The effects of measurement error and testing frequency on the fitness-fatigue model applied to resistance training: a simulation approach.

STEPHENS HEMINGWAY, B.H.S., BURGESS, K.E., ELYAN, E. and SWINTON, P.A.

2020

The Effects of Measurement Error and Testing Frequency on the Fitness Fatigue Model Applied to Resistance Training: A simulation approach

Benedict H. Stephens Hemingway [a], Katherine E. Burgess [a], Eyad Elyan [b], Paul A. Swinton [a]*

[a] *School of Health Sciences, Robert Gordon University, Aberdeen, UK*
[b] *School of Computing Science and Digital Media, Robert Gordon University, Aberdeen, UK*

Corresponding Author*

Paul A Swinton
School of Health Sciences Office, Faculty of Health and Social Care
Robert Gordon University
Garthdee Road, Garthdee, Aberdeen
United Kingdom
AB10 7QG

Telephone: 01224 262 3361

Email: p.swinton@rgu.ac.uk

Disclosure of Funding

No funding was received for this work

Conflicts of Interest

None

Running Title

Fitness Fatigue Model Simulation

Abstract

This study investigated the effects of measurement error and testing frequency on prediction accuracy of the standard Fitness-Fatigue Model. A simulation-based approach was used to systematically assess measurement error and frequency inputs commonly used when monitoring the training of athletes. Two hypothetical athletes (intermediate and advanced) were developed and realistic training loads and daily 'true' power values generated using the fitness-fatigue model across 16-weeks. Simulations were then completed by adding Gaussian measurement errors to true values with mean 0 and set standard deviations to recreate more and less reliable measurement practices used in real-world settings. Errors were added to the model training phase (weeks 1-8) and sampling of data used to recreate different testing frequencies (every day to once per week) when obtaining parameter estimates. In total, 210 sets of simulations (n=10^4 iterations) were completed using an iterative hill-climbing optimisation technique. Parameter estimates were then combined with training loads in the model testing phase (weeks 9-16) to quantify prediction errors. Regression analyses identified positive associations between prediction errors and the linear combination of measurement error and testing frequency (R^2_{adj}=0.87-0.94). Significant model improvements (P<0.001) were obtained across all scenarios by including an interaction term demonstrating greater deleterious effects of measurement error at low testing frequencies. The findings of this simulation study represent a lower-bound case and indicate in real-world settings where a fitness-fatigue model is used to predict training response, measurement practices that generate coefficients of variation greater than ≈4% will not provide satisfactory results.

Introduction

The fitness-fatigue model has been used for decades primarily as a conceptual framework that describes the training process [1]. In its most basic form, the model posits that a single bout of training creates two antagonistic after-effects including a positive long-lasting and low-magnitude fitness effect, and a negative short-lasting and high-magnitude fatigue effect. These antagonist components then combine to describe an athlete's performance and state of preparedness. Several mathematical implementations of the fitness-fatigue model have also been developed to more directly inform training program design [2-5]. Each fitness-fatigue model is described by a mathematical equation tailored to an individual athlete that links quantification of training load (input) to a performance measure (output). Individual tailoring is achieved by setting parameters in the equation to match the magnitude and decay rate of the positive and negative after-effects experienced by the specific athlete. This is achieved in practice by performing a period of training and performance measurement with parameters retrospectively fit to best match the input and output data generated. This process is referred to as the 'model training' phase and once complete, the model and fitted parameters can be used to predict future responses to physical training and inform its design [6]. Accurate quantification of training load and regular best-effort criterion trials (e.g. timed run or load lifted) revealing the athlete's current capabilities are therefore required [7]. fitness-fatigue models have traditionally been applied to endurance athletes including runners [7,8], swimmers [9-11] and triathletes [12,13] as training loads are relatively simple to calculate and criterion trials closely match sporting performance.

A small number of studies have investigated the fit of fitness-fatigue models with performance in individual and team sports where strength and power are the primary fitness components [3,14,15]. Busso et al. [3,14] reported adequate to strong fit (r=0.53 to 0.97) between modelled and actual clean and jerk performance of 6 elite weightlifters over a 1-year period. Strong fit (r=0.95) was also obtained for a 37-week case study comprising a hammer thrower engaged in diverse training that included resistance exercise, sport-specific weighted movements, sprints and plyometrics [16]. A more recent study by Graham et al. [15] fitted individual fitness-fatigue models to training load and match performance data across a team of Australian Rules Football players throughout a 24-week in-season macrocycle. Moderate to strong model fit ($r = 0.56 - 0.89$) was demonstrated depending upon the method used to quantify training load. However, each of these studies only assessed the ability to retrospectively fit input and output data as part of the model training phase and did not include 'hold-back' data sets to quantify prediction accuracy. This represents a major limitation of the research base as the central premise of mathematical fitness-fatigue models is to predict future responses to training.

Increased use of the fitness-fatigue model with team sport athletes may require a shift in emphasis where training data is used to predict response in terms of fitness variables [17] rather than sporting performance which is likely to demonstrate more complex relationships with training loads. Such a change in emphasis would also match the conceptual framework adopted by many strength and conditioning coaches where capacity of an athlete is viewed in terms of dimensions of fitness (e.g. strength and power) [17, 18]. Activities such as the vertical jump or bench press which are widely used to monitor athletes and assess improvements in fitness could then be used to fit models and characterise individual response [17]. Previous research has shown that multiple factors including precision and frequency of measurements influence accuracy of parameter estimates [19]. As the fitness-fatigue model is a non-linear model, it has been suggested that between 15 and 200 performance tests over a training period may be required to obtain stable estimates. As a result, the ability to model and predict fitness of team sport athletes as suggested above, may require performance tests that can be completed at high weekly frequencies. The vertical jump is the most popular means of obtaining frequent assessments of an athlete's physical capability and can be used daily without causing acute or chronic declines in performance [20] Additionally, a range of mechanical variables (e.g. impulse, power, rate of force development) can be extracted during vertical jumps to assess various features of the neuromuscular system [17, 21].

