International Biomechanics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbbe20

Selecting the number of trials in experimental biomechanics studies
Stephanie E. Forrester
a
a Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, UK
Published online: 01 Jun 2015.

To cite this article: Stephanie E. Forrester (2015) Selecting the number of trials in experimental biomechanics studies, International Biomechanics, 2:1, 62-72, DOI: 10.1080/23335432.2015.1049296
To link to this article: http://dx.doi.org/10.1080/23335432.2015.1049296

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Versions of published Taylor & Francis and Routledge Open articles and Taylor & Francis and Routledge Open Select articles posted to institutional or subject repositories or any other third-party website are without warranty from Taylor & Francis of any kind, either expressed or implied, including, but not limited to, warranties of merchantability, fitness for a particular purpose, or non-infringement. Any opinions and views expressed in this article are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor & Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Terms & Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

It is essential that you check the license status of any given Open and Open Select article to confirm conditions of access and use.
Selecting the number of trials in experimental biomechanics studies

Stephanie E. Forrester*

Wolfson School of Mechanical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, UK

(Received 27 August 2014; accepted 20 April 2015)

Experimental biomechanics studies often involve the comparison of mean values from individuals across two or more experimental conditions. The purpose of this study was to evaluate two existing methods for determining the number of trials necessary to estimate these means. The sequential estimation technique (SET) was investigated in terms of the influence of input data distribution on the outcome. Paired samples $t$-tests were investigated in terms of the interaction between the number of subjects and number of trials necessary to achieve an acceptable level of statistical power. Simulation models were developed to perform SET and paired samples $t$-tests on representative synthetic input data. The SET results confirmed that the number of trials to achieve a stable estimate of the mean is independent of the input distribution provided the mean and standard deviation are fixed. For the commonly used 20 reference trials and 0.25 standard deviation threshold $9 \pm 8$ trials were needed to achieve stability. The paired $t$-test results confirmed that both number of subjects and number of trials can have a marked effect on the statistical power, e.g. a power of 0.80 can be achieved for effect size of 0.80 using 15 subjects and at least 19 trials or 20+ subjects and only 3 trials. The SET method suffers from arbitrary convergence criteria and neglecting intra-subject variance and, thus, should be applied with extreme caution. In contrast, statistical power can provide a more objective and conclusive means for determining the number of trials required for a given experimental situation.

Keywords: sequential estimation technique; statistical power; experimental biomechanics

Introduction

The averaging of key performance variables across multiple trials of an individual is a common procedure in experimental biomechanics studies. The number of trials typically varies from 3 to 5 in jumping (e.g. Unick et al. 2005; Walsh et al. 2007) to 30+ for treadmill walking/running (e.g. Foissac et al. 2009; McGhee et al. 2013). However, even within the same type of investigation the number of trials used can vary widely between studies. For example, within the scientific literature on breast biomechanics during treadmill running the number of gait cycles analysed ranges from 5 (White et al. 2011), through 15 (Zhou et al. 2012) to 30 (McGhee et al. 2013). Notably none of these studies provide strong justification for the number of trials used. Generally it is assumed that a sufficient number of trials have been used such that the resulting mean provides an ‘acceptable’ estimate of the key performance variable (Bates et al. 1983; Salo et al. 1997). Since variability is intrinsic to all human movement, using too few trials may result in a mean that poorly represents the individual’s long-term technique. While increasing the number of trials can help to overcome this limitation, there are often restrictions on the number of trials an individual can complete, e.g. due to fatigue.

From a statistical perspective, there are well-established methods for relating the number of trials to the confidence intervals on the estimated mean (Triola 2004). For unknown population variances (as can be considered the norm in experimental biomechanics studies), the $t$-statistic is recommended (Figure 1). Regardless of method, the overall trend is for the confidence intervals to decrease exponentially with increasing sample size. Although this reinforces the argument noted above regarding the use of too few trials, it does little on its own to help guide how many trials are needed for a given experimental situation.

A number of studies have used the sequential estimation technique (SET, Bates et al. 1983) to estimate the number of trials necessary to provide a stable estimate of a key performance variable. These have covered a range of movements including running (Bates et al. 1983), walking (Hamill & McNiven 1990), vertical jumping (Rodano & Squadrone 2002), landing (James et al. 2007), continuous jumping (Racic et al. 2009) and over-arm throwing (Taylor et al. 2015). In brief, SET requires a reference number of trials and standard deviation threshold to be selected. The number of trials required is then defined as where the cumulative mean first falls within this reference mean ± standard deviation.

