s for 2004 was 7% faster than the actual time of 509.06s.

DISCUSSION

The results indicate the ability to predict performances into the near future based on past performances is possible, however both sets of

Table 3. Men's and women's freestyle swimming events, type of mathematical function and predictive equations.

Gender	Event	Type	Equation	R^2 value
Men	50m	Inverse	$y = -44.77 + 134064.199 / year$	.543 ns
	100m	Compound	$y = 42747.22 \times .99^{year}$	.972 *
	200m	Sigmoidal	$y = e^{-1.33 + 11971.98 / year}$	.862 *
	400m	Cubic	$y = 12572.6 - 8.79 \times year + 6.53 \times 10^{-7} \times year^3$	.977 *
	1500m	Cubic	$y = 40457 - 27.31 \times year + 1.88 \times 10^{6} \times year^3$	.963 *
Women	50m	Inverse	$y = -71.68 + 193578 / year$	.732 ns
	100m	Cubic	$y = 7417.11 - 5.48 \times year + 4.5 \times 10^{-7} \times year^3$	.959 *
	200m	Sigmoidal	$y = e^{-0.54 + 8451.02 / year}$	.812 *
	400m	Cubic	$y = 1565 - 1.68 \times 10^{-7} \times year^3$	.976 *
	800m	Sigmoidal	$y = e^{-1.26 + 14938.8 / year}$	.692 *

ns: due to the small degrees of freedom; * p < 0.05

Table 4. Actual 1996 means, predicted and actual means for 2000 and 2004 Olympics. All times are in seconds.

Event		Actual 1996	Predicted/Actual 2000	Predicted/Actual 2004
Men's	50m Freestyle	22.47	22.26/22.19	22.13/22.11
	100m Freestyle	49.31	48.17/48.95	47.53/48.80
	200m Freestyle	108.40	105.57/107.43	104.32/106.47
	400m Freestyle	230.57	219.03/226.21	215.29/225.92
	1500m Freestyle	910.29	854.72/901.67	835.76/898.05
Women's	50m Freestyle	25.37	25.11/25.01	24.91/24.96
	100m Freestyle	55.37	54.81/54.76	54.53/54.58
	200m Freestyle	119.95	117.90/118.89	116.91/118.69
	400m Freestyle	250.00	238.98/249.46	233.91/248.05
	800m Freestyle	514.89	496.38/507.82	489.03/509.06

results derived from athletics and swimming indicate that in many events the improvements in performances were overestimated. Although improvement in the majority of athletic and swimming events studied did occur in absolute times or distances, the rate of improvement was slower than predicted by the numerous equations generated. The reality is, in some events performances have been static such as men's and women's high jump and women's 200m freestyle or actually declined such as women's 400m and women's 800m freestyle. The two events where actual performances exceeded, very slightly the predicted performances, were the men's 100m and men's 50m freestyle.

Recent emphasis on successful drug testing may have impacted more on women athletes than men, where the slowing down, and in some cases the declines in actual performances were noted for women.

Performances are expected to improve over a period of time due to a number of interacting sports scientific, ontogenetic (lifespan) and pharmacological factors, such as:

1. The use of more efficient running, jumping and swimming techniques, a biomechanical construct.
2. Improved training programs, which are exercise physiological and functional anatomical construct.
3. Enlarged population of athletes due to increased participation by more nations from which high performance athletes and swimmers are drawn. This will result in an increased sample from the human gene pool, a genetic construct.
4. Improved talent identification programs designed and implemented by national sporting organisations and sports institutes that will select

and develop tomorrow's high performance athletes.
5. Changes in human physiology, such as the recent ontogenetic trends of increasing height and weight in Australia.
6. The use of performance enhancing drugs legal or illegal, especially androgenic- anabolic steroids and human growth hormone, which have a masculinising effect on women or the use of neutraceuticals (functional foods).

However, the exact mathematical trends these improvements and interactions take can be plotted mathematically to reveal future trends across athletic and swimming events. In effect, the actual performance is a summary of all these factors. In some cases the event can be predicted with a high degree of accuracy, such as the sprint swimming events for both men and women, whereas in other events such as 400m women and 800m freestyle are not predicted well as the events are currently displaying performance declines which were not identified by the mathematical models. It must be emphasised that all the models across all events indicated consistent improvements over time.