Several challenges exist in researching the effectiveness of the fitness-fatigue model to predict athlete fitness and identify the importance of factors such as measurement error and testing frequency. Within standard sport science designs the primary challenge is the recruitment of large sample sizes to perform daily testing across various measurement procedures to accurately isolate the effects of measurement error and frequency on prediction accuracy. In addition, the existence of error in all measurements precludes 'true' underlying performance of an athlete to be known [22] placing limits on the ability to assess predictions. Due to these challenges an alternative approach employing simulation techniques is applied in the present study to systematically investigate the effects of plausible measurement errors and frequencies. The approach adopted represents a best-case scenario with the assumption that response to training is completely specified by the fitness-fatigue model and 'observed' performances deviate due to measurement error only. Adopting this best-case approach sets a lower bound whereby similar practices in real-world settings can on average only lead to greater predictive errors. By simulating thousands of training responses across a distributed range of measurement errors and frequencies, this study provides unique insight into the use of the fitness-fatigue model and was also designed to identify whether certain measurement errors and testing frequencies used in practice or research could be identified to have no potential validity in real-world settings.

Materials and Methods

Experimental approach to the problem

A simulation-based approach was adopted to quantify the effects of measurement error and testing frequency on fitness-fatigue model parameter estimates and performance prediction for two hypothetical athletes (intermediate and advanced). The vertical jump was selected as the performance measurement tool due to its popularity in athlete monitoring and potential to be used daily [20]. A range of mechanical variables including power, impulse and jump height were considered for the simulation study. However, each of the variables demonstrated similar relative profiles with regards to change in magnitude across an intervention compared to measurement error and therefore each outcome would result in the same conclusions produced by this study. Power produced during the vertical jump was ultimately selected for simulation as the measurement has previously been used in mathematical models to predict player fitness in response to training dose [17].

Two popular training load distributions ("summated microcycles" and "wave-like" distribution; Figure 1) were combined with athlete specific parameters to generate realistic daily power values over a 16-week period. The generated data represented known 'true' values that are accessible only within simulation-based approaches. The generated values were split in half to create an initial 'model training' set (weeks 1-8) and a 'hold back' data set (weeks 9-16) to assess prediction error. The effects of measurement error and testing frequency were assessed by adding realistic errors to true values in the training set whilst fitting the fitness-fatigue model to varying proportions of the augmented data (observed = true + error). The process replicated the situation adopted in real-world settings where observed scores on a physical test comprise the athlete's true score and measurement error [22]. The simulation approach, however, incorporated the additional assumption that the fitness-fatigue model completely specified the athlete's response to training. Parameter estimates obtained from the model training set were then combined with training loads from the hold back data to obtain predicted power values and their associated prediction errors. Finally, extensive simulations were completed for each scenario to obtain distributional estimates. A detailed flowchart illustrating the simulation process is presented in Figure 2.

Development of Hypothetical athletes

Simulated data representing true performance change for two hypothetical athletes (intermediate and advanced) were developed based on the inverse relationship between experience and improvement [23]. Research investigating change in vertical jump power from a single intervention has demonstrated that improvements in peak power (W) for moderately trained athletes generally range between 0 and 20% [24], whereas improvements for advanced athletes generally range between 0 and 5% [25, 26]. Based on these findings, increases of 15% and 5% were chosen for the intermediate and advanced athletes over the 16-week period, respectively. The same research base [24-26] was also used to identify realistic baseline values.

Development of Training Loads

Two characteristic training load distributions (TRIMP values) were developed for each athlete. The first (TRIMP-1) followed a summated microcycles distribution, in which each 4-week mesocycle comprised 3 weeks of progressive loading followed by 1 week of deloading [27]. The second (TRIMP-2) followed a wave-like pattern where training load gradually increased and oscillated over each 4-week mesocycle [28, 29]. TRIMP values and their scaling across the two hypothetical athletes are presented in figure 1.

Development of Athlete Specific Parameters

The standard fitness-fatigue model (eq.1) was used to fit all models in the present study.

$$p(t) = p(0) + k_1 \sum_{s=0}^{t-1} e^{-\frac{t-s}{\tau_1}} \cdot w(s) - k_2 \sum_{s=0}^{t-1} e^{-\frac{t-s}{\tau_2}} \cdot w(s) \qquad (\text{eq}.1)$$

Where $p(t)$ is the performance on day t, k_1 and k_2 are weighting factors that translate the units of the training load to the fitness and fatigue effects of the performance measure (power measured in Watts), respectively; τ_1 and τ_2 are decay constants controlling the decay time of fitness and fatigue effects, respectively; and w is the daily TRIMP value. Athlete specific parameters were obtained through a process of systematic parameter-space exploration. Briefly, desired end-performance values following 16-weeks of training were calculated for each athlete and an interval of $\pm$ 75 W constructed to provide an initial screening threshold. Simulations were run with 3.8×10^6 parameter sets $(k_1, k_2, \tau_1, \tau_2)$ constructed by incrementing values in a grid-like fashion. Approximately 2000 potential parameter sets were obtained for each athlete in which the end-performance value resided within the threshold set. These parameter sets were then plotted and visually investigated for realistic developments over the 16-weeks. This process reduced the number of parameter sets to approximately 10 for each athlete (Table 1; parameter ranges intermediate athlete k_1: 0.5-2.5, k_2: 1.0-4.0, τ_1: 14-37, τ_2: 5-19; parameter ranges advanced athlete k_1: 0.5-4.5, k_2: 1.0-5.0, τ_1: 6-25, τ_2: 5-15). A final selection was made (Table 2) based on further visual comparison to create upward trends with plateau, and ensuring parameter values and their ratios (k_1/k_2 and τ_1/τ_2) were consistent with previous research [30].

Model Simulations

Power values were generated for each athlete with the training loads and parameters described above using the fitness-fatigue model in eq.1 to represent true performance (Figure 3). Repeated simulations were then used to investigate the effects of error magnitude and training frequency. Measurement error was added to each true power value (in the initial 8-week model training block) to replicate testing in a real-world setting. Errors were added by random draws from a Gaussian distribution with mean zero and standard deviation representative of that obtained during a vertical jump. A review of literature identified that power is frequently measured via a force platform or linear position transducer [21, 31]. The former measurement tool calculates power via force and velocity values obtained through integration, and the latter calculates power via force and velocity values obtained from differentiation of displacement data. Reliability studies have reported coefficients of variation (CV) ranging from approximately 2 to 10%, with superior reliability obtained when using a force platform [21, 31]. As a result, standard deviations for Gaussian errors were derived for both hypothetical athletes by multiplying each CV value (2, 4, 6, 8, 10%) by the athlete's initial baseline power value ($p(0)$) and dividing by 100 (Table 3). An additional set of simulations for the advanced athlete incorporating the same absolute error used in the intermediate case were completed to facilitate further comparisons.

