

Position-specific allometric adjustment of NFL Combine tests: evaluation across year cohorts

Thomas Davin Fahey

discusdoc@aol.com

California State University

Article

Keywords: allometric scaling, American football, body mass, W-scores, cross-validation, performance assessment

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11176926/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

To determine whether position-specific body-mass adjustment improves interpretation of NFL pre-draft performance across tests and year cohorts.

Methods

We analyzed 2,333 records from 2020–2026 using separate eligibility for six tests and 47 position–test log–log models. W-scores expressed residuals in units of the training residual standard error. Leave-one-year-out evaluation compared mass-adjusted predictions with position-specific mean baselines. Sensitivity analyses excluded 2021; exploratory analyses included bootstrap intervals, fixed-exponent comparators, and a retrospective 2026 holdout trained on earlier years.

Results

Twenty-six models reduced held-out squared error, and 21 increased it. Pooled reductions were 26.0% for sprint, 9.7% for vertical jump, 16.6% for broad jump, 7.2% for bench press, 13.9% for shuttle, and 16.1% for three-cone. Without defensive linemen, gains fell to 4.6%, – 0.5%, approximately 0.0%, – 0.8%, and 4.3% for sprint, vertical jump, broad jump, shuttle, and three-cone, respectively. In 2026, mean W-scores were + 0.50, + 0.75, + 0.45, and + 0.52 for sprint, vertical jump, broad jump, and bench press, indicating positive cohort offsets. Prediction error increased for broad jump and bench press in that holdout. Bench press was secondary; its year-cluster intervals included zero, and fixed exponents outperformed the estimated exponent in pooled cross-validation.

Conclusion

Adjustment benefits varied by position and task and were concentrated among defensive linemen in body-transport tests. Later-cohort score offsets require calibration checks before adjusted scores supplement raw performance.

Introduction

NFL pre-draft physical testing assesses prospects' strength, endurance, jumping, sprinting, and change-of-direction performance. Interpretation requires attention to substantial differences in body size and physical demands across playing positions. Positional groups differ in both anthropometry and performance, so comparisons across the pooled prospect population may reflect positional specialization as well as individual physical capacity (Nuzzo 2015; Robbins and Young 2012). Raw test results remain relevant descriptions of what an athlete achieves. Body-mass-adjusted results address a

complementary question: how does that performance compare with the expectation for an athlete of similar mass within the same positional group? Neither interpretation, by itself, establishes football ability or subsequent professional success (Kuzmits and Adams 2008; Teramoto et al. 2016).

The relationship between body mass and performance need not be consistent across tests or positions. Bench-press repetitions at a fixed external load and tasks requiring acceleration or displacement of the body place different demands on an athlete. Furthermore, the range of body mass within a position can differ substantially from that across the entire prospect population. A single adjustment applied to all athletes may therefore obscure relationships that matter within particular position–test combinations. Previous football research has examined allometric strength scaling, body-mass normalization, and position-specific reference values, providing a foundation for empirically evaluating these relationships (Gillen 2023; Nuzzo 2015; Oba et al. 2014).

Simple ratio scaling, such as dividing a performance measure by body mass, imposes a specific relationship between the variables. When that relationship is inappropriate, the resulting score can retain or introduce size-related associations (Packard and Boardman 1988; Tanner 1949). Allometric modeling offers an alternative by estimating the exponent of a power relationship between performance and body mass (Jaric 2002; Nevill et al. 1992). In a log–log regression, the exponent quantifies the proportional association between mass and expected performance. Its magnitude and uncertainty should be evaluated for the population and task of interest rather than assumed universal.

Regression–residual expresses performance relative to a body-mass-specific expectation. Related applications in strength, powerlifting, and peak oxygen uptake provide methodological context but do not establish that adjustment transfers equally well across football positions and tasks (Fahey 2026a, 2026c; Fahey and Rivera 2026).

Evaluation beyond the estimation sample is essential. Within the estimation sample, ordinary least-squares residuals are mathematically orthogonal to the fitted predictor, log body mass. Consequently, a near-zero in-sample association is not independent evidence that the adjustment will generalize to other cohorts. It also does not establish independence from untransformed mass or rule out nonlinear dependence. Held-out assessment can instead examine whether scores remain appropriately centered and scaled, whether residual associations with mass persist, and whether incorporating mass improves prediction relative to a position-specific model without mass. Comparing these outcomes within each position and test avoids reliance on pooled results that may mask weaker performance in particular subgroups.

Incomplete test participation introduces an additional analytical issue. Requiring complete results across an entire test battery discards otherwise usable observations from athletes who completed only some tests. Separate-test analyses retain each valid body-mass–performance pair and provide a direct basis for interpreting the corresponding task. This approach does not eliminate selective participation, but it avoids making eligibility for one analysis depend on completing unrelated tests. Differences

between year cohorts also warrant evaluation, including sensitivity to the inclusion of records labeled 2021, whose testing comparability remains uncertain.

The study aimed to determine when position-specific mass–performance relationships improve prediction and score interpretation across tests and year cohorts relative to position-specific mean baselines. Each test used its own set of eligible observations. Primary analyses estimated the positional relationships and evaluated them by leaving out one year at a time; sensitivity analyses excluded 2021. Exploratory analyses examined uncertainty, fixed-exponent comparators, a temporally ordered 2026 holdout, residual diagnostics, and within-position mass spread. The contribution is the assessment of transfer across positions, tasks, and cohorts, including settings in which adjustment offers little benefit. Figure 1 summarizes the workflow.

Methods

Study design and data source

This retrospective observational study examined NFL pre-draft anthropometric and performance records from 2020 to 2026. Data were compiled from the publicly accessible Pro Football Reference NFL Combine Results archive (Sports Reference LLC; <https://www.pro-football-reference.com/draft/combine.htm>). Original retrieval dates were not retained. The reconciled source file contained 2,333 records: 324 from 2020, 439 from 2021, 315 from 2022, 306 from 2023, 315 from 2024, 322 from 2025, and 312 from 2026. The unit of analysis was a prospect record within a testing year; unique record identifiers were used throughout processing. Identity matching across years was not independently verified. Figure 1 summarizes the analytical workflow. The Zenodo project record was published on 22 September 2026 (Fahey 2026b). The cleaned analysis snapshot used here was prepared on 23 September 2026; this date identifies the snapshot, not the original website retrieval. All analyses used the same 2,333-record snapshot.

The study was approved by the Institutional Review Board at California State University, Chico (protocol IRB-2026-20, 13 February 2026). Athlete names were omitted from the analysis files and replaced with record identifiers. Record identifiers (NFL00001 onward) index rows in the compiled analytical file; they are not independently verified person identifiers. Therefore, the same person could appear in more than one year.

Position groups and test-specific eligibility

The positional classification in the compiled dataset was retained: defensive backs (DB; 454 records), defensive linemen (DL; 423), linebackers (LB; 228), offensive linemen (OL; 409), quarterbacks (QB; 110), running backs (RB; 214), tight ends (TE; 154), and wide receivers (WR; 341). Each performance test was analyzed separately within each position. Eligibility required a positive body mass and a valid positive result for that test. Availability of other performance tests, height, arm length, or hand size was not

required. Missing measurements were not imputed, and records with missing values for one test remained eligible for other tests.