*Email: s.forrester@lboro.ac.uk

© 2015 The Author(s). Published by Taylor & Francis.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
The estimated mean is expressed non-dimensionally as \( (m_N - \mu)/s_N \), where \( m_N \) is the mean of the \( N \) trials, \( \mu \) is the true mean and \( s_N \) is the standard deviation of the \( N \) trials. The estimated mean is given by the solid black line and the 95% confidence intervals by the dashed black lines. The light-shaded region represents the range of trials used in previous studies investigating breast kinematics during treadmill running while the darker shaded region represents the range of trials recommended to give a stable estimate of the mean as determined by SET analysis over a range of movements.

threshold \( \times \) reference standard deviation for all subsequent cumulative mean values. Typically, 20 reference trials (range 10–25) and a standard deviation threshold of 0.25 (range 0.25–0.30) have been employed resulting in a recommendation for 8–12 trials. However, little rationale is provided for the selection of reference number of trials and standard deviation threshold; thus, although the SET method provides a means for objectively selecting the number of trials, it does so using arbitrary criteria. Furthermore, the similarity in the outcomes of this analysis across studies is somewhat unsurprising. Based on the criterion given above and provided the input experimental data have a fixed distribution (i.e. an invariant mean and standard deviation), then it can be expected that the outcome will depend only on the reference number of trials and standard deviation threshold selected and not on the specific dataset inputted. Thus, the repeated use of the SET method for different movements appears unnecessary since the outcome can be determined without the need for an experimental assessment.

Fewer studies have used intra-class correlation coefficient (ICC) to estimate the number of trials necessary for stability. These have included landing (James et al. 2007), continuous jumping (Racic et al. 2009) and swimming (Connaboy et al. 2010). Although ICC may provide a more traditional means of determining stability, it also suffers from the arbitrary selection of what constitutes an acceptable reliability coefficient (James et al. 2007). Typically, ICC analysis has resulted in an estimated four trials for stability, suggesting it to be less conservative than the SET. Based on the \( t \)-statistic (Figure 1), this value is in the region where the confidence intervals on the mean are changing quite rapidly with number of trials and a similar observation has been made based on experimental data (Taylor et al. 2015). Hence, the ICC method holds the risk of under-estimating the number of trials required to achieve stability.

Perhaps of greater relevance is that the number of trials can impact the outcome of subsequent statistical analysis performed on the same data (Bates et al. 1992; Dufek et al. 1995). Fewer trials result in wider confidence intervals on the estimated means (Figure 1) and, therefore, a greater variance in the difference between means across different experimental conditions. This can impact the power (i.e. the probability of correctly rejecting the null hypothesis of equal population means) in subsequent inferential statistical testing. Despite the early work of Bates et al. (1992) and Dufek et al. (1995) and the opportunity to provide a less arbitrary outcome in the selection of an appropriate number of trials, the downstream statistical analysis of the (group) data rarely seems to be considered. This early work demonstrated the potentially marked effect of number of trials on the power of statistical testing; however, the focus was on interpretation of experimental results (type-II errors, i.e. ensuring adequate statistical power when accepting the null hypothesis) rather than as a tool to support the experimental design process. Furthermore, although various statistical tests were considered a focus on the simplest inferential statistical test, i.e. paired \( t \)-test, may have provided more tangible results from an experimental design perspective. Statistical analysis requires the selection of a target power and significance level; however, these probability variables are arguably more constrained and less arbitrary than those required for SET and ICC.

The overall aim of this study was to evaluate both SET and statistical power as a basis for selecting the number of trials to use in experimental biomechanics studies. For the SET method, the influence of the input data distribution on the number of trials needed to obtain a stable estimate of the mean was investigated. While for paired samples \( t \)-tests, the interaction between the number of subjects and number of trials necessary to achieve an acceptable level of statistical power was investigated. This aim was captured in the following objectives: (1) to confirm that for fixed input data distributions the outcome of the SET analysis is dependent only on the reference number of trials and standard deviation threshold employed; (2) to determine the effect of number of trials

![Figure 1. The effect of number of trials (N) on the 95% confidence intervals of the estimated mean obtained using the t-statistic. The estimated mean is expressed non-dimensionally as \( (m_N - \mu)/s_N \), where \( m_N \) is the mean of the \( N \) trials, \( \mu \) is the true mean and \( s_N \) is the standard deviation of the \( N \) trials. The estimated mean is given by the solid black line and the 95% confidence intervals by the dashed black lines. The light-shaded region represents the range of trials used in previous studies investigating breast kinematics during treadmill running while the darker shaded region represents the range of trials recommended to give a stable estimate of the mean as determined by SET analysis over a range of movements.](image)
on the statistical power for paired $t$-tests using data-sets typical of an experimental biomechanics study; (3) to explore the interaction between the number of subjects and number of trials in terms of achieving an acceptable statistical power for paired $t$-tests. These objectives were addressed using computer simulations; however, a typical biomechanics experimental data-set was included to provide relevant inputs for these simulations.

**Methods**

**Experimental data**

A typical experimental data-set, obtained during a separate study in our laboratories, was included as a comparison to the simulation data. For the SET method this data-set, in combination with experimental results from the literature, provided a comparison to the simulation outputs. For the paired $t$-test simulations, the experimental data were used to guide the simulation inputs as well as evaluation of the outputs.