To simulate different testing frequency states, a proportion of power values were isolated to recreate the real-world setting of measuring performance once per week to each day (in unit increments). For example, every 7th power value with error was selected from the model training block when simulating the once per week condition. A total of 210 scenarios were investigated with each comprising 10^4 simulations (2.1×10^6 total simulations). Parameter estimates for each simulation were fitted under a parallel computing framework using least-squares regression via a limited-memory modification of the BFGS quasi-Newton method [32], in the optimization package *Optim* (a part of the R *stats* package, v.3.4.4). Parameter estimates for each simulation were combined with the corresponding TRIMPs across the entire 16-week block [33], with predictions for the hold-back data used to obtain prediction errors for subsequent analysis.

Statistical Analyses

For each set of 10^4 simulations, prediction errors were transformed into summary statistics representing distributional centrality ($\mathrm{DPE_M}$) and spread ($\mathrm{DPE_S}$) by calculating the median and distance between the 0.16 and 0.84 quantiles, respectively. Relationships between dependent variables (centrality: $\mathrm{DPE_M}$, spread: $\mathrm{DPE_s}$) and centred independent variables (measurement error and testing frequency) were quantified by multiple linear regression. Initially, the linear combination of measurement error and testing frequency expressed as continuous variables were entered into regression models. A second series of models featuring the linear combination and product of measurement error and testing frequency (interaction effect) were then included. Fit and suitability of each linear model was assessed with adjusted R^2 and residual analysis, respectively. Distributions of parameter estimates were described using descriptive statistics and ill-conditioning assessed via calculation of Pearson correlation coefficients [19, 30].

Quality Control

Systematic examination of the simulation source code (code review) was performed pre- and post-simulation deployment, to detect inaccuracies that would prevent successful implementation or cause erroneous results. A sensitivity analysis was conducted to assess the effects of different initial values on the non-linear least squares optimisation function. The sensitivity analysis comprised fitting the least squares algorithm with 100 different starting values across the parameter space and no substantive changes were noted from code featuring a single set of starting values comprising the true parameters. Optimisation convergence was set using a tolerance of 10^{-8} in the objective and found to be successful for approximately 99% of total parameters estimated within the experiment.

Results

Prediction errors

All analyses revealed positive associations between dependent variables (prediction error centrality and spread) and the independent variables measurement error and testing frequency. DPE_M was well explained by the linear combination of the two independent variables (Adjusted R^2 = 0.89-0.94) across all six athlete-TRIMP groupings (Figure 4). Regression coefficients for testing frequency ($\beta_1 = 19.1 - 26.4$) and measurement error ($\beta_2 = 20.8 - 25.4$) were shown to be significant ($P<0.001$) for each model assessed. The inclusion of an interaction term significantly ($P<0.001$) improved the fit of each model (Adjusted R^2 = 0.96-0.98) and demonstrated that the deleterious effect of increased measurement error on DPE_M was increased at lower testing frequencies. Similar results were obtained for DPE_S (Figure 5), with strong linear relationships obtained with testing frequency and measurement error (Adjusted R^2 = 0.87-0.91). Again, each regression coefficient was found to be significant ($P<0.001$) and all models were improved ($P<0.001$) with an interaction effect demonstrating greater deleterious effects of measurement error at low testing frequencies.

Comparisons of prediction errors between the advanced and intermediate athlete demonstrated a dependence on testing frequency. For high measurement frequencies the advanced athlete simulations for both TRIMP distributions demonstrated systematically lower prediction errors compared to the intermediate athlete. This finding was obtained for both DPE_M (mean±sd = 90±47 vs. 103±44 W, respectively) and DPE_S (mean±sd = 168±110 vs. 193±93 W, respectively), despite larger absolute measurement error values inputted to advanced athlete simulations. However, when testing frequency was low, prediction errors were similar for both athletes, and in some cases, slightly larger for the advanced athlete. When simulations were repeated using the same absolute error magnitude for both athletes, centrality and spread of prediction error were consistently lower for the advanced athlete across all conditions.

Model parameter estimates

In general, model parameter estimates $(k_1, \tau_1, k_2, \tau_2)$ displayed a range of different distributions across simulation scenarios. Estimates for the gain parameters (k_1, k_2) were unstable for both athletes, with boundary values frequently obtained when testing frequency was low. This effect was magnified when low testing frequency was combined with high measurement error. Distribution of the decay parameters (τ_1, τ_2) became increasingly right-skewed for both athletes as measurement error increased. This effect became more pronounced when large measurement errors were combined with low testing frequency. Correlations between parameter estimates (N=350,000 per athlete-TRIMP grouping; Table 4) revealed strong associations between gain parameters (k_1 and k_2: r = 0.88-0.99) for both athletes, thereby demonstrating ill-conditioning. Low to moderate strength negative correlations were also obtained between k_1 and τ_1(r = -0.63 to -0.29), and k_2 and τ_1 (r = -0.64 to -0.31) for both athletes.

Discussion

The present study comprised a unique and efficient design to investigate prediction accuracy of the standard fitness-fatigue model in a strength and conditioning context. The simulation approach provided an effective method to systematically assess the effects of two key challenges in athlete training modelling, namely, controlling measurement error and identifying appropriate measurement frequencies [30, 34, 35]. Whilst it is unrealistic to expect an athlete will respond deterministically to a series of training loads, the approach and underlying assumptions adopted in the present study provide important general information and informative lower bound cases for researchers and practitioners to consider. That is, the approach can identify and rule out specific practices that have no potential to be successful in real-world settings (but not rule in other practices).

The key findings of this study indicate that increased measurement error and reduced testing frequency across standard ranges encountered in practice meaningfully increase prediction errors. Additionally, variation in prediction errors were well explained by the simple linear combination of measurement error and testing frequency (Adjusted R^2 = 0.87-0.94). Regression coefficients showed that for every 1% increase in CV, the distribution of prediction errors increased by approximately 21-25W in centrality (DPE_M), and 45-63W in spread (DPE_S). Similarly, models showed that for a single day reduction in testing frequency, the distribution of prediction errors increased by approximately 19-26W in centrality (DPE_M) and 42-65W in spread (DPE_S). Theoretically, the standard fitness-fatigue model and traditional non-linear least squares methods used to obtain parameter estimates should demonstrate poor performance with high measurement error. The results of this simulation support this notion and show that if observed scores in a given performance test comprise error of more than 2-5% of an athletes baseline score, they are unlikely to be suitable for use with the fitness-fatigue model due to unacceptable prediction accuracy even in this most optimistic scenario where performance is directly specified by the model. The results demonstrate that when the model is fit to moderately inaccurate data (comprising error more than 5% CV), predictive errors become unacceptably high across all frequency conditions. For example, if measurements comprised ~6% CV, the results of this study suggest that even under high testing frequencies prediction errors of +150 W should be expected for intermediate/advanced athletes. For further context, an error of 150 W is equal to approximately 3% of the baseline scores, with total improvement across the entire training phase set at 5 and 15% for the advanced and intermediate athlete, respectively. Furthermore, these simulations represent a lower bound case, where additional real-world factors will further increase prediction errors.