The resulting mass–performance samples comprised 1,700 sprint, 1,665 vertical-jump, 1,568 broad-jump, 885 bench-press, 911 shuttle, and 854 three-cone observations. These samples overlapped and were not treated as independent groups. The single QB bench-press observation was retained in the descriptive table but could not be used for regression or validation. Body mass was available for 2,310 of the 2,333 source records. Table 1A presents each available anthropometric measurement separately; Table 1B presents performance for records meeting the mass–test eligibility criteria.

Measurement processing and performance variables

The six performance measures were the 40-yard sprint, vertical jump, broad jump, bench-press repetitions at a 225-lb load, and the shuttle and three-cone tests. Body mass, recorded in pounds, was converted to kilograms by multiplying by 0.45359237. Jump measurements and anthropometric lengths were converted from inches to centimeters by multiplying by 2.54. Timed results were reported in seconds for descriptive reporting.

The source contained both decimal measurements and encoded lengths. Four-digit height codes were decoded into feet, inches, and eighths of an inch; arm-length and hand-size codes were decoded into whole inches, with a fractional numerator and denominator. Encoded broad-jump values were decoded into feet followed by inches. Two noncanonical height codes, 6018 and 6058, were excluded from height summaries only; they did not affect test-specific eligibility because height was not a regression predictor. One noncanonical broad-jump code, 912, was excluded from the broad-jump analysis without excluding the record from other tests. The primary analyses retained these field-specific exclusions rather than assigning unverified corrected values.

For modeling, performance P was defined as the reciprocal of time ($1/t$, s^{-1}) for the sprint, shuttle, and three-cone tests; jump height or distance in centimeters for the jumping tests; and repetitions for bench press. Thus, higher P consistently indicated better performance. All logarithms were natural. Existing score columns in the source workbook were not used; model coefficients and W -scores were recalculated from the processed measurements. Bench press was retained as a secondary descriptive analysis because repetitions are discrete and the Gaussian log-residual model was not assumed to yield a calibrated count distribution.

Allometric models and W -scores

Separate ordinary least-squares regressions were fitted for each position–test combination, relating log performance to log body mass M :

$$\ln(P_i) = \alpha + b \ln(M_i) + \varepsilon_i$$

Here, α is the intercept, b is the allometric exponent, and ϵ_i is the residual. A full-sample model required at least three eligible observations and variation in body mass. Forty-seven of the 48 possible position–test combinations were estimable; QB bench press was the exception. Model summaries included sample size, intercept, exponent, a conventional two-sided 95% t confidence interval for the exponent, the coefficient of determination (R^2), and the residual standard error (SEE). The exponent interval used $n - 2$ degrees of freedom and the ordinary least-squares slope standard error. Intervals were not adjusted for multiple comparisons or clustered dependence.

Residual standard error and W-scores were calculated as:

$$SEE = \sqrt{[\sum \epsilon_i^2 / (n - 2)]}$$

$$W_i = [\ln(P_i) - \alpha - b \ln(M_i)] / SEE$$

W-scores were expressed in residual-standard-error units, without transformation to a mean of 50 and an SD of 10. A positive score indicates performance above the fitted, position-specific expectation at that mass. Full-sample equations are reported in Table 2, and exponents and confidence intervals are shown in Figure 2. In-sample residual orthogonality to log body mass was assessed as an implementation property, not as independent validation.

Leave-one-year-out validation

Validation was conducted separately for each position and test. One testing year was omitted, and the allometric equation and SEE were estimated using eligible observations from the remaining years. The fitted equation then predicted log performance for the omitted-year records, and their W-scores were calculated using the training SEE. This procedure was repeated for each available year. Omitted-year performance observations did not contribute to their own fitted coefficients or residual standard error.

A validation fold required at least 20 training observations and nonzero variation in body mass. This served as a pragmatic safeguard against very sparse training samples, not an established criterion for model adequacy. Folds that failed this rule were recorded and excluded from validation summaries. All 47 estimable models in the all-years analysis produced held-out predictions for every eligible observation. The bench-press validation sample therefore comprised 884 records, whereas its descriptive sample included 885.

For comparison, an intercept-only baseline predicted the mean log performance for the same training sample at the corresponding position and for the test. Held-out errors were defined as observed minus predicted log performance. Prediction accuracy was summarized by root mean squared error (RMSE) and by the percentage reduction in summed squared error relative to the baseline:

$$RMSE = \sqrt{[\sum e_i^2 / n]}$$

$$\Delta \text{ error (\%)} = 100 \times [1 - \sum e_i^2 / \sum e_0^2]$$

Here, n_i is the number of held-out predictions, e_i is the allometric-model prediction error, and e_0 is the corresponding baseline error. Positive percentages indicate a lower squared error than the baseline; negative values indicate worse predictive performance. The measure does not represent a percentage change in athletic performance.

We calculated the held-out W-score mean, sample SD, Pearson correlation with log body mass, and regression slope against log body mass. Metrics were summarized within year–position–test folds and across omitted years for each position–test combination. Additional summaries pooled predictions across positions within each test; pooled correlations were not treated as evidence of successful adjustment within each position. Correlations required at least three predictions and variation in both variables. Table 3 reports the principal validation metrics, and Figure 3 displays position-specific error reductions and residual mass correlations.

Calibration was described using body-mass quintiles within each position–test combination across its held-out records; tied boundaries could reduce the number of bins. Bins did not enter model fitting. The largest absolute quintile means are summarized in Supplementary Table S4. LOYO folds share training data, and some train on years later than the omitted year. Consequently, LOYO evaluates transfer across observed cohorts rather than prospective forecasting.

Sensitivity analysis excluding 2021

Individual testing conditions for records labeled 2021 could not be independently verified. To assess sensitivity to that cohort, all 439 2021 records were removed, leaving 1,894 records. Test-specific eligibility, position-specific model fitting, baseline estimation, and leave-one-year-out validation were repeated with the remaining data. Full-cohort coefficients and SEE values were not reused.

This analysis retained 1,321 sprint, 1,286 vertical-jump, 1,201 broad-jump, 563 bench-press, 586 shuttle, and 518 three-cone pairs. For RB three-cone, four folds with 20 omitted-year observations failed the minimum-training rule, leaving one eligible held-out prediction. Consequently, three-cone validation comprised 498 predictions. The single RB prediction was included in the overall error calculation, but its subgroup correlation and percentage error reduction were suppressed in Table 3 and Figure 3 because a single observation does not support a meaningful subgroup validation summary.

