Statistical Analysis and Results

# 1. Statistical Analysis

All code utilised for data preparation and analyses are available in either the Open Science Framework page for this project <https://osf.io/zrxjp/> or the corresponding GitHub repository <https://github.com/jamessteeleii/isokinetic_isometric_agreement_reliability>. We cite all software and packages used in the analysis pipeline using the grateful package (Rodriguez-Sanchez et al., 2023) which can be seen here: <https://osf.io/pgx6v>. This project was not pre-registered, but had an exploratory estimation goal. All analyses have been conducted within a Bayesian posterior estimation framework and all posterior estimates and their precision, along with conclusions based upon them, will be interpreted continuously and probabilistically, considering priors, data quality, and all within the context of each outcome and the assumptions of the model employed as the estimator (Kruschke & Liddell, 2018). Given that most utilisation of strength measurement is for the purpose of determining whether a change has occurred from test-to-test, we focused on the *agreement* of measurements as opposed to the typical *reliability* statistics which instead reflect the ability of measurements to distinguish between individuals (Berchtold, 2016; Kottner & Streiner, 2011; Vet et al., 2006). However, we report on the reliability in the form of variance decomposition ratios for our Bayesian models calculated directly from the posterior predictve distributions which are comparable to intraclass correlation coefficients (ICC) in order to compare to prior research. Here we also employed multivariate mixed effects methods for examining agreement by variance components enabling us to model all three exercises, chest press, leg press, and row, simultaneously extending previous approaches (Schluter, 2009) thus offering greater precision, robustness, and efficiency of estimates. Two sets of models, each detailed below, were used to examine both the between- and within-day agreement for isometric outcomes, and for between-day agreement for isokinetic outcomes. In each model we employed informative yet weakly regularising priors which are detailed below. All models were fit with four Markov Chain Mote Carlo chains using 2000 warmup and 6000 sampling iterations. Trace plots were produced along with values to examine whether chains had converged, and posterior predictive checks for each model were also examined to understand the model implied distributions. Note, all values are in Newtons of force.

## 1.1 Isometric Outcomes

Given that for isometric outcomes we had both three repeated days of testing, and two repeated trials within each day, we opted to adapt the methods described by Jones et al. (2011) and Christensen et al. (2020) to derive the limits of agreement with the mean. Typically where there are two measurements to compare in terms of agreement the traditional Bland-Altman Limits of Agreement approach can be employed (Bland & Altman, 1986). When there are multiple measurements (whether multiple methods, observers, or tests, or whether there are replicates within these) it is more difficult to apply these typical models. Instead, we can model the outcomes directly and derive the agreement with the *mean* value for the participant over the repeated measurements made. In the case where we can assume there is no bias for a particular measurement (in our case no particular bias across days for example), then we can assume that the mean reflects a good estimate of the true value and the 95% limits of agreement then reflect the range over which we would expect measurements to fall about the true value 95% of the time.

In the present case we sought to partition the variance components such that we could determine separately the between-participant variance (i.e., below), the between-day variance (i.e., below), and the residual variance which here reflects the within-day variance (i.e., below). We estimate these variance components through a multivariate mixed effects model of the joint three exercise outcomes observed (chest press, leg press, and row; see [Equation 1](#eq-isometric-model)). The model included a population (i.e., fixed) effect for day which was Helmert coded; This meant that for the three days we have two coefficients in the model for each outcome with the first, below, reflecting the difference between the mean of day one and the mean of day two and three, and the second, below, reflected the difference between mean of day two and the mean of day three. This allowed us to examine whether there was any systematic bias, and in this case the Helmert coding was specifically used because we anticipated that any bias would manifest in terms of a “familiarisation” effect whereby participants improved systematically with repeated measurement. The model also included random (i.e., group level terms) intercepts for participants (i.e., below), and random intercepts for day nested within participant (i.e., below). Each of these were modelled as correlated between outcomes reflecting the models assumption that typically participants that are stronger are stronger across each exercise tested (i.e., strength is correlated between exercises), and also that variation across days was likely to also be related reflecting that lower/higher values on a given exercise on one day would likely be related to lower/higher values on a different exercise on that day. Lastly, the residual errors were also modelled as correlated. The model for isometric measurements can thus be represented as follows in [Equation 1](#eq-isometric-model):

where each exercise outcome is represented by a superscript for observations and model parameters (i.e., ), and for a given exercise the subscripts reflect the th measurement (), from the th day () for the th participant (). Population and group (i.e., fixed and random) parameters are described above. The covariance matrices for observations, random intercepts for participant, and random intercepts for day within participant are given by , , .