Whilst a simple linear combination of the two independent variables explained most of the variation in prediction errors, significant improvements in model fit were obtained in all cases though inclusion of an interaction term. In each model the interaction demonstrated that deleterious effects of increased measurement error on prediction accuracy were amplified at lower testing frequencies. Viewed from the opposing perspective, the interaction effect demonstrated that very low measurement error (~2% CV) offered a protective effect on prediction accuracy even at low measurement frequencies. These findings suggest that if practitioners maintain very low measurement error (e.g. ≤2% CV) then use of the standard fitness-fatigue model may be viable despite low measurement frequencies (i.e. every 5-7 days). The lower bound case identified in this study suggested average prediction errors of approximately 50-100 W over an 8-week period with 2% CV and testing once per week. This finding aligns with previous research where adequate to strong fit (r = 0.53-0.97) was obtained with the fitness-fatigue model applied to resistance training data and performance measured once per week [3, 14]. The authors measured performance with 1RM tests which have been shown to demonstrate very low measurement error with CV values between 1 and 3% reported in literature [36, 37].

It is important to note that almost all previous studies conducted with the fitness-fatigue model have only included a model training phase and therefore potential for overfit given the four parameters available is likely. In contrast, the potential value of the fitness-fatigue model if its functional from is appropriate and parameters can be reliably estimated is to project into the future predicting an individual's response and thereby guide training prescription. Notably, the use of a cross-validation or 'hold back' data set is required to assess predictive capacity and should be considered compulsory for all future studies that assess fitness-fatigue models in practice. This recommendation corresponds with recent studies [33,38] demonstrating that fitness-fatigue models can generate moderately accurate predictions using data collected outside of a laboratory.

Parameter estimates obtained from the study were tested with correlations, with high values observed between the two magnitude factors, k_1 and k_2 across all simulations. This finding supports criticisms of ill-conditioning partially due to parameter inter-dependency, first raised by Hellard et al. [19] and further supported by Pfeiffer [30]. As discussed in detail by Hellard et al. [19], inter-dependency removes practical meaning from parameters as representations of an athlete's physiological state, and instead indicates the model has likely overfit, at the expense of accurate future predictions. Given the objective of uncovering meaningful parameters that characterise an individual's response to training, further research investigating fitness-fatigue models in a simulation environment may consider alternative parameter search methods, penalisation techniques to reduce parameter variability, and model reparameterisations [19].

Prediction errors and distributions of parameters estimates in the present study were similar across the two TRIMP designs for both athletes. In general, prediction errors for the advanced athlete were either less than or similar to those obtained for the intermediate athlete, despite greater absolute measurement errors applied for the former due to the scaling effect of CV. However, controlling for this effect by applying the same absolute measurement errors for both athletes resulted in lower prediction errors across all scenarios for the advanced athlete. These results indicate that a range of factors including heteroscedasticity in measurement error, absolute performance level and adaptive rate are likely to combine to influence the suitability of predictions. Further simulation work encompassing a wider range of training programs and realistic athlete development profiles should be conducted to identify cases where fitness-fatigue models have the greatest potential utility.

Conclusions

Whilst the simulation results presented here provide novel insights into the effects of measurement error and testing frequency on fitness-fatigue model prediction accuracy, there are a range of complex additional factors that would be expected to influence model prediction of an athlete's response to training. The primary aim of this study was to create lower bound estimates, where even in the absence of these additional factors, certain measurement errors and testing frequencies conditions would be considered ineffective to warrant use in practice or research. Collectively, the results of this study indicate that practitioners and researchers should focus on relevant performance tests that generate highly reliable data (≤4% CV). Additional processes including taking the average of multiple trials and filtering techniques can also increase reliability and should be considered. Even under high frequency conditions, the results of this study demonstrate that accurate predictions are not likely if measurement error is not minimised. Prior to investing time in data collection, it is recommended that practitioners and researchers adopt a simulation approach like the one applied here, where various measurement error and testing frequencies can be applied to training loads and adaptive rates realistic to each athlete being studied. Finally, it is recommended that future research investigating the use of fitness-fatigue models report prediction accuracy using cross-validation to appropriately evaluate the utility of the model to practitioners within the field of sport science.