Exploratory comparisons and temporal holdout

Comparator and temporal-holdout analyses were exploratory. The free exponent was compared with an intercept-only model ($b = 0$), a ratio specification ($b = 1$ for P/M), and fixed benchmarks ($b = -1/3$ for reciprocal-time tests, $b = 0$ for jumps, and $b = 2/3$ for bench repetitions). These benchmarks were sensitivity choices, not established exponents for these tasks; the jump benchmark equals the no-mass

model. Each fixed- b model fitted its own intercept as $\text{mean}[\ln(P) - b \ln(M)]$ using exactly the same training records as the free model. Its residual SEE used $n - 1$ degrees of freedom; the free model used $n - 2$. Held-out error and residual–mass correlation were then evaluated identically. Using $b = 1$ on reciprocal time means dividing $1/t$ by mass, not dividing time by mass.

For a temporally ordered check, each position–test model was fitted using records from 2020–2025 and applied only to 2026. The 2021-exclusion variant removed that year from training. Both analyses used the same minimum-training rule. The 2026 holdout was specified after the full dataset had been examined and is therefore a retrospective temporal evaluation, not a prospectively specified external validation.

Uncertainty and diagnostic analyses

Two complementary percentile bootstrap analyses estimated 95% intervals for LOYO error reduction, RMSE, and residual–mass correlation. First, 2,000 record bootstrap samples resampled records with replacement within each position–year cell, preserving cell counts, and refitted the entire LOYO procedure in each replicate. This captures record-sampling and fitting variability conditional on the observed cohorts; record-level independence is assumed. Second, 5,000 paired year-cluster samples resampled the seven observed years (six without 2021), keeping all positional predictions from each selected year together. Model and baseline errors were resampled as pairs. These year-cluster intervals condition on the fitted predictions, reflect variation in cohort composition, and do not capture refitting uncertainty. With few clusters and overlapping training samples, neither analysis provides a definitive population-level uncertainty estimate. Intervals are exploratory and were not adjusted for multiple comparisons.

For each eligible position–test sample, the mass range, mass coefficient of variation, and SD of $\ln(\text{mass})$ were calculated. Spearman associations of log-mass SD with absolute exponent and held-out gain were described across positions within each test; no causal inference or formal hypothesis testing was applied. Pooled gains were also recalculated without DL predictions to quantify their contribution. Full-sample diagnostics included $\ln(P)$ versus $\ln(M)$, residual versus fitted, and normal Q–Q plots for all 47 models. Residual skewness, excess kurtosis, rank association between absolute residuals and log mass, and the increase in in-sample R^2 after adding a centered quadratic log-mass term were recorded descriptively. No alternative mean function was selected based on these checks.

Software and computational checks

Analyses used Python 3.12.14, NumPy 2.3.5, pandas 2.2.3, and SciPy 1.17.0. Computational checks verified that record identifiers were unique, that training and omitted-year record identifiers were separated within each fold, and that there was at most one held-out prediction per record, test, and scenario. Recalculated full-sample exponents were compared with the reconciled model outputs. Table 3

metrics were compared directly with the saved held-out predictions. Full-precision estimates were retained for calculations; rounding was applied only for presentation. Regression fitting used `scipy.stats.linregress`. Bootstrap refitting used equivalent least-squares moment equations. Comparator calculations reproduced all 13,037 primary held-out predictions across both scenarios to numerical tolerance. Bootstrap random seeds were 260924 (year clusters) and 260925 (record resampling).

AI assistance

ChatGPT (OpenAI) assisted with manuscript copy editing, analytical programming, and figure preparation. The author is responsible for the accuracy, integrity, and final approval of the manuscript, analyses, figures, and references.

Results

Across the 47 estimable position–test models, 26 reduced LOYO squared prediction error relative to the position-specific mean, while 21 increased it. Pooled benefits were therefore not uniform across positions (Table 3; Figure 3).

Analytical samples and descriptive characteristics

The source cohort comprised 2,333 records from 2020–2026 across eight playing-position groups. Body mass was available for 2,310 records, height for 1,796, and arm length and hand size for 1,793 each. Mean body mass ranged from 88.69 ± 4.74 kg in defensive backs to 142.02 ± 6.84 kg in offensive linemen. Anthropometric characteristics and measurement-specific sample sizes are presented in Table 1A.

Applying the separate-test eligibility criteria shown in Figure 1 yielded 1,700 sprint, 1,665 vertical-jump, 1,568 broad-jump, 885 bench-press, 911 shuttle, and 854 three-cone observations. These samples overlapped because records could contribute to more than one test. Within positions, sample sizes varied substantially: sprint samples ranged from 55 quarterbacks to 349 defensive backs, whereas three-cone samples ranged from 47 quarterbacks to 198 offensive linemen. The single quarterback bench-press observation was retained for descriptive purposes but excluded from regression and validation, leaving 884 modeled bench-press observations. Test-specific means, SDs, and sample sizes are reported in Table 1B.

Position-specific allometric relationships

Forty-seven position–test models were estimated. The direction and magnitude of the body-mass exponent varied across tests and positions (Table 2; Figure 2). Sprint exponents were negative across all

eight positions, ranging from -0.387 in defensive linemen to -0.041 in quarterbacks. The 95% confidence intervals included zero for quarterbacks and tight ends but excluded zero for the other six sprint models.

For vertical jump, exponents ranged from -0.785 in defensive linemen to 0.353 in quarterbacks; for broad jump, they ranged from -0.562 in defensive linemen to 0.144 in tight ends. In both jumping tests, the confidence intervals excluded zero for defensive and offensive linemen but included zero for the remaining positional groups. Bench-press exponents were positive across all seven estimable positions, ranging from 0.097 in linebackers to 2.192 in running backs. The linebacker interval included zero, whereas the other six bench-press intervals excluded zero (Table 2; Figure 2A–D).

Shuttle exponents ranged from -0.380 among defensive linemen to 0.240 among tight ends, and three-cone exponents ranged from -0.357 among defensive linemen to 0.092 among defensive backs. The corresponding confidence intervals and fitted coefficients are shown in Table 2 and Figure 2E–F. Model fit also varied: defensive-lineman models explained 57.8% of the variance in log sprint performance, 40.5% of the vertical-jump variance, 53.6% of the broad-jump variance, 49.1% of the shuttle variance, and 45.0% of the three-cone variance. For bench press, the highest R^2 was 0.201 among running backs. Several other position–test models had R^2 values close to zero (Table 2).

Held-out prediction and W-score distributions

Leave-one-year-out validation produced a prediction for every eligible observation in the 47 estimable all-years models. When predictions were pooled across positions within each test, the allometric models reduced the summed squared log-performance error relative to the position-specific intercept-only baselines by 26.0% for sprint, 9.7% for vertical jump, 16.6% for broad jump, 7.2% for bench press, 13.9% for shuttle, and 16.1% for three-cone (Table 3). The all-years analysis evaluated 319 non-empty eligible folds with no failures across the 47 estimable models. Excluding 2021 left 268 eligible folds and four failed folds; QB bench was excluded before fold construction in both scenarios (Supplementary Table S1).