The experimental data comprised of three-dimensional breast displacement ranges, expressed in the local (torso) co-ordinate system, during treadmill running ($2.78 \text{ m s}^{-1}$) in two different sports bras. The local (torso) co-ordinate system was defined with the origin at the suprasternal notch marker, the $z$-axis (superoinferior, SI) passed through the mid-point of markers positioned on the anteroinferior aspect of the right and left 10th ribs and the suprasternal notch marker, the $x$-axis (antero-posterior, AP) was perpendicular to the plane formed by these three markers and the $y$-axis (mediolateral, ML) was perpendicular to the $x$ and $z$-axes (Milligan et al. 2014). Breast motion was obtained from a marker positioned on the sports bra over the right nipple (Scurr et al. 2009), while gait cycles were identified from a marker positioned on the right heel (Zeni et al. 2008). A total of ten subjects (bra size 36DD) completed the within subjects experiment and a total of 50 gait cycles were captured for each subject and sports bra. The experimental data were analysed to give the breast range of motion in each co-ordinate direction in each gait cycle.

**SET model simulations**

To address Objective 1 a computer model was developed to predict the outcomes of the SET analysis for establishing the number of trials required to achieve a stable mean. A range of input distributions were investigated including two with a fixed mean and standard deviation (normal distribution and gamma distribution) and one with a variable mean (normal distribution with a linear drift in the mean; Table 1). In each case, the sample was submitted to the SET analysis to find the number of trials ($N_{STABLE}$) where the normalised cumulative mean and all subsequent normalised cumulative means fell within the specified threshold:

$$\frac{|m_{N\rightarrow REF} - m_{REF}|}{s_{REF}} \leq \text{THRES}$$

where $m_{N\rightarrow REF}$ are all the cumulative means from $N$ to $N_{REF}$ trials, $m_{REF}$ is the mean of the $N_{REF}$ reference trials, $s_{REF}$ is the standard deviation of the $N_{REF}$ trials and THRES is the threshold value. The simulations were carried out for $N_{REF}$ between 3 and 50 and THRES between 0.1 and 0.5. Each sample involved 100,000 computer generated data-sets from which the mean and 95% confidence intervals for $N_{STABLE}$ were evaluated (Taylor et al. 2015). The simulation model was validated by using threshold values of 0 (expect $N_{STABLE} = N_{REF}$) and 2 (expect $N_{STABLE} \rightarrow 1$). The simulation outputs were compared across the different input distributions as well as with the experimental data detailed above and from the literature.

**Paired $t$-test power analysis**

To address Objectives 2 and 3 a computer model was developed to predict the statistical power of paired

| Table 1. Input data distributions used for the SET analysis simulations (all abbreviations are as defined in Appendix 1). |
| --- |
| Distribution | Mean, $m$ | SD, $s$ | Shape parameter, $a$ | Scale parameter, $B$ | Drift parameter, $k$ |
| Normal: $m + s \times \text{randn}(N_{REF}, 1)$ | 0 | 1 | | | |
| 0 | 1 | 2 | | | |
| 1 | 1 | 2 | | | |
| 2 | 1 | | | |
| Gamma (skewed): $\text{gammld}(a, B, N_{REF}, 1)$ | 2 | 3 | 1 | | |
| Normal with baseline drift: $m + s \times \text{randn}(N_{REF}, 1)$ | $s \times (i - 1)$ for $i = 1, \ldots, N_{REF}$ | 1 | | | ± 0.01 |
| Drift parameter, $k$ | | | 0.5 | | ± 0.02 |

Note: Random samples for each of the distributions were defined in MATLAB according to the built-in functions listed above, where $N_{REF}$ is the reference number of trials.
samples $t$-tests. In paired samples $t$-tests, the statistical power depends on: (1) type-I error probability, i.e. the probability of falsely rejecting the null hypothesis ($\alpha$, fixed at 0.05 throughout); (2) number of subjects ($N_2$); (3) mean of the difference between conditions ($m_{DIFF}$); and (4) standard deviation of the difference between conditions ($s_{DIFF}$). The final two can be combined to provide the dimensionless Cohen’s $d_z$ effect size ($ES$, Lakens 2013):

$$ES = \frac{m_{DIFF}}{s_{DIFF}}$$

(2)

The value for $s_{DIFF}$ is related to the standard deviation in the means of each condition ($s_1$ and $s_2$) and the correlation coefficient between conditions ($r$) according to (Cohen 1988):

$$s_{DIFF} = \sqrt{s_1^2 + s_2^2 - 2 \times r \times s_1 \times s_2}$$

(3)

The most commonly used benchmark values for effect size are small ($ES = 0.20$), medium ($ES = 0.50$) and large ($ES = 0.80$; Cohen 1988). Fewer trials will increase the variance in the estimated means for the individual subjects (Figure 1). While the SET method was based on using arbitrary criteria to define when this variance reached an acceptable level, the paired $t$-tests method allows the effects of this increased variance on the statistical power to be quantified. In general, this increased variance will increase the standard deviation in the difference between conditions ($s_{DIFF}$), through an increase in $s_1$ and $s_2$ and a decrease in $r$ and, therefore, decreases the effect size ($ES$) and statistical power (Figure 2, Equations (2) and (3)).