References

1. Chiu LZ and Barnes JL. The fitness-fatigue model revisited: Implications for planning short-and long-term training. *Strength Cond J* 2003; 25: 42-51.
2. Banister EW, Calvert, TW, Savage MV and Bach T. A systems model of training for athletic performance. *Aust J Sports Med* 1975; 7: 57-61.
3. Busso T et al. A systems model of training responses and its relationship to hormonal responses in elite weight-lifters. *Eur J Appl Physiol Occup Physiol* 1990; 61: 48-54.
4. Calvert TW, Banister EW, Savage MV and Bach T. A systems model of the effects of training on physical performance. *IEEE Syst Man Cybern* 1976; 6: 94-102.
5. Taha T and Thomas SG. Systems modelling of the relationship between training and performance. *Sports Med* 2003; 33: 1061-1073.
6. Schaefer D, Asteroth A and Ludwig M. Training plan evolution based on training models. In *Innovations in Intelligent SysTems and Applications* 2015; 1-8.
7. Morton RH, Fitz-Clarke JR and Banister EW. Modeling human performance in running. *J Appl Physiol* 1990; 69: 1171-1177.
8. Banister EW and Hamilton CL. Variations in iron status with fatigue modelled from training in female distance runners. *Eur J Appl Physiol Occup Physiol* 1985; 54: 16-23.
9. Mujika I, Busso T, Lacoste L, Barale, F, Geyssant A and Chatard JC. Modeled responses to training and taper in competitive swimmers. *Med Sci Sports Exerc* 1996; 28: 251-258.
10. Hellard P, Avalos M, Millet G, Lacoste L, Barale F and Chatard JC. Modeling the residual effects and threshold saturation of training: A case study of Olympic swimmers. *J Strength Cond Res* 2005; 19: 67-75.
11. Ishii H, Takahashi S, Chiba T, Maeda A and Takahashi Y. Prediction of swim performance in junior female swimmers by dynamic system model. *Hum Perform Measurement* 2008; 5: 1-8.
12. Banister EW, Carter JB and Zarkadas PC. Training theory and taper: validation in triathlon athletes. *Eur J Appl Physiol Occup Physiol* 1999; 79: 182-191.
13. Millet GP, Candau RB, Barbier B, Busso T, Rouillon JD and Chatard JC. Modelling the transfers of training effects on performance in elite triathletes. *Int J Sports Med* 2002; 23: 55-63.
14. Busso T, Häkkinen K, Pakarinen A, Kauhanen H, Komi PV and Lacour JR. Hormonal adaptations and modelled responses in elite weightlifters during 6 weeks of training. *Eur J Appl Physiol Occup Physiol* 1992; 64: 381-386.
15. Graham SR, Cormack S, Parfitt G and Eston R. Relationships between model predicted and actual match performance in professional Australian footballers during an in-season training macrocycle. *Int J Sports Physiol Perform* 2017; 14:1-23.
16. Busso T, Candau R. and Lacour JR. Fatigue and fitness modelled from the effects of training on performance. *Eur J Appl Physiol Occup Physiol* 1994; 69: 50-54.
17. Revie M, Wilson K, Holdsworth R and Yule, S. On modelling player fitness in training for team sports with application to Glasgow Warriors Rugby Club. *Int J Sports Sci Coach.* 2017; 2: 183-193.

18. Carlock, J. M. et al. The relationship between vertical jump power estimates and weightlifting ability: a field-test approach. *J Strength Cond Res* 2004; 18: 534-539.
19. Hellard P, Avalos M, Lacoste L, Barale F, Chatard JC and Millet GP. Assessing the limitations of the Banister model in monitoring training. *J Sports Sci Med* 2006; 24: 509-520.
20. Watkins CM et al. Determination of vertical jump as a measure of neuromuscular readiness and fatigue. *J Strength Cond Res.* 2017; 31: 3305-3310
21. Cormack SJ, Newton RU, McGuigan MR and Doyle TL. Reliability of measures obtained during single and repeated countermovement jumps. *Int J Sports Physiol Perform* 2008; 3: 131-144.
22. Swinton PA, Hemingway BS, Saunders B, Gualano B and Dolan E. A statistical framework to interpret individual response to intervention: Paving the way for personalised nutrition and exercise prescription. *Front Nutr* 2018; 5: 41.
23. Appleby B, Newton RU and Cormie P. Changes in strength over a 2-year period in professional rugby union players. *J Strength Cond Res* 2012; 26: 2538-2546.
24. McBride JM, Triplett-McBride T, Davie A and Newton, RU. The effect of heavy-vs. light-load jump squats on the development of strength, power, and speed. *J Strength Cond Res* 2002; 16: 75-82.
25. Harris GR, Stone MH, O'Bryant HS, Proulx CM and Johnson RL. Short-term performance effects of high power, high force, or combined weight-training methods. *J Strength Cond Res* 2000; 14: 14-20.
26. Mangine GT, Ratamess NA, Hoffman JR, Faigenbaum AD, Kang J and Chilakos A. The effects of combined ballistic and heavy resistance training on maximal lower-and upper-body strength in recreationally trained men. *J Strength Cond Res* 2008; 22: 132-139.
27. Plisk SS and Stone MH. Periodization Strategies. *Strength Cond J* 2003; 25; 19-37.
28. Baker D. Applying the in-season periodization of strength and power training to football. *Strength Cond J* 1998; 20: 18-27.
29. Baker D. Cycle-length variants in periodized strength/power training. *Strength Cond J* 2007; 29: 10-17.
30. Pfeiffer M. Modeling the relationship between training and performance-a comparison of two antagonistic concepts. *Int J Comput Sci Sport* 2008; 7: 13-32.
31. Cronin JB, Hing RD and Mcnair PJ. Reliability and validity of a linear position transducer for measuring jump performance. *J Strength Cond Res* 2004; 18: 590-593.
32. Byrd RH, Lu P, Nocedal J and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput* 1995; 16: 1190-1208.
33. Ludwig M and Asteroth A. Predicting performance from outdoor cycling training with the fitness-fatigue model. In *Workshop Modelling in Endur Sports* 2016; 3-7.
34. Bourdon, P.C. et al. Monitoring athlete training loads: consensus statement. *Int J Sports Physiol Perform* 2017; 12: S2161-S2170.
35. Clarke DC and Skiba PF. Rationale and resources for teaching the mathematical modelling of athletic training and performance. *Adv Physiol Educ* 2013; 37: 134-152.

36. McCurdy K, Langford G, Jenkerson D and Doscher M. The validity and reliability of the 1RM bench press using chain-loaded resistance. *J Strength Cond Res* 2008; 22: 678-683.
37. Urquhart BG, Moir GL, Graham SM and Connaboy C. Reliability of 1RM split-squat performance and the efficacy of assessing both bilateral squat and split-squat 1RM in a single session for non–resistance-trained recreationally active men. *J Strength Cond Res* 2015; 29: 1991-1998.
38. Williams S, West S, Howells D, Kemp SP, Flatt AA and Stokes K. Modelling the HRV response to training loads in elite rugby sevens players. *J Sport Sci Med* 2018; 17: 402-408.

Table 1: Athlete specific parameter sets $(k_1, k_2, \tau_1, \tau_2)$ creating realistic improvements

Athlete	k_1	τ_1	k_2	τ_2	Athlete	k_1	τ_1	k_2	τ_2
INT	0.5	18	1.0	5	ADV	4.5	8	5.0	7
INT	2.5	14	3.5	9	ADV	3.5	6	4.0	5
INT	2.5	19	4.0	11	ADV	0.5	12	1.0	15
INT	0.5	37	1.5	9	ADV	1.5	10	2.0	7
INT	1.0	22	2.0	9	ADV	2.5	10	3.0	8
INT	1.0	19	1.5	10	ADV	0.5	20	1.0	9
INT	0.5	31	1.0	11	ADV	0.5	25	1.0	11
INT	1.5	26	2.0	17	ADV	1.0	19	1.5	12
INT	2.5	25	3.0	19	ADV	2.5	16	3.0	13

INT: Intermediate | **ADV**: Advanced. Top row for each athlete includes the parameter set used for simulations.