Aggregate gains were not uniform across positions. Twenty-six of the 47 position–test combinations had lower held-out squared error than their baselines; 21 had higher error. Defensive-lineman models reduced error across all six tests, with reductions of 56.9% for sprint, 39.4% for vertical jump, 52.4% for broad jump, 4.1% for bench press, 48.6% for shuttle, and 43.9% for three-cone. In contrast, linebacker three-cone and offensive-lineman bench-press models increased squared error by 15.8% and 6.7%, respectively. Smaller positive and negative differences occurred in other combinations (Table 3; Figure 3A).

Across positions, mean held-out W-scores ranged from -0.076 for three-cone to -0.018 for vertical jump, and SDs ranged from 1.033 for vertical jump to 1.130 for bench press. The pooled correlations between W-scores and log body mass ranged from approximately zero for sprint and broad jump to 0.041 for bench press. Position-specific correlations showed greater variation, including -0.256 for linebacker

three-cone, -0.204 for linebacker shuttle, 0.245 for offensive-lineman bench press, and 0.211 for offensive-lineman shuttle (Table 3; Figure 3B).

Descriptive calibration across body-mass quintiles

Held-out W-score means across all-years mass quintiles ranged from -0.658 to 0.472 . The most negative mean occurred in the middle tight-end shuttle quintile ($n = 15$), and the largest positive mean occurred in the second quarterback three-cone quintile ($n = 6$). These are descriptive extrema, not validated thresholds. Supplementary Table S4 provides sample sizes and mass ranges for the ten largest absolute deviations in each scenario; all quintiles are provided in the numerical supplement.

Sensitivity analysis excluding 2021

Removing the 439 records labeled 2021 left 1,894 source records. The test-specific samples comprised 1,321 sprint, 1,286 vertical-jump, 1,201 broad-jump, 563 bench-press, 586 shuttle, and 518 three-cone pairs. Four running-back three-cone validation folds failed the minimum-training rule, leaving only one eligible running-back prediction and 498 three-cone predictions overall. Therefore, the corresponding running-back subgroup correlation and percentage error reduction are not displayed (Table 3; Figure 3C–D).

After refitting and validation without 2021, pooled squared-error reductions remained positive across all six tests: 30.2% for sprint, 9.2% for vertical jump, 18.4% for broad jump, 1.9% for bench press, 21.6% for shuttle, and 21.2% for three-cone. Pooled mean W-scores ranged from -0.076 to -0.006 , SDs from 1.023 to 1.112, and correlations with log body mass from -0.015 to 0.012 . Position-level performance continued to vary. Defensive-lineman models maintained lower prediction error across all six tests, whereas several other combinations remained worse than their baselines. Residual mass correlations included -0.247 for linebacker bench press and -0.170 for quarterback shuttle (Table 3; Figure 3C–D). These scenario comparisons are descriptive and were not formally tested for significance.

Uncertainty and positional contribution

For the sprint, the pooled error reduction of 26.0% had a record-refit interval of 20.0–30.2% and a conditional year-cluster interval of 19.5–35.6%. The corresponding bench intervals were 0.7–11.8% and -1.3 –10.9%; without 2021, both bench intervals included zero. Year-cluster intervals for the five non-bench DL models remained above zero, whereas the DL bench interval included zero (Table 4; Supplementary Table S2). These intervals capture exploratory uncertainty under the specified resampling schemes.

When DL predictions were excluded from the all-years pooled summaries, gains were 4.6% for sprint, -0.5% for vertical jump, approximately 0.0% for broad jump, 7.8% for bench press, -0.8% for shuttle, and

4.3% for three-cone. Thus, the concentration of benefit in DL was pronounced for locomotor and jumping tests but did not explain the bench-press result.

Comparator scaling and the 2026 holdout

Free-exponent models outperformed the chosen fixed-b benchmarks in pooled LOYO error for sprint, jumps, shuttle, and three-cone. For bench press, however, pooled error reductions were 7.2% for the free exponent, 9.5% for $b = 1$, and 7.9% for $b = 2/3$. Pooled bench residual–mass correlations were 0.041, 0.049, and 0.082, respectively. Ratio $b = 1$ markedly increased prediction error for the other five tasks; fixed-b results do not establish a universal preferred exponent (Table 4 and numerical supplement).

The 2026 holdout contained 189 sprint, 155 vertical-jump, 108 broad-jump, 39 bench, 58 shuttle and 45 three-cone predictions across 43 non-empty position–test cells, all meeting the training rule. Error reductions were 24.4%, 11.2%, –2.4%, –16.3%, 28.5% and 17.9%, respectively. Excluding 2021 from training preserved the direction of these six comparisons. Pooled 2026 mean W-scores were 0.499 for sprint, 0.749 for vertical jump, 0.451 for broad jump and 0.520 for bench, indicating cohort offsets despite some prediction gains. No 2026 broad-jump results were available for DL or LB, and no 2026 bench results were available for DB or TE. The aggregate temporal comparisons therefore also reflect test-participation composition; the strongest DL broad-jump equation was not tested in 2026. Full positional results are provided in Supplementary Table S3.

Mass spread and residual diagnostics

Across test-specific DL samples, the SD of $\ln(\text{mass})$ ranged from 0.090 to 0.100, compared with 0.030–0.033 for TE and 0.045–0.050 for QB (Supplementary Table S5). Across the eight positions, rank correlations between mass spread and LOYO gain were 0.55 for sprint, 0.43 for vertical jump, 0.45 for broad jump, 0.19 for shuttle, and 0.31 for three-cone. The corresponding correlation for bench was –0.39 across seven positions. Wider mass variation was associated with higher gains in several tests but was not a sufficient explanation across all outcomes.

Bench residual diagnostics showed marked tails in DL (skewness –2.22; excess kurtosis 15.03) and RB (–1.77; 5.79). Adding a quadratic log-mass term increased in-sample R^2 by 0.087 and 0.093 in those cells, respectively; these were descriptive checks, not validated alternative models. Dispersion was not uniform: the absolute-residual–log-mass rank association across bench cells ranged from –0.214 to 0.139. All 47 diagnostic panels and numerical summaries are provided in the supplement.

Discussion

The principal finding was heterogeneity in the value of mass adjustment. Twenty-six of 47 position–test models improved held-out prediction, and 21 worsened it. Defensive linemen accounted for much of the aggregate benefit in sprint, jump, and change-of-direction tests, whereas bench press showed a different

pattern (Tables 3–4; Fig. 3). The 2026 holdout further narrowed the conclusion: pooled gains did not carry forward for broad jump or bench press. The absence of DL broad-jump observations in 2026 limits direct comparison of that pooled result with LOYO. These results support testing the usefulness and calibration of each adjustment before interpreting it as an assessment score.

Analyzing tests separately made substantially more of the available data usable. From 2,333 source records, individual test samples contained 854 to 1,700 valid mass–performance pairs, and 884 observations supported bench-press modeling after excluding the single quarterback observation (Tables 1A–1B; Fig. 1). Eligibility for one test did not depend on completing the other five. This approach reflects the dataset's practical structure and provides estimates for the specific task under interpretation. It preserves observations without assuming that every prospect completes a common test battery.