The primary purposes behind this type of predictive research are that we might understand the realistic limits to human improvement in many sports, to set new and realistic goals that athletes will have to achieve to make representative teams and Olympic finals, to provide a more coherent understanding of what performances of the past suggest about performances of the future, to understand if different events represent changes which reflect developments of human biomechanical, exercise physiological, motor learning and sport psychological functions as expressed in sport; and as a intellectual exercise to understand more completely the complex trends that underpin human evolution and training adaptation that are expressed in the sports arena.

CONCLUSION

As a heuristic exercise the derivation of mathematical-statistical models that predict changes in human sporting performance both in the near and distant future occupies definitely the minds of mathematicians and statisticians and if "we get it right" we will have a crystal ball into the future of sport. The problem is it just takes time to find out how good we are at solving such problems.

REFERENCES

Athletics Australia (2004) *Nomination criteria for 2004 Olympic Games Athens, Greece – August 2004 Athletics Australia*. Available from URL: *http://www.athletics.org.au/events/selection/dp_details.cfm?ObjectID=108.*

Banister, E. and Calvert, T. (1980) Planning for future performance: Implications for long term training. *Canadian Journal of Applied Sports Science* **5**, 170-176.

Edwards, D. and Hopkins, W. (1979) Targets for the future. *Modern Athlete and Coach* **17(3),** 34-36.

Federation Internationale de Natation Amateur. (1997) Welcome to FINA. Available form URL: *http://www.fina.org/.*

Heazlewood, I. and Lackey, G. (1996) The use of mathematical models to predict elite athletic performance at the Olympic Games. In: *Third Conference on Mathematics and Computers in Sport, 30 September - 2 October, Bond University, Queensland, Conference Proceedings.* Ed: deMestre, N. 185-206.

Hopkins, W. (2000) Limits to performance. *SportScience.* Available form URL: *http://www.sportsci.org/jour/0002/inbrief.html#limits*

Jokl, E. and Jokl, P. (1976a) Running and swimming world records. *Olympic Review* **107/108**, 536-543.

Jokl, E. and Jokl, P. (1976b) Running and swimming world records. *British Journal of Sports Medicine* **10**, 203-208.

Jokl, E. and Jokl, P. (1977) Running and swimming world records. *Journal of Sports Medicine and Physical Fitness* **17**, 213 - 229.

Lackey, G. and Heazlewood, I. (1998) The use of mathematical models to predict elite swimming performance. In: *Fourth Conference on Mathematics and Computers in Sport. 9-12 July, Bond University, Queensland, Conference Proceedings.* Eds: deMestre, N and Kumar, K. .79-110.

Norušis, M. (1993) *SPSS for Windows: Base Systems User's Guide. Release 6.0.* SPSS Inc, Chicago, USA.

Péronnet, F. and G. Thibault, G. (1989) Mathematical analysis of running performance and world running records. *Journal of Applied Physiology* **67**, 453-465.

Prendergast, K. (1990) What do world running records tell us? *Modern Athlete and Coach* **28**, 33-36.

Rothstein, A. (1985) *Research design and statistics for physical education*. Prentice Hall, Englewood Cliffs, USA.

SPSS Inc. (1994) *SPSS for Windows Release 6.1-Computer Software*. SPSS Inc, Sydney, Australia.

Wallechinsky, D. (1996) *The complete book of the Olympics 1996 edition*. Aurum Press, London, UK.

Wikipedia. (2006) Results of the 2000 and 2004 Olympic Games. Available form URL: *http://en.wikipedia.org/wiki/2004_Summer_Olympics.*

KEY POINTS

- Prediction of future Olympic performance based on previous performance trends.
- Application of non-linear mathematical equations resulting in better fitting models.
- Application of mathematical predictive models to the Olympic sports of athletics and swimming.
- Accuracy of mathematical models in predicting sprint events in running and swimming.
- A research approach to predict future Olympic performance and set future performance standards that could be applied to other sports.

AUTHOR BIOGRAPHY

Timothy Heazlewood
Employment
Course Coordinator, School of Exercise Science NSW, Faculty of Health Sciences, Australian Catholic University.
Degree
B.Sc. (Hons), Dip. Ed., M.Ed., Ph.D. (Syd).
Research interests
Mathematical modelling, laboratory predictors of competition performance, research methods and statistics as applied to exercise and sport science, especially with high performance athletes.
E-mail: t.heazlewood@mackillop.acu.edu.au

✉ **Dr Timothy Heazlewood**
School of Exercise Science, Australian Catholic University, Strathfield, NSW, Australia