Table 2: Athlete specific parameters $(k_1, k_2, \tau_1, \tau_2)$, initial starting values $p(0)$ and end-values $p(112)$.

		Baseline performance (W)	Final performance (W)	True Parameters			
Athlete	Change (%)	$p(0)$	$p(112)$	k_1	τ_1	k_2	τ_2
INT	~ 15%	4500	~ 5175	0.50	18	1.0	5
ADV	~ 5%	5250	~ 5500	4.50	8	5.0	7

INT: Intermediate | **ADV**: Advanced

Table 3: Standard deviation (Watts) of the Gaussian error distribution with mean 0, from which random measurement error was drawn and applied to each known true-value within every individual simulation. Categorised by error (CV%) condition and athlete.

	SD of Gaussian error distributions by error (CV%) state				
Athlete	2%	4%	6%	8%	10%
Intermediate	90 W	180 W	270 W	360 W	450 W
Advanced	105 W	210 W	315 W	420 W	525 W

Table 4: Correlations between estimated model parameters for simulated scenarios within athlete-TRIMP groupings ($\mathrm{N} = 10000$ parameter sets, per scenario).

	Correlation Coefficient							
	INT: TRIMPS-1		**INT: TRIMPS-2**		**ADV: TRIMPS-1**		**ADV: TRIMPS-2**	
Model Parameter	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
$k_1 - k_2$	0.99	0.009	0.990	0.013	0.98	0.018	0.978	0.025
$k_1 - \tau_1$	-0.57	0.031	-0.560	0.036	-0.38	0.057	-0.382	0.056
$k_1 - \tau_2$	0.28	0.231	0.237	0.221	-0.05	0.112	-0.084	0.089
$k_2 - \tau_1$	-0.58	0.032	-0.576	0.036	-0.39	0.049	-0.398	0.050
$k_2 - \tau_2$	0.23	0.253	0.186	0.244	-0.11	0.125	-0.151	0.100
$\tau_1 - \tau_2$	0.16	0.354	0.217	0.348	0.63	0.278	0.689	0.219
$\tau_1 - k_1/k_2$	-0.09	0.174	-0.073	0.165	-0.13	0.233	-0.049	0.195
$\tau_2 - k_1/k_2$	0.06	0.225	0.039	0.213	0.19	0.114	0.150	0.104

INT: Intermediate | **ADV**: Advanced | **SD:** Standard Deviation

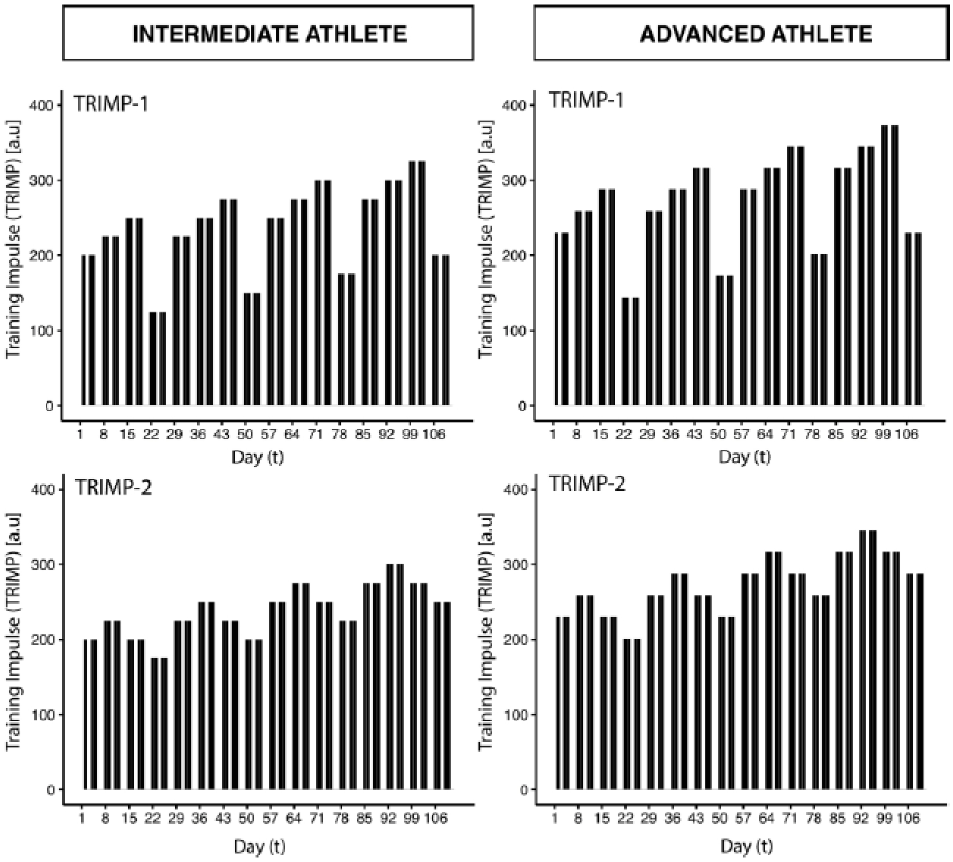
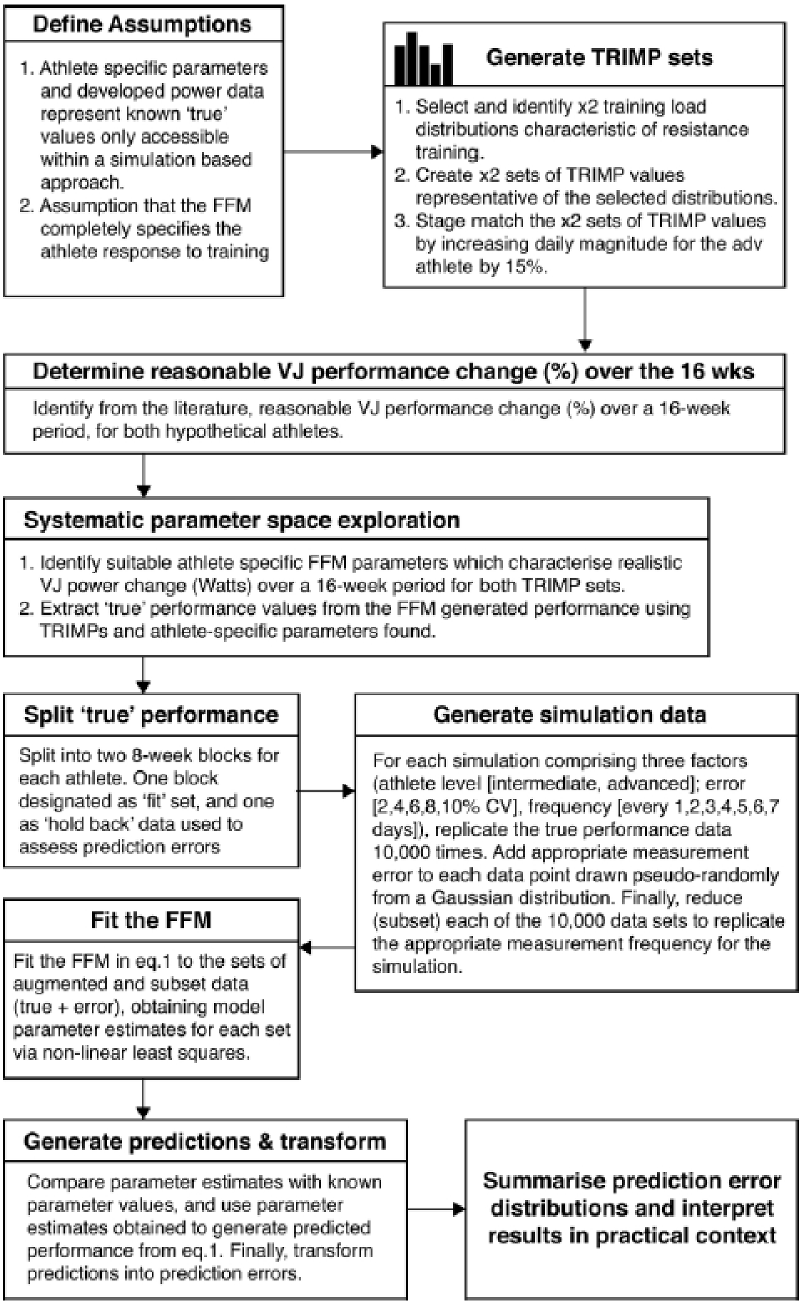