As mentioned above we adopted informative yet weakly regularising priors. Default priors in the brms R package used to fit the model are weakly regularising on all intercept terms (i.e., ) and are set such that they are centred and scaled using a distribution with and represent the expected response value when all predictors are at their means, all group level terms are set with a distribution with , a , and scaled to the expected response values, and all correlation matrices are set with an distribution. The remaining population effect coefficients are by default set with an improper flat prior on the reals (i.e., ) and thus we opted to set our own informative weakly regularising priors based on the raw data descriptives. Typically a 10% coefficient of variation is deemed acceptable for strength measures test-retest variation (Nuzzo et al., 2019) and so we opted for a slightly more skeptical prior with a location set at 20% of the sample arithmetic mean of all observations for each exercise outcome. Further, we assumed a simple propagation of error approach for two independent measurements (i.e., ignoring covariance and thus reflecting a lack of knowledge about the exact nature of it) again utilising the sample variance of all observations for each exercise outcome. Thus, the priors for the model in [Equation 1](#eq-isometric-model) were:

The between- and within-day 95% limits of agreement with the mean where calculated from the posterior draws of the relevant variance components. These were calculated adapting the approach of Christensen et al. (2020) adjusting for the degrees of freedom based upon the number of days, and number of replicates within days. The between-day limits of agreement with the mean were calculated using the between day variance component for each exercise outcome, , as:

And the within-day limits of agreement with the mean utilising the remaining within day residual variance component for each exercise outcome, , as:

For each posterior draw the limits of agreement with the mean were calculated and then the mean and 95% quantile intervals (QI) determined. These were then presented graphically with a limits of agreement with the mean plot following the methods described by Jones et al. (2011) and Christensen et al. (2020) where the participant mean for each exercise outcome, , was plot on the x-axis and the difference between each observation for each exercise outcome with the mean, , plot on the y-axis with the corresponding limits of agreement with the mean for both between- and within-day plot about these. We also present the posterior distribution for the bias reflected by the and coefficients along with their corresponding mean and 95% quantile interval. The variance decomposition ratios were calculated for both between- and within-day by calculating the ratio between the variance for draws from the posterior predictive distribution not conditioned on random (i.e., group level terms) and the variance for draws conditioned on the appropriate random effects. The mean and 95% quantile interval for these were then calculated.

## 1.2 Isokinetic Outcomes

For the isokinetic outcomes we only had two repeated days of testing, and for each day a single “best” repetition measured for each exercise outcome and for both concentric and eccentric phases. We opted to model the concentric and eccentric phases separately as we suspected, whilst they may be correlated, the between-day agreement might differ for either. As such, given we only had two measurements between days for each exercise outcome and muscle action a traditional Bland-Altman Limits of Agreement approach could be employed (Bland & Altman, 1986). We did however still adopt a multivariate approach allowing for the residual errors were also modelled as correlated which in this case, given our model described below had no other predictors (i.e., they include an intercept only for each outcome) these are the correlations between the between-day differences in each outcome. Where for each exercise outcome in each muscle action is the difference between days for the th participant () i.e., , the model is:

where is the residual covariance matrix (omitted due to size).

Weakly regularising default priors were used again on all intercept terms (i.e., ) for each outcome and set such that they were centred and scaled using a distribution with representing the expected response value when all predictors are at their means which in this case meant the raw means. The residual correlation matrix was set with an distribution.

In this model the 95% limits of agreement are calculated simply as for each outcome. Again this was calculated for each outcome for each posterior draw and the mean and 95% quantile intervals calculated. The bias for each outcome then are the intercept terms and the corresponding means and 95% quantile intervals for these were also determined. These were plot together in a traditional Bland-Altman limits of agreement plot where the participant mean for each exercise outcome, , was plot on the x-axis and the difference between days for each exercise outcome, , plot on the y-axis with the corresponding limits of agreement and mean bias plot about these.

For the variance decomposition ratios for isokinetic outcomes a separate multivariate model was fit for for each exercise outcome and muscle action as follows:

where is the residual covariance matrix and the random intercept covariance matrix (both also omitted due to size).

Weakly regularising default priors were used again on all intercept terms (i.e., ) for each outcome and set such that they were centred and scaled using a distribution with representing the expected response value when all predictors are at their means which in this case meant the raw means. The residual correlation matrix was set with an distribution.

Variance decomposition ratios were then calculated for between-day by calculating the ratio between the variance for draws from the posterior predictive distribution not conditioned on random (i.e., group level terms) and the variance for draws conditioned on the appropriate random effects. The mean and 95% quantile interval for these were then calculated.

# 2. Results

## 2.1 Isometric Outcomes

The isometric model showed chain convergence with values all and posterior predictive checks were good. Model diagnostics can be seen in the supplementary materials here: <https://osf.io/ypt59>.