For Objective 2, the input values for the simulations are presented in Table 2 and were based on data from the experimental biomechanics study detailed above. For Objective 3, the interaction between the number of subjects and number of trials was investigated through additional simulations in which the intra-subject and inter-subject variabilities were systematically changed from their original values. For both Objectives samples were drawn from each of the two conditions and inputted to a paired samples $t$-test for testing the null hypothesis that the conditions have equal means. All analyses were two-tailed, a $p$-value $\leq 0.05$ was considered statistically significant and a power of 0.80 was assumed adequate (as typically used in experimental biomechanics studies for estimating sample size, e.g. McLean et al. 2004; Milner et al. 2006). Each case was repeated 10,000 times from which the statistical power was determined. All analyses were conducted in MATLAB (R2010b, The MathWorks, Inc., Natick, MA, USA).

Results

SET simulations

The SET simulations results confirmed that the number of trials required to obtain a stable estimate of the mean was independent of the input data-set provided this data had a fixed distribution (Figure 3). The outcome was dependent only on the number of reference trials and standard deviation threshold. This was evidenced through the identical simulation results for all four fixed normal and three gamma input data distributions in terms of the mean and confidence intervals for the number of trials to achieve stability (Figure 3, thick solid lines). In contrast, when the input data-set included a drift in the baseline mean then the number of trials to achieve stability was no longer independent of the input data-set but also depended on the magnitude of drift (Figure 3, thick dashed lines). As the magnitude of drift increased then the number of trials to achieve stability also increased.

The SET simulation results demonstrated an overall trend for the number of trials to achieve stability to increase as the number of reference trials increased or the standard deviation threshold decreased (Figure 3(a)–(c)). For low standard deviation thresholds, the number of trials to achieve stability continued to increase even at 50 reference trials (Figure 3(a)); however, for high standard deviation thresholds this increase plateaued above ~20 reference trials (Figure 3(c)). For the extreme cases, as the standard deviation threshold tended towards zero then the number of trials to achieve stability tended towards the number of reference trials (Figure 3(d)) and as the standard deviation threshold became very large then the number of trials to achieve stability tended towards one (Figure 3(e)).

The SET results from previous studies obtained from the literature across a range of movements and for the experimental breast kinematics data-set detailed herein all fell within the 95% confidence intervals of the simulated results (Figure 3(a)–(c)). However, there was a tendency for some of the experimental results to exceed the simulation results for the fixed input data-set distribution perhaps indicating an underlying drift in the experimental data.
Effect of number of trials on the statistical power of paired \( t \)-tests

The initial paired \( t \)-test simulations, based on the experimental data, confirmed that the number of trials affected the statistical power of the test particularly for powers in the range 0.40–0.80 and small number of trials \((N_{TR} < 10; \text{Figure 4})\). In contrast, for very high or low statistical powers (<0.30 or >0.90) and \( N_{TR} > 20 \), then the number of trials had a negligible effect on the power. Simulation data are presented for 5, 10 and 20 subjects, representing the range typically used in experimental biomechanics studies, and for the three experimental effect sizes (Table 2). For the smallest effect size (AP direction, \( ES = 0.081 \) very small), a power of 0.80 was not achievable even with 20 subjects and 50 trials. For the middle effect size (ML direction, \( ES = 0.80 \) large) then 20 subjects and 3 trials would have been necessary to achieve a power of 0.80. While for the largest effect size (AP direction, \( ES = 1.45 \) very large) then 10 subjects and 3 trials were sufficient. These results also demonstrate the risk of committing a type-II error in failing to reject the null hypothesis for the middle effect size where 10 subjects only generated a statistical power of ~0.5–0.6 irrespective of number of trials.

Interaction between number of subjects and number of trials

The preceding results confirmed the relevance of quantifying the interaction between the number of subjects and number of trials in terms of achieving an acceptable statistical power. Furthermore, these results allowed the effect sizes (the two largest considered here, Table 2) and number of subjects (5–25) where this interaction appeared most relevant to be identified. Thus, the simulations for the third objective investigated the number of trials required to achieve a statistical power of 0.80 with a focus on these effect sizes and number of subjects with greater resolution than presented in Figure 4. In addition, the effects of inter- and intra-subject variability on the interaction were considered. Consequently, these results may have greater relevance over a broader range of experimental conditions.