Position and task specificity

Previous research has established the relevance of positional reference values and body-mass scaling for American football assessment (Gillen 2023; Nuzzo 2015; Oba et al. 2014; Robbins and Young 2012). The present findings extend this work by evaluating equations in omitted-year cohorts and comparing their predictions against position-specific baselines. The considerable variation in estimated exponents and held-out performance shows that a pooled relationship across all prospects cannot establish the value of adjustment within an individual position. Nuzzo (2015) examined Combine mass-normalization techniques and reference values; Gillen (2023) examined absolute and allometrically scaled positional percentiles. The present contribution evaluates cohort transfer and incremental prediction within individual position–test cells. It does not validate replacement percentile norms.

Sprint exponents were negative across all positions, indicating that higher mass was associated with lower reciprocal-time performance, although the confidence intervals included zero for quarterbacks and tight ends. Jumping relationships were also heterogeneous: negative exponents were most clearly supported among defensive and offensive linemen, while intervals included zero for the other positions (Table 2; Fig. 2). Bench-press exponents were positive across all seven estimable positions. These differences align with the contrasting demands of moving the body and completing repetitions against a fixed external load, but the observational models do not identify the underlying mechanisms. Body composition, segment dimensions, training, and positional selection could contribute to the observed relationships (Jaric 2002, 2003). A between-person exponent should therefore not be interpreted as the performance change an individual would experience after gaining or losing mass.

The bench-press findings also highlight the distinction between model fit and transfer to other cohorts. Positive fitted exponents did not guarantee better omitted-year prediction. For example, the offensive-lineman bench model increased the held-out squared error by 6.7% relative to its baseline. Across all positions, bench-press error decreased by 7.2%, falling to 1.9% after excluding 2021 (Table 3). Thus, the direction of an association alone is insufficient to recommend an adjustment. In addition, both fixed-

exponent bench comparators outperformed the free-exponent model in pooled LOYO prediction. The study therefore provides no basis for claiming that estimating a separate slope is always preferable.

Mass range and task mechanics

Mass range affects how much predictive information a slope can contribute. In a correctly specified linear log model, a wider predictor distribution increases the variation captured by the fitted mass term; when within-position variation is narrow, slope estimation and incremental prediction can be less informative. The observed DL mass spread is consistent with part of its larger gains, but modest cross-position associations and the opposite bench pattern rule out a simple claim that range explains most model failures. A wider predictor range alone does not imply a steeper biological exponent.

Tasks requiring acceleration or displacement of the body differ mechanically from repetitions against a fixed external load. Lean tissue, adiposity, limb geometry, and positional selection could influence the observed relationships (Jaric 2002, 2003). However, these were not measured mechanisms in this study. In particular, the data do not establish that DL have greater adiposity or that restricted mass ranges cause near-zero true exponents in skill positions.

Meaning and validation of W scores

W-scores quantify deviation from a mass-specific fitted expectation, building on residual methods and allometric modeling (Batterham and George 1997; Henry 1956; Nevill et al. 1992). Ratio scaling imposes an exponent of one and can introduce size associations when proportionality is inappropriate (Packard and Boardman 1988; Tanner 1949). The comparator results illustrate this problem for reciprocal-time and jump outcomes, while showing that a simpler fixed slope can perform better for bench repetitions. The chosen fixed benchmarks were sensitivity specifications, not physiological laws.

There is also a direct link between W-scores and conventional allometric adjustment. In a fitted model, the residual is $\ln(P/M^\alpha) - \alpha$, and W is this residual divided by SEE. Consequently, W-scores preserve the ranking of P/M^α within that model; standardization changes the scale, not the ordering. Their practical contribution is an explicit reference expectation and a common residual-error scale. They differ from an ordinary z-score of raw performance because the expected value depends on mass. Research using z-scores and composite athleticism scores addresses related but distinct assessment questions (Zhao et al. 2025).

Recent strength, independently validated powerlifting, and participant-separated oxygen-uptake studies support evaluating residual scores in their intended population while distinguishing statistical adjustment from physiological or external validity (Fahey 2026a, 2026c; Fahey and Rivera 2026).

Held-out evaluation is central to that distinction. With an intercept, ordinary least-squares residuals are orthogonal to log body mass in the estimation sample. Near-zero in-sample correlations therefore partly reflect score construction. Here, coefficients and SEE were estimated without the omitted-year observations. Pooled held-out correlations were small, but position-specific associations remained,

including -0.256 for linebacker three-cone and 0.245 for offensive-lineman bench press (Table 3; Fig. 3B). Pooling can obscure opposing associations within subgroups, so aggregate correlations should not determine whether a particular positional equation is satisfactory.

Across the pooled LOYO predictions, W-score SDs of 1.033 – 1.130 indicate greater held-out dispersion than the training SEE on average. This may reflect both estimation uncertainty and cohort differences. The 2026 shifts in W-score means further show that pooled multi-year calibration can mask a later-cohort offset. The scores should not be treated as standard-normal percentiles. Record-refit and year-cluster intervals address different uncertainty questions; their disagreement for bench press illustrates sensitivity to the sampling framework rather than a definitive finding of improvement.

Testing year and selective participation

NFL reporting described college pro-day workouts during the altered 2021 assessment program (Patra 2021). Individual testing conditions for the 439 records labeled 2021 could not be verified. Removing these records and repeating model fitting and validation preserved positive pooled error reductions across all six tests, though the magnitude changed, particularly for bench press (Table 3; Fig. 3C–D). The aggregate finding therefore did not depend entirely on including 2021. However, this sensitivity analysis cannot establish equivalence between testing settings or attribute the differences to protocol changes, because removing a year also changes sample composition and training-set size.

Separate-test analysis does not eliminate selection bias. Prospects with an observed result may differ from those without one, and the reasons for missing tests were unavailable. Accordingly, the equations describe the observed participants in each position–test combination, not necessarily all eligible prospects. Sparse participation also remained consequential: no quarterback bench model was estimable, and after excluding 2021, only one running-back three-cone observation met the validation training rule. A larger total sample size therefore improves data use without resolving every subgroup limitation.

Practical interpretation and limitations

Raw results and adjusted scores address complementary questions. Raw time, distance, or repetition count describes the recorded performance; W quantifies how much it deviates from the position-specific expectation at that mass. A positive W-score does not imply a superior absolute result, and equal W-scores across tasks or positions do not indicate equal football value. The models do not justify combining scores into an overall athleticism index or automatically converting them into normal-distribution percentiles. For an exploratory application, both the raw result and its adjusted score should be reported, along with the applicable positional reference and validation evidence. A worked example illustrates the distinction. Using the full-cohort DL sprint equation, two hypothetical prospects both running 4.80 s at masses of 110 and 140 kg receive W-scores of approximately -1.18 and $+1.63$, respectively. Their raw times are identical; the heavier prospect is further above the model expectation. This example falls within the observed DL mass range and is not evidence that increasing body mass improves performance.