Figure 1: Distributions of scaled TRIMP values for hypothetical athletes over 16-weeks (Measured in arbitrary units [a.u])



Define Assumptions

1. Athlete specific parameters and developed power data represent known 'true' values only accessible within a simulation based approach.
2. Assumption that the FFM completely specifies the athlete response to training

Generate TRIMP sets

1. Select and identify x2 training load distributions characteristic of resistance training.
2. Create x2 sets of TRIMP values representative of the selected distributions.
3. Stage match the x2 sets of TRIMP values by increasing daily magnitude for the adv athlete by 15%.

Determine reasonable VJ performance change (%) over the 16 wks

Identify from the literature, reasonable VJ performance change (%) over a 16-week period, for both hypothetical athletes.

Systematic parameter space exploration

1. Identify suitable athlete specific FFM parameters which characterise realistic VJ power change (Watts) over a 16-week period for both TRIMP sets.
2. Extract 'true' performance values from the FFM generated performance using TRIMPs and athlete-specific parameters found.

Split 'true' performance

Split into two 8-week blocks for each athlete. One block designated as 'fit' set, and one as 'hold back' data used to assess prediction errors

Generate simulation data

For each simulation comprising three factors (athlete level [intermediate, advanced]; error [2,4,6,8,10% CV], frequency [every 1,2,3,4,5,6,7 days]), replicate the true performance data 10,000 times. Add appropriate measurement error to each data point drawn pseudo-randomly from a Gaussian distribution. Finally, reduce (subset) each of the 10,000 data sets to replicate the appropriate measurement frequency for the simulation.

Fit the FFM

Fit the FFM in eq.1 to the sets of augmented and subset data (true + error), obtaining model parameter estimates for each set via non-linear least squares.

Generate predictions & transform

Compare parameter estimates with known parameter values, and use parameter estimates obtained to generate predicted performance from eq.1. Finally, transform predictions into prediction errors.

Summarise prediction error distributions and interpret results in practical context

Figure 2: Flowchart describing the simulation process

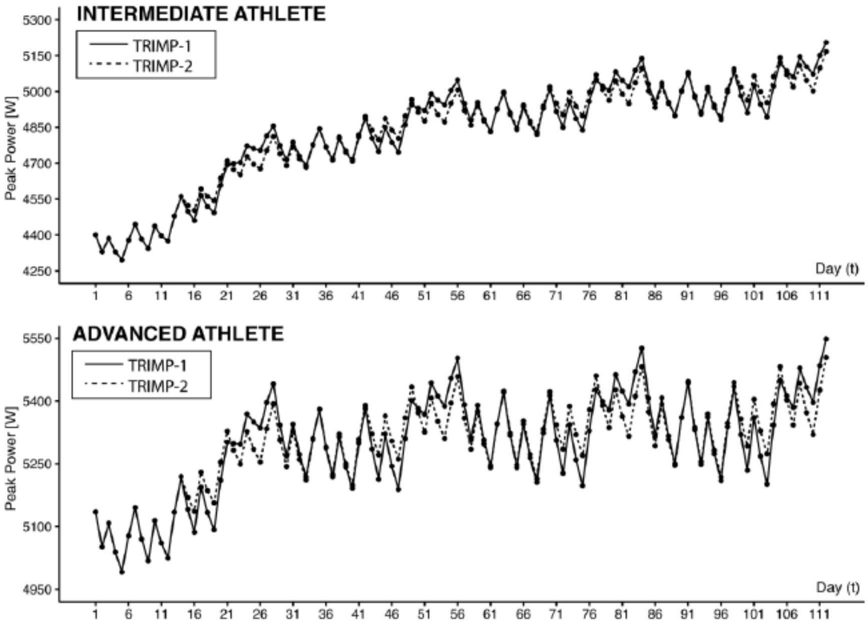


Figure 3: True vertical jump power values simulated across the 16-weeks with two training load distributions (TRIMP-1, TRIMP-2).