Contour plots of statistical power (≥0.80) as a function of number of subjects and number of trials enabled the interaction regions, where power was most affected by the number of trials, to be identified, i.e. the lower left-hand corner of the shaded regions in Figure 5. As the numbers of trials decreased towards the minimum of 3 power decreased, in some cases to the point where it

---

Table 2. Inputs for the simulated paired \( t \)-tests and a summary of the experimental data on which they were based (all abbreviations are as defined in Appendix 1).

| Dependent variable | Simulations | Experimental data |
|--------------------|-------------|-------------------|
|                    | Breast displacement ranges during treadmill running | Sports bra design (×2) |
|                    | 5, 10, 20   | 10                |
| No. of subjects    | 3, 5, 10, 15, 20, 30, 40, 50 | 50                |
| No. of data-sets   | 10,000      | 1                 |
| Direction          | AP          | ML                |
|                    | SI          | SI                |

| Direction          | AP          | ML    | SI |
|--------------------|-------------|-------|----|
|                   | 20.9        | 17.5  | 32.6 |
| Sports Bra 1 group results: Group mean, \( m_1 \) | 0 | 0 | 0 | 20.9 | 17.5 | 32.6 |
| Group SD, \( s_1 \) | 1 (0.75, 1.25) | 1 (0.75, 1.25) | 1 (0.75, 1.25) | 7.2 | 5.2 | 8.0 |
| Individual SD, \( s_{IND} \) | 0.5 (0.25, 0.75) | 0.5 (0.25, 0.75) | 0.5 (0.25, 0.75) | 3.4 | 3.0 | 5.0 |
| Sports Bra 2 group results: Group mean, \( m_2 \) | 0.916 | 0.506 | 0.051 | 28.0 | 20.4 | 33.0 |
| Group SD, \( s_2 \) | \( s_1 \) | \( s_1 \) | \( s_1 \) | 7.5 | 7.8 | 8.1 |
| Individual SD, \( s_{IND} \) | \( s_{IND} \) | \( s_{IND} \) | \( s_{IND} \) | 4.0 | 3.1 | 4.8 |
| Bra 1 vs. Bra 2 group results: Correlation coeffi, \( r \) | 0.8 | 0.8 | 0.8 | 0.78 | 0.92 | 0.76 |
| Difference mean, \( m_{DIFF} \) | 0.916 | 0.506 | 0.051 | 7.0 | 2.9 | 0.4 |
| Difference SD, \( s_{DIFF} \) | 0.632 (0.474, 0.791) | 0.632 (0.474, 0.791) | 0.632 (0.474, 0.791) | 4.9 | 3.6 | 5.5 |
| Cohen’s \( d \) Effect Size, \( ES \) | 1.45 (1.93, 1.16) | 0.80 (1.07, 0.64) | 0.081 (0.11, 0.66) | 1.45 | 0.80 | 0.08 |

Notes: To define the simulation inputs, initially \( s_1 \) and \( s_2 \) were set to 1, Equation (3) was then used to give \( s_{DIFF} \) which was combined with the experimental effect sizes to give \( m_2 - m_1 \). The value of \( s_{IND} \) was set then set to 0.5 based on the experimental values relative to \( s_1 \) and \( s_2 \). Simulation inputs given in parentheses were only used within the sensitivity analysis part of Objective 3. The rows below the dashed line represent variables that were not directly inputted to the model but were calculated from those that were directly inputted. To check the validity of the computer model simulations was also ran for \( m_1 = m_2 = 0 \), i.e. the null hypothesis holds. The experimental mean and standard deviation data are all in mm.
dropped below 0.80 leading to the rounding off of the corner. Increasing inter-subject variability ($s_1$, $s_2$) principally increased the minimum number of subjects required to achieve a statistical power of 0.80, i.e. moved the shaded region upwards (Figure 5(a)–(c)) but had little effect on the characteristics of the interaction region. This increase in the minimum number of subjects resulted from an increased $s_{DIFF}$, decreased ES and decreased power (Figure 2, Equations (2) and (3)). In contrast, increasing intra-subject variability ($s_{IND}$) principally increased the extent of the interaction region, i.e. increased the rounding in the corner of the shaded region.
but had little effect on the minimum number of subjects required to achieve a statistical power of 0.80. This increase in curvature resulted from a decreased correlation coefficient relating the two conditions ($r$) which increased $s_{\text{DIFF}}$, decreased $ES$ and decreased power (Figure 2, Equations (2) and (3)).

These results suggest that generally for higher inter-subject variability more subjects will be required to achieve the required statistical power, while for higher intra-subject variability then simply more trials may be sufficient to achieve the required power. These observations are reinforced by looking at the minimum number of trials needed to achieve a statistical power of 0.80 as a function of number of subjects (Figure 6). The predominant right shift in the curves as inter-subject variability increased (Figure 6(a) and (b)) and increasing extent of curvature as intra-subject variability increased (Figure 6(c) and (d)) are again evident. For example, to achieve a power of 0.80 in the AP direction (Figure 6(a) and (c)) based on the original simulation conditions required at least seven subjects and six trials. Increasing inter-subject variability increased the number of subjects to nine, while the number of trials remained at six. Increasing the intra-subject variability did not affect the number of subjects ($N_S = 7$), but the number of trials increased to 11.