Application to individual position and test combinations

The strongest candidates for further assessment are the DL sprint, vertical jump, broad jump, shuttle, and three-cone equations, which showed substantial LOYO gains and positive conditional year-cluster intervals. Weak or negative gains in other cells warrant retaining raw performance as the principal reported result until further validation. Bench remains a secondary descriptive application. A rigid use/do-not-use list or a universal residual-correlation cutoff is not validated here. Positive incremental prediction should be considered alongside calibration, uncertainty, temporal stability, and practical purpose; Supplementary Tables S2–S3 provide evidence for each cell.

The study did not examine draft selection, playing time, injury, or subsequent professional performance. Earlier investigations examined draft outcomes and subsequent NFL performance (Kuzmits and Adams 2008; Robbins 2010; Teramoto et al. 2016), while a recent analysis reported limited predictive power of physical performance components for draft outcomes (Doyle et al. 2026). The present results do not establish incremental prediction of football success or improved scouting decisions. Such claims would require a separate validation study linking draft predictions to outcomes.

Several additional limitations affect generalizability. The source records were compiled retrospectively; original retrieval dates were not retained, and athlete identity was not independently reconciled across years. Unique record identifiers therefore do not rule out repeated individuals across validation folds. Measurement errors, source coding, and unrecorded testing conditions may influence the estimates. The project archive is cited for traceability (Fahey 2026b); the numerical supplement accompanies this manuscript. Surface, timing method, and direct fat-free-mass measurements were unavailable. Archiving cannot resolve unrecorded testing conditions. Body mass was the only regression predictor and could not distinguish lean tissue from adiposity or account for height, limb dimensions, and other relevant characteristics.

Log–log equations and a constant SEE remain simplified descriptions. Diagnostic plots show heavy residual tails, particularly for DL and RB bench press, and departures from a linear mean in some cells. Bench W-scores are therefore secondary descriptive indices, not validated count-model residuals. The 47 exponent intervals are exploratory and do not directly test differences between positions. The record bootstrap assumes independent records within position–year cells and conditions on the observed cohorts; uncertain person matching remains a limitation. The year-cluster bootstrap has only seven or six clusters, retains fitted predictions, and cannot capture refitting uncertainty or make overlapping LOYO folds independent. The retrospective 2026 holdout avoids training on later years but does not provide independent prospective validation. Adding a noisy slope can worsen prediction when mass contributes little incremental information, so negative gains are a reason to question that cell's adjustment, not proof against residual scoring in general. A new, independently identified cohort is required before establishing operational scoring rules.

Conclusion

Body-mass adjustment was selectively useful across NFL positional groups and tests. Separate-test eligibility preserved available observations, whereas cohort-based evaluation revealed limitations concealed by pooled results and in-sample residual orthogonality. W-scores can complement raw performance when the relevant positional model demonstrates transferable prediction and adequate calibration. The findings do not establish universal mass independence, normal-distribution percentiles, or prediction of football success.

Abbreviations

CI, confidence interval; DB, defensive backs; DL, defensive linemen; IRB, Institutional Review Board; LB, linebackers; LOYO, leave-one-year-out; NFL, National Football League; OL, offensive linemen; QB, quarterbacks; RB, running backs; RMSE, root mean squared error; SD, standard deviation; SEE, residual standard error; TE, tight ends; WR, wide receivers.

Declarations

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

TDF was responsible for study design, data preparation, analysis, interpretation and writing.

Data availability

We obtained source data for the 2020–2026 cohorts from Pro-Football-Reference.com, operated by Sports Reference LLC, via its annual NFL Combine results pages for 2020, 2021, 2022, 2023, 2024, 2025, and 2026. The 2021 page reports pro-day numbers because the NFL Combine was canceled that year.

Project data are deposited on Zenodo (Fahey 2026b; <https://doi.org/10.5281/zenodo.22884447>). Supplementary Information 1 contains Tables S1–S5; Supplementary Information 2 contains analysis data, code, numerical outputs (including comparators_and_holdout.csv), and the table/figure file map; Supplementary Information 3 contains residual diagnostics.

References

1. Batterham AM, George KP (1997) Allometric modeling does not determine a dimensionless power function ratio for maximal muscular function. *J Appl Physiol* 83:2158–2166. <https://doi.org/10.1152/jappl.1997.83.6.2158>
2. Doyle BP, Stanelle ST, Riechman SE, Mann JB (2026) The NFL Scouting Combine explains within-position physical performance variance but is a poor predictor of draft outcomes. *J Strength Cond Res*. <https://doi.org/10.1519/JSC.0000000000005654>
3. Fahey TD (2026a) Body-mass normalization of powerlifting performance: Development and independent validation of sex- and age-specific regression-residual scores. *Eur J Appl Physiol*. <https://doi.org/10.1007/s00421-026-06427-y>
4. Fahey TD (2026b) NFL Combine 2020–2026: Data and code for body-mass-adjusted performance analysis [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.22884447>
5. Fahey TD (2026c) Sex-specific allometric modeling and regression-residual normalization for size-adjusted strength assessment. *Eur J Appl Physiol* 126:4581–4592. <https://doi.org/10.1007/s00421-026-06276-9>
6. Fahey TD, Rivera MA (2026) Residual body-mass dependence in $\dot{V}O_{2\text{peak}}$ normalization across ratio, allometric and regression-residual methods. *Sci Rep*. <https://doi.org/10.1038/s41598-026-70210-4>
7. Gillen ZM (2023) Percentile rankings and position differences for absolute and allometrically scaled performance measures from the National Football League Scouting Combine. *J Strength Cond Res* 37:e613–e624. <https://doi.org/10.1519/JSC.0000000000004546>
8. Henry FM (1956) Evaluation of motor learning when performance levels are heterogeneous. *Research Quarterly. American Association for Health, Physical Education and Recreation* 27:176–181. <https://doi.org/10.1080/10671188.1956.10612868>
9. Jaric S (2002) Muscle strength testing: Use of normalisation for body size. *Sports Med* 32:615–631. <https://doi.org/10.2165/00007256-200232100-00002>
10. Jaric S (2003) Role of body size in the relation between muscle strength and movement performance. *Exerc Sport Sci Rev* 31:8–12. <https://doi.org/10.1097/00003677-200301000-00003>
11. Kuzmits FE, Adams AJ (2008) The NFL combine: Does it predict performance in the National Football League? *J Strength Cond Res* 22:1721–1727. <https://doi.org/10.1519/JSC.0b013e318185f09d>
12. Nevill AM, Ramsbottom R, Williams C (1992) Scaling physiological measurements for individuals of different body size. *Eur J Appl Physiol Occup Physiol* 65:110–117. <https://doi.org/10.1007/BF00705066>
13. Nuzzo JL (2015) The National Football League Scouting Combine from 1999 to 2014: Normative reference values and an examination of body mass normalization techniques. *J Strength Cond Res* 29:279–289. <https://doi.org/10.1519/JSC.0000000000000755>