[Figure 1](#fig-isometric-plot) shows the mean bias and limits of agreement with the mean for both between- and within-day for each exercise. There was no clear evidence of a “familiarisation” biasing effect between days given that the sign of the contrasts both between- and within-exercises was variable and the posterior distributions typically all ranged from both small positive to negative effects. As might be expected, the between-day limits of agreement with the mean were greater than the within-day agreement. The between-day limits of agreement with the mean for the chest press were 46.19 [95%QI: 38.62, 56.12], for the leg press were 68.4 [95%QI: 56.57, 84.47], and for the row were 31.41 [95%QI: 25.75, 39.15]. The within-day limits of agreement with the mean for the chest press were 39.18 [95%QI: 31.4, 48.84], for the leg press were 52.67 [95%QI: 41.95, 67.25], and for the row were 22.64 [95%QI: 18, 28.84].

The variance decomposition ratio (comparable to the ICC) for the chest press between-day was 0.937 [95%QI: 0.892, 0.963] and within-day was 0.939 [95%QI: 0.895, 0.964]. For the leg press the variance decomposition ratio between-day was 0.967 [95%QI: 0.942, 0.981] and within-day was 0.968 [95%QI: 0.944, 0.981]. For the row the variance decomposition ratio between-day was 0.969 [95%QI: 0.945, 0.983] and within-day was 0.97 [95%QI: 0.947, 0.983].

|  |
| --- |
| Figure 1: Bias and limits of agreement with the mean for isometric outcomes for both between- and within-day. |

## 2.2 Isokinetic Outcomes

The isokinetic model also showed chain convergence with values all and posterior predictive checks were good. Model diagnostics can be seen in the supplementary materials here: <https://osf.io/wzh9m>.

[Figure 2](#fig-isokinetic-plot) shows the mean bias and limits of agreement between-day for each exercise and muscle action. There was no clear evidence of a “familiarisation” biasing effect between days for most exercises and muscle actions, perhaps with the exception of the eccentric leg press which showed somewhat of an improvement from day one to day two: 5.13 [95%QI: -19.45, 29.76]. The limits of agreement for concentric muscle actions for the chest press were 34.86 [95%QI: 24.26, 51.28], for the leg press were 100.85 [95%QI: 74.76, 137.59], and for the row were 24.29 [95%QI: 16.41, 36.49]. The limits of agreement for eccentric muscle actions for the chest press were 50.78 [95%QI: 35.17, 74.84], for the leg press were 162.4 [95%QI: 116.74, 230.53], and for the row were 38.73 [95%QI: 27, 56.76].

The variance decomposition ratio (comparable to the ICC) for concentric muscle actions for the chest press between-day was 0.954 [95%QI: 0.89, 0.984], for the leg press was 0.912 [95%QI: 0.797, 0.968], and for the row was 0.971 [95%QI: 0.927, 0.991]. For eccentric muscle actions the variance decomposition ratio for the chest press was 0.931 [95%QI: 0.828, 0.977], for the leg press was 0.855 [95%QI: 0.652, 0.95], and for the row was 0.937 [95%QI: 0.851, 0.978].

|  |
| --- |
| Figure 2: Between day bias and limits of agreement for isokinetic outcomes. |

# 3. References

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 2059799116672875. <https://doi.org/10.1177/2059799116672875>

Bland, J. M., & Altman, D. G. (1986). [Statistical methods for assessing agreement between two methods of clinical measurement](https://www.ncbi.nlm.nih.gov/pubmed/2868172). *Lancet (London, England)*, *1*(8476), 307–310.

Christensen, H. S., Borgbjerg, J., Børty, L., & Bøgsted, M. (2020). On Jones et al.’s method for extending Bland-Altman plots to limits of agreement with the mean for multiple observers. *BMC Medical Research Methodology*, *20*(1), 304. <https://doi.org/10.1186/s12874-020-01182-w>

Jones, M., Dobson, A., & O’Brian, S. (2011). A graphical method for assessing agreement with the mean between multiple observers using continuous measures. *International Journal of Epidemiology*, *40*(5), 1308–1313. <https://doi.org/10.1093/ije/dyr109>

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, *64*(6), 701–702. <https://doi.org/10.1016/j.jclinepi.2010.12.001>

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective [Journal Article]. *Psychonomic Bulletin & Review*, *25*, 178–206.

Nuzzo, J. L., Taylor, J. L., & Gandevia, S. C. (2019). CORP: Measurement of upper and lower limb muscle strength and voluntary activation. *Journal of Applied Physiology*, *126*(3), 513–543. <https://doi.org/10.1152/japplphysiol.00569.2018>

Rodriguez-Sanchez, F., cre, cph, Jackson, C. P., Hutchins, S. D., & Clawson, J. M. (2023). *Grateful: Facilitate Citation of R Packages*.

Schluter, P. J. (2009). A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies. *BMC Medical Research Methodology*, *9*(1), 6. <https://doi.org/10.1186/1471-2288-9-6>

Vet, H. C. W. de, Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. <https://doi.org/10.1016/j.jclinepi.2005.10.015>