**Discussion**

Experimental biomechanics studies often involve the comparison of the mean values from multiple subjects across two or more experimental conditions. However, studies typically provide little rationale for the number of trials selected, despite the potential importance of this variable, and neglect existing methods proposed for estimating an appropriate value. This study aimed to evaluate two of these existing methods for selecting the number of trials in order to provide recommendations for future experimental biomechanics investigations. In particular, the SET method was investigated in terms of the influence of the input data distribution on the outcome. While paired samples $t$-tests were investigated in terms of the interaction between the number of subjects and number of trials necessary to achieve an acceptable level of statistical power. This was captured in the following objectives: (1) to confirm that for fixed input data distributions the outcome of the SET analysis is dependent only on the reference number of trials and standard
deviation threshold employed; (2) to determine the effect of number of trials on the statistical power of paired t-tests for data-sets typical of an experimental biomechanics study; (3) to explore the interaction between the number of subjects and number of trials in terms of achieving an acceptable statistical power.

Regarding the first objective, the SET simulations confirmed that provided the input data have a fixed distribution (normal or skewed) the outcome was independent of the variables describing this distribution. Thus, the number of trials to achieve stability for a given reference number of trials and standard deviation threshold can be obtained directly from this simulation model for any movement without the need to invest the time and resource in collecting experimental data. The currently used criteria of 20 reference trials and a threshold of 0.25 resulted in $9 \pm 8$ trials (mean ± 95% confidence intervals; Figure 3) to achieve a stable estimate of the mean which encompassed all the results from the experimental studies (Bates et al. 1983; Hamill & McNiven 1990; James et al. 2007; Racic et al. 2009; Taylor et al. 2015). Interestingly, some of the experimental SET results tended to lie above the mean simulation line particularly as the reference number of trials increased. This is worthy of further investigation and, based on a similar trend in the simulated data that included a drift in the baseline mean, may be indicative of a drift in the human performance data, perhaps due to warm up, fatigue, or some other process. Notably, the presence of any drift in the human performance data questions the validity of the SET approach since a truly stable performance may not occur in reality. A similar observation has been reported for ground reaction force variables during overground running where only 29–57% of the cumulative means were found to demonstrate convergence (Oriwol et al. 2012). Oriwol et al. (2012) found little evidence supporting a systematic pattern of convergence or divergence and suggested that one of the contributors may be the presence of time-dependent variations in the biological signals. Thus, further study is needed to investigate more broadly whether or not human performance variables tend converge to a stable performance and, if not, the mechanisms contributing to the divergence.

SET necessarily requires the arbitrary selection of the number of reference trials and standard deviation threshold. This threatens the validity of SET in terms of ensuring sufficient trials are included to achieve an acceptable level of statistical power in subsequent inferential analysis of the data. Strictly, the full statistical analysis needs to be considered and, by doing so, may lead to a more

Figure 6. The simulation results for the minimum number of trials ($N_{TR}$) necessary to achieve a statistical power of 0.80 versus number of subjects ($N_S$) for the larger two effect sizes represented in the experimental data (anteroposterior and mediolateral directions in Table 2). Subplots (a)–(b) represent the results for different inter-subject variability: (a) anteroposterior, (b) mediolateral; while subplots (c)–(d) represent the results for different intra-subject variability: (c) anteroposterior, (d) mediolateral. These curves serve to quantify the number of subjects region where the number of trials is a necessary consideration in order to achieve the required statistical power. For higher inter-subject variability (or a smaller effect size) more subjects are required to achieve the required statistical power (the curves are shifted to the right), while for higher intra-subject variability then simply more trials may be sufficient to achieve the required power (range of curvature increases). The remaining inputs are as given in Table 2.
robust and conclusive outcome on the required number of trials as illustrated, in this study, for paired t-test comparisons.

Regarding the second objective, the results confirmed that the number of trials can have a marked influence on the statistical power of paired t-tests under certain experimental conditions, in agreement with Bates et al. (1992) and Dufek et al. (1995). Reducing the number of trials increased the variance in the means for the individual subjects (according to Figure 1). This, in turn, increased the variance in the difference between conditions ($s_{DIFF}$) leading to a reduced effect size and statistical power (Figure 2, Equations (2) and (3)). For example, for the effect size representing the ML direction experimental data ($ES = 0.80$ large), the results indicated that the current experimental design lacked the statistical power to identify a significant difference between bras. Even using 50 trials the power did not exceed 0.60 for the 10 subject experimental design and more subjects would have been required to avoid a type-II error (incorrectly concluding no significant difference between bras). Given the relative ease of collecting 20+ trials compared to recruiting extra subjects for this type of experiment, then the optimal solution may have been to select the number of subjects that allowed the power to be achieved in 10–20 trials, e.g. 16 subjects and 13 trials (Figure 6). However, in other experimental situations 10–20 trials may not be feasible due to biological constraints (e.g. fatigue); and the optimal solution would have been to increase the number of subjects until a feasible number of trials was reached, in this example 20+ subjects and 3 trials.