14. Oba Y, Hetzler RK, Stickley CD, Tamura K, Kimura IF, Heffernan TP Jr (2014) Allometric scaling of strength scores in NCAA Division I-A football athletes. *J Strength Cond Res* 28:3330–3337. <https://doi.org/10.1519/JSC.0000000000000548>
15. Packard GC, Boardman TJ (1988) The misuse of ratios, indices, and percentages in ecophysiological research. *Physiol Zool* 61:1–9. <https://doi.org/10.1086/physzool.61.1.30163730>
16. Patra K (2021) Around 150 draft prospects to undergo in-person medical evaluations in Indianapolis. NFL.com. <https://www.nfl.com/news/around-150-draft-prospects-to-undergo-in-person-medical-evaluations-in-indianapo>
17. Robbins DW (2010) The National Football League (NFL) combine: Does normalized data better predict performance in the NFL draft? *J Strength Cond Res* 24:2888–2899. <https://doi.org/10.1519/JSC.0b013e3181f927cc>
18. Robbins DW, Young WB (2012) Positional relationships between various sprint and jump abilities in elite American football players. *J Strength Cond Res* 26:388–397. <https://doi.org/10.1519/JSC.0b013e318225b5fa>
19. Tanner JM (1949) Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J Appl Physiol* 2:1–15. <https://doi.org/10.1152/jappl.1949.2.1.1>
20. Teramoto M, Cross CL, Willick SE (2016) Predictive value of National Football League Scouting Combine on future performance of running backs and wide receivers. *J Strength Cond Res* 30:1379–1390. <https://doi.org/10.1519/JSC.0000000000001202>
21. Zhao Y, Wei X, Zhao K (2025) Analysis of Z-score and total score of athleticism on drafted and undrafted players from the NFL scouting combine. *Sci Rep* 15:21742. <https://doi.org/10.1038/s41598-025-07383-x>.

Tables

Tables 1 to 4 are available in the supplementary files section.