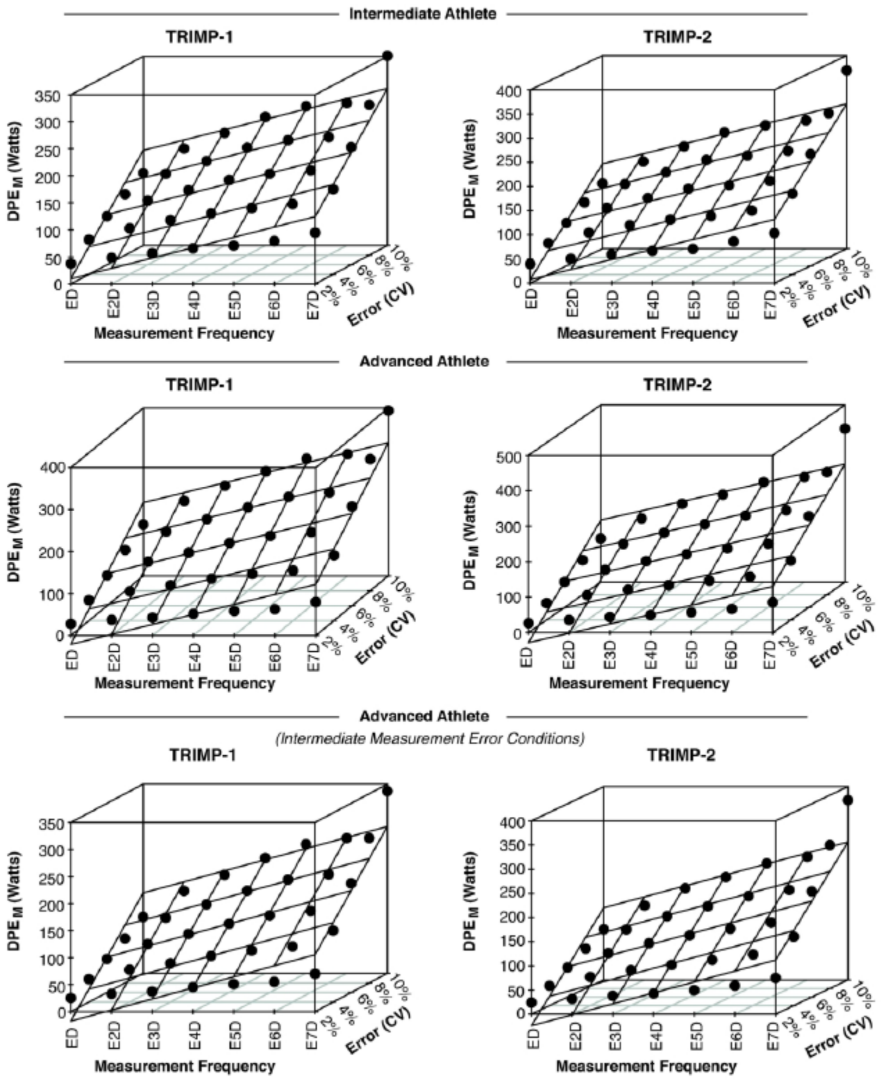


Figure 4: Regression planes illustrating relationships between prediction errors (centrality of distribution) and independent variables (measurement error and testing frequency).

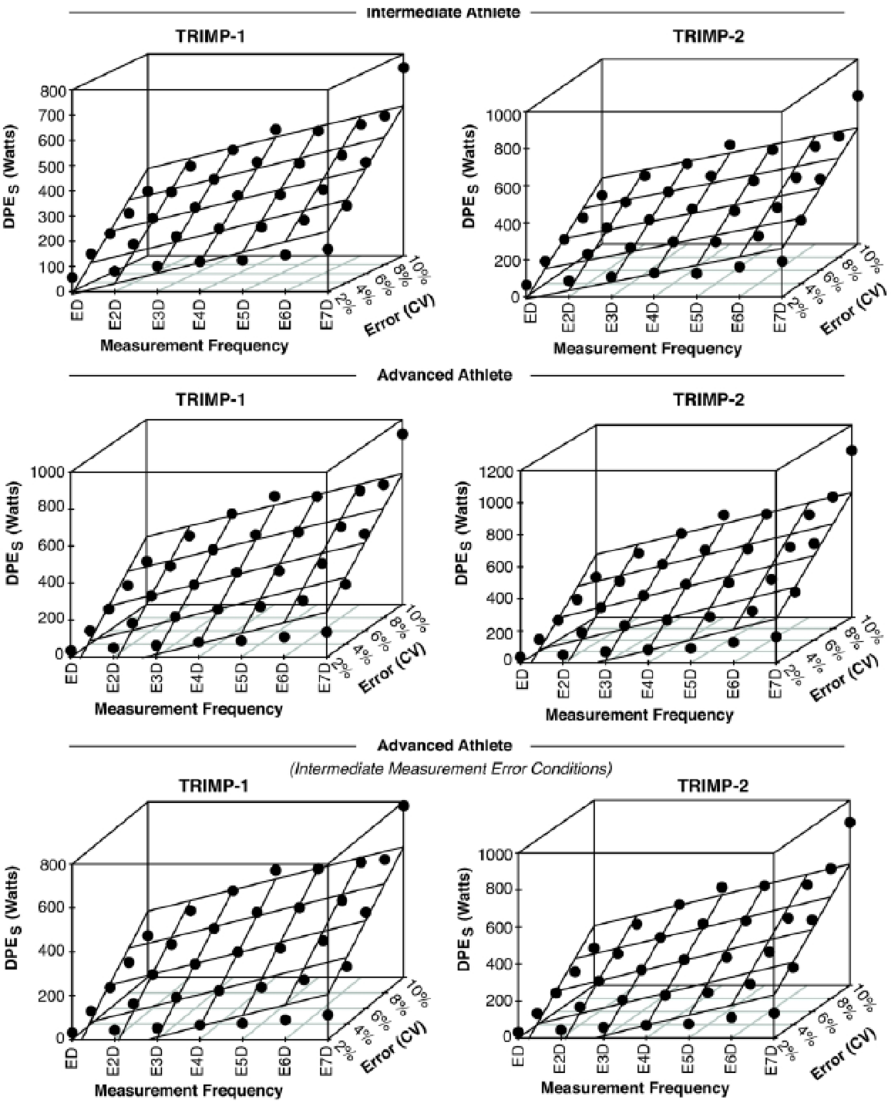


Figure 5: Regression planes illustrating relationships between prediction errors (spread of distribution) and independent variables (measurement error and testing frequency).