Regarding the third objective, the simulations allowed the range of experimental conditions where the number of trials is most relevant to be identified. Statistical enquiry can be used to determine the minimum number of subjects necessary to achieve a power of 0.80 for a given effect size. Starting at this minimum, then a region exists where the number of trials needed to achieve this power is important, until the number of subjects is reached where only three trials are needed. In practical terms, this region represents the range of experimental conditions where increasing the number of trials can be considered a realistic alternative to achieve a statistical power of 0.80, particularly where access to subjects is limited. The size of this region appears large enough under typical experimental biomechanics conditions to require consideration. In addition to the values from the data-set considered here, the sensitivity of this region to both intra- and inter-subject variability was explored with the intention of demonstrating the relevance over a wider range of realistic experimental conditions. Increasing inter-subject variability had negligible effect on the extent of this region, its primary effect was to reduce statistical power (through an increased $s_{DIFF}$ and decreased $ES$) resulting in the fundamental requirement for more $ES$) resulting in the extent of the region, i.e. for a given number of subjects more trials were needed to achieve the required power in order to account for the reduced correlation coefficient between conditions.

A limited comparison between the current results and those presented by Bates et al. (1992) is possible. Bates et al. (1992) found that 6, 2 and 1 trials were needed to achieve a statistical power of 0.80 for 5, 10 and 20 subjects, respectively. This was for an effect size of 1 (the adjacent comparisons condition); however, the magnitude of intra-subject variability was not provided. These numbers are close to the 6, 3 and 3 trials needed to achieve a statistical power of 0.80 for 7, 10 and 20 subjects, respectively, for the largest effect size represented in the experimental data ($ES = 1.45$; AP direction in Figure 6). Given potential differences in the simulation inputs and minimum number of trials considered, these results appear to be in reasonable agreement.

The paired t-test simulations can be used to further evaluate the validity of SET. The paired t-test results, based on the core experimental data, indicated that the interaction between number of trials versus number of subjects to achieve a statistical power of 0.80 gave a minimum number of trials requirement of 5–20 (Figure 6 mid-grey lines); further increases in the number of trials (up to 50) did not allow the power to be achieved with fewer subjects. This is similar to the output of 9 ± 8 trials from SET based on the typically used convergence criteria. However, the paired t-test results also indicated that the extent of this interaction was highly sensitive to intra-subject variability; if intra-subject variability was halved then the interaction almost completely disappeared (Figure 6(c) and (d) black lines), while if intra-subject variability was doubled then the extent of the interaction increased to approximately 10–35 trials. This highlights a further limitation of SET; since it uses a standard deviation-based threshold within the convergence criteria (Equation (1)) the outcome is insensitive to changes in intra-subject variability; despite this variable having a marked effect on statistical power.

The paired t-tests simulation model employed a number of assumptions, principally normal distributions of the individual data-sets and subject means, a constant correlation coefficient ($r$) and a constant intra-subject variability ($s_{IND}$) for all subjects. These assumptions appeared reasonable based on the current experimental data-set (e.g. using the Kolmogorov–Smirnov test for normality, the null hypothesis could not be rejected for 67% of the data-sets), although Dufek et al. (1995) argued for a non-constant intra-subject variability model based on experimental observations. Furthermore, the universal appropriateness of group analyses in experimental
biomechanics has also been raised with the observation that different individuals may respond differently to an intervention, i.e. the concept of performer strategies (Dufek et al. 1995). A group analysis under such circumstances may fail to detect significant differences due to low statistical power (resulting from high group variance in Equations (2) and (3)); while significant differences may be present at the individual performer level. However, the assumptions applied are unlikely to have influenced the key outcome that under certain experimental situations the number of trials is a relevant experimental design factor to consider. Although the current framework has been demonstrated for a paired comparisons it could equally be applied to other statistical tests, e.g. repeated measures ANOVA, two-sample t-test and non-normal distributions.