Figures

Position-specific assessment of six performance tests

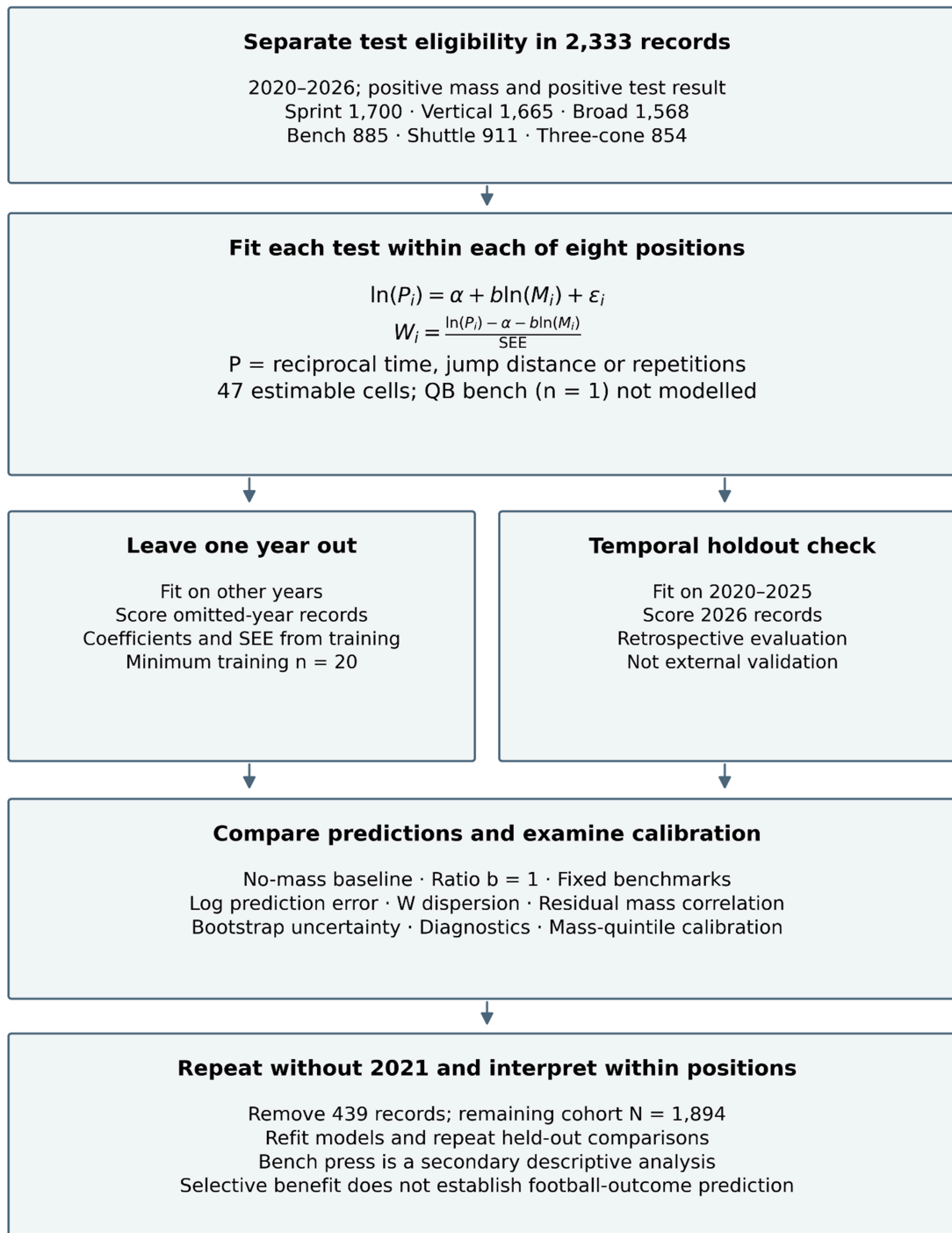


Figure 1

Figure legend not provided with this version

Body-mass exponents by test and position

2020–2026 • Points: exponent b • Bars: 95% confidence interval

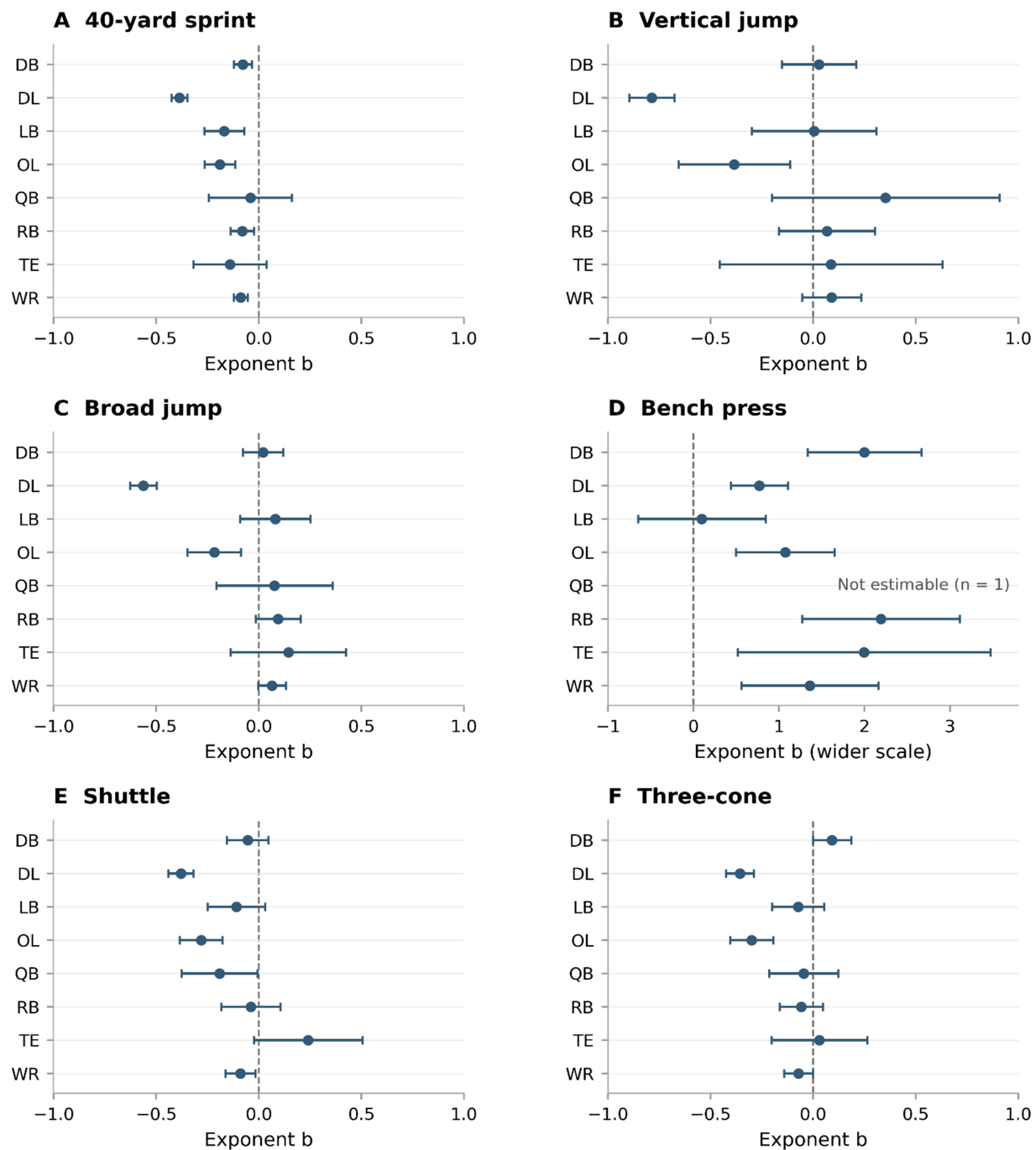


Figure 2

Figure legend not provided with this version

Held-out validation by test and position

Leave-one-year-out predictions • Results correspond to Table 3

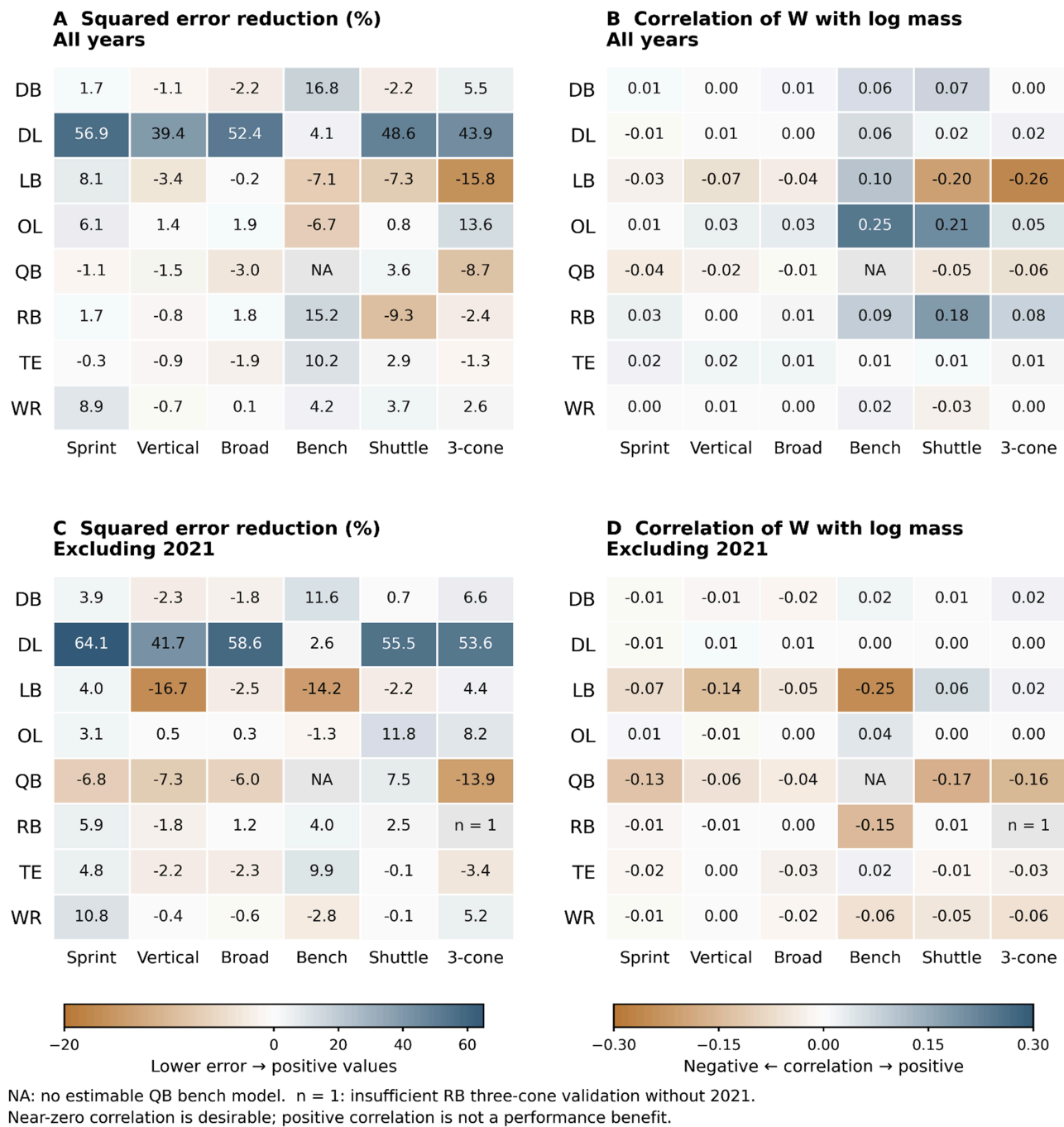


Figure 3

Figure legend not provided with this version

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1AB.docx](#)
- [Table2.docx](#)
- [Table3.docx](#)
- [Table4.docx](#)
- [SupplementaryTables.docx](#)