Conclusion

The SET simulations confirmed that provided the input data has a fixed distribution then the number of trials to achieve a stable estimate of the mean is independent of this distribution and can be determined directly from simulation data. However, arbitrary convergence criteria combined with the neglect of intra-subject variability suggest that this method lacks validity and should be applied with extreme caution. The paired t-test simulations confirmed that both number of subjects and number of trials can have a marked effect on statistical power. Indeed, the number of subjects predicted by traditional sample size calculations does not necessarily guarantee that the required power is achieved. Under such circumstances the method presented herein, which accounts for the number of trials and provides an objective means for determining an appropriate combination of number of subjects and trials, can provide a better solution.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Bates BT, Dufek JS, Davis HP. 1992. The effect of trial size on statistical power. Med Sci Sports Exerc. 24:1059–1065.
Bates BT, Osternig LR, Sawhill JA, James SL. 1983. An assessment of subject variability, subject-shoe interaction, and the evaluation of running shoes using ground reaction force data. J Biomech. 16:181–191.
Cohen J. 1988. Statistical power analysis for the behavioural sciences. 2nd ed. New York (NY): Routledge Academic.
Connaboy C, Coleman S, Moir G, Sanders R. 2010. Measures of reliability in the kinematics of maximal undulatory underwater swimming. Med Sci Sports and Exerc. 42: 762–770.
Dufek JS, Bates BT, Davis HP. 1995. The effect of trial size and variability on statistical power. Med Sci Sports Exerc. 27:288–295.
Foissac M, Millet GY, Geysant A, Freychat P, Belli A. 2009. Characterization of the mechanical properties of backpacks and their influence on the energetics of walking. J Biomech. 42:125–130.
Hamill J, McNiven SL. 1990. Reliability of selected ground reaction force parameters during walking. Hum Mov Sci. 9:117–131.
James CR, Herman JA, Dufek JS, Bates BT. 2007. Number of trials necessary to achieve performance stability of selected ground reaction force variables during landing. J Sports Sci Med. 6:126–134.
Lakens D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol. 4:863.
McGhee DE, Steele JR, Zealey WJ, Takacs GJ. 2013. Bra-breast forces generated in women with large breasts while standing and during treadmill running: Implications for sports bra design. Appl Ergon. 44:112–118.
Mclean SG, Lipter SW, Van Den Bogert AJ. 2004. Effect of gender and defensive opponent on the biomechanics of sidestep cutting. Med Sci Sports Exerc. 36:1008–1016.
Milligan A, Mills C, Scurr J. 2014. Within-participant variance in multiplanar breast kinematics during 5 km treadmill running. J Appl Biomech. 30:244–249.
Milner CE, Ferber R, Pollard CD, Hamill J, Davis IS. 2006. biomechanical factors associated with tibial stress fracture in female runners. Med Sci Sports Exerc. 38:323–328.
Oriwol D, Milani TL, Maiwald C. 2012. Methodological issues associated with the mean value of repeated laboratory running measurements. Footwear Sci. 4:183–190.
Racic V, Pavic A, Brownjohn JM. 2009. Number of successive cycles necessary to achieve stability of selected ground reaction force variables during continuous jumping. J Sports Sci Med. 8:639–647.
Rodano R, Squadrone R. 2002. Stability of selected lower limb joint kinetic parameters during vertical jump. J Appl Biomech. 18:83–89.
Salo A, Grimshaw PN, Viitasalo JT. 1997. Reliability of variables in the kinematic analysis of sprint hurdles. Med Sci Sports Exerc. 29:383–389.
Scurr J, White J, Hedger W. 2009. Breast displacement in three dimensions during walking and running gait cycles. J Appl Biomech. 25:322–329.
Taylor PG, Lee K-Y, Landeo R, O’Meara DM, Millett E. 2015. Determining optimal trial size using sequential analysis. J Sports Sci. 33:300–308.
Triola MF. 2004. Elementary statistics using the graphing calculator: for the TI-83/84 plus. 1st ed. Boston (MA): Pearson.
Unick J, Kieffer H, Cheesman W, Feeney A. 2005. The acute effects of static and ballistic stretching on vertical jump performance in trained women. J Strength Cond Res. 19:206–212.
Walsh M, Böhm H, Butterfield M, Santhosam J. 2007. Gender bias in the effects of arms and countermovement on jumping performance. J Strength Cond Res. 21:362–366.
White J, Scurr J, Hedger W. 2011. A comparison of three-dimensional breast displacement and breast comfort during overground and treadmill running. J Appl Biomech. 27:47–53.
Zeni JA, Richards JG, Higginson JS. 2008. Two simple methods for determining gait events during treadmill and overground walking using kinematic data. Gait Posture. 27:710–714.
Zhou J, Yu W, Ng S-P. 2012. Studies of three-dimensional trajectories of breast movement for better bra design. Text Res J. 82:242–254.
Appendix 1. List of nomenclature

| Symbol | Definition |
|--------|------------|
| AP     | anteroposterior direction |
| $a$    | shape parameter for the gamma distribution |
| $B$    | scale parameter for the gamma distribution |
| $ES$   | Cohen’s $d$, effect size |
| $ICC$  | intra-class correlation coefficient |
| $k$    | drift parameter for the normal distribution with baseline drift |
| $ML$   | mediolateral direction |
| $N$    | number of trials |
| $r$    | correlation coefficient |
| $SD$   | standard deviation |
| $SET$  | sequential estimation technique |
| $SI$   | superoinferior direction |
| $THRES$| threshold |
| $m$    | mean |
| $s$    | standard deviation |

Greek symbols

| Symbol | Definition |
|--------|------------|
| $\alpha$ | significance level (Type-I error probability) |
| $\mu$  | true mean |

Subscripts

| Symbol | Definition |
|--------|------------|
| $DIFF$ | difference between conditions |
| $IND$  | individual subject |
| $N$    | number |
| $REF$  | reference |
| $S$    | subject |
| $STABLE$ | stability |
| $TR$   | trial |
| $1, 2$ | condition 1, condition 2 |