International Journal of
EXERCISE SCIENCE

Original Research

Thirty-year Trends of Study Design and Statistics in Applied Sports and Exercise Biomechanics Research

GEORGE VAGENAS‡1, DIMITRIA PALAIOTHODOROU†1, and DUANE KNUDSON‡2

[1]Faculty of Physical Education & Sport Science, National and Kapodistrial University of Athens, Athens, GREECE;
[2]Department of Health and Human Performance, Texas State University, San Marcos, Texas, USA.

†Denotes doctoral student author, ‡Denotes professional author.

ABSTRACT

International Journal of Exercise Science 11(1): 239-259, 2018. This study documented the change in study design and statistics employed in applied sports and exercise biomechanics research from 1985 to 2014. The sample comprised 676 data based original research reports published in the *Journal of Applied Biomechanics* (JAB) from 1985 to 2014. Eight design and 10 statistical criteria were extracted from each study. Descriptive statistics were calculated and change in study criteria over time were documented. Design criteria that did not change over time, remaining at relatively low levels of rigor, were widespread (71%) use of small (2-20) sample sizes and examination of numerous dependent variables (26.6% with >13). The number of experimental groups and independent variables also did not change with typically 1 to 2 reported. There was a significant 61% linear increase in randomization of participants into groups, however by 2014 still a minority (39%) of studies were not reporting randomized assignment. Types of statistical analysis showed positive changes over time with a 48% quadratic decrease in descriptive analyses, a 3% linear increase in nonparametric statistics, and a 45% linear increase in reporting parametric statistical analysis. Changes in specific statistical methods included a 9% linear decrease in bivariate correlation and a 73% linear increase in ANOVA. Reporting of assumptions had a 35% linear increase, yet in 2014 sixty-five percent still did not report on meeting statistical assumptions. Changes in test statistics included a linear 56% increase of reporting observed P values and a quadratic 29% increase in reporting effect sizes beginning in the late 1990s. It was concluded there was evidence of small improvements in research design and statistics in JAB over the last 30 years; however, there is still room for improvement to meet higher levels of research rigor and current recommendations on statistical analysis and reporting.

KEY WORDS: Assumptions, bias, effect size, sample size, P values, quality, rigor

INTRODUCTION

Scientific research strives to provide better models of reality and eventually apply that knowledge to improve human wellbeing. The quality of research design and statistical analysis in research directly affects this advancement of knowledge. Despite impressive developments in human knowledge and its application in technology, there have been growing concerns about errors and bias in research reports in many disciplines that undermine repeatability and advancement of knowledge (24, 35, 64, 55, 13, 32, 7). To what extent have these problems influenced advancements in the field of kinesiology/sport and exercise science? This field, hereafter referred to as kinesiology, is an expanding discipline with research on all aspects of human physical activity that also might have a body of knowledge threatened by accumulation of erroneous results.

Several studies that have examined the methodological rigor of research reports in kinesiology sub-disciplines have been published. Reports investigated the rigor of the research designs and the statistics in adapted physical activity, biomechanics, medicine, nutrition, pedagogy, and psychology. Observations on the quality of kinesiology published reports include weaknesses in reporting statistical assumptions and statistical analysis (15, 16, 49, 57, 63, 74) and weaknesses in experimental designs (27, 49, 57, 74). In sports medicine, there appears to be improvement in research rigor over time (12), however randomized control trials remain rare (10). Despite advancements in statistical analysis, substantial percentages of published kinesiology research have weak designs, small samples, and errors in statistical analysis and interpretation. The kinesiology sub-discipline of applied sports and exercise biomechanics, however, is a highly quantitative field with little qualitative research, that has a history of several articles calling for improvements in research design and statistical analysis.

Early biomechanics articles and commentary discussed methodological problems like uncorrected statistical testing of multiple dependent variables (53) and greater use of within-subject rather than between-subject designs in biomechanics research (3, 4, 38). Subsequent studies continued to point out that these methodological and statistical problems were still apparent in biomechanics research reports (54, 40, 41, 42, 45, 66). Recently, studies of applied biomechanics reports indicated that despite significant increases in coauthorship over twenty years there have been no changes in sample size (43, 44). Both these observations have been reported in other kinesiology sub-disciplines (44). Continued errors in design and statistical analysis in applied biomechanics reports may pose a confidence crisis in knowledge development in this field (45).

The present review builds upon this recent work to confirm the rigor of design and statistical analyses of applied biomechanics research reports over a 30-year period. It was hypothesized that the design and statistical rigor of applied biomechanics research reports would not have significantly changed over the last 30 years. The study provided evidence on potential improvements in research design and analysis in applied research in biomechanics over time. According to Ioannidis (37) "the study of the trajectory of the credibility of scientific findings and of ways to improve it is an important scientific field on its own."

METHODS

Applied sports and exercise biomechanics research is published in several multidisciplinary kinesiology journals, however in limited numbers compared to biomechanics and other scientific journals. The journal selected for this study was the *Journal of Applied Biomechanics* (JAB) because it represents the longest continuously published source of applied biomechanics research related to kinesiology, beginning publication in 1985 as the *International Journal of Sport Biomechanics*. Thus, the initial source of sports and exercise biomechanics papers comprised the 866 papers published in JAB during the first three decades of its circulation (1985 to 2014). After excluding editorials, letters to the editor, technical notes, case studies, modeling, and reviews, a final sample of 676 data based original research reports was retained for analysis. No attempt was made to exclude reports relationship to kinesiology when authors chose to submit under clinical, neuroscience, or ergonomics review areas of JAB beginning in 1993.

Each retained research report was analyzed using eight design specific (Table 1) and five statistical analyses specific (Table 2) criteria. A sub-analysis was performed by classifying the statistical analysis used into five groups (Table 3). The frequency (f) and proportion (%) of studies for the eighteen criteria were calculated annually and for the 30-year sample.

The progression of each variable across the 30 years studied was statistically tested using polynomial trend analysis via GLM-ANOVA. We, thus, conducted 18 tests in two sets (families): (i) eight tests for the study design criteria and (ii) ten tests for the study statistics criteria. Multiplicity implied protection against inflation of the type I error rate (i.e. 8, 59). Thus, to keep the family-wise alpha level for each of the two sets of analyses at 0.05, we tested the significance of each conducted statistical test at the Šidàk-Bonferroni (61) adjusted probability of $1-(1-0.05)^{1/8} = 0.006391$ for the design variables, and $1-(1-0.05)^{1/10} = 0.005116$ for the statistics variables.

To assess the size of each analyzed variable's percent (%) value and change (Δ%) overall and at specific points in the 30-year span studied, Batterham and Hopkins (6) extension of Cohen's (19) scale of assessing outcome statistics on frequencies were employed: 0-10% (very low), 11-30% (low), 31-50% (medium), 51-70% (high), 71-90 (very high), 91-100% (excellent). All statistical analyses were performed with SPSS version 23.

RESULTS

Study Design

Almost all reviewed studies (99.6%) used nonrandom samples, typically (71%) comprising 2-20 participants (Table 1). Most all (91%) studies examined 1-2 groups, with more fixed (69.5%) than randomized group assignment (30.5%). About half of the studies (54%) collected data using 1-3 trials per participant, with 15% of the studies collecting 9 or more trials per participant. Most of the studies examined 1-2 independent variables (76%). Studies often examined numerous dependent variables (DVs), 45% reporting 3-8 dependent variables and 27% of studies reporting

13 or more DVs. In 57% and 64% of the studies, the text included some information on limitations or recommendations, respectively.

Table 1. Categories and counts (f, %) for the study design criteria (N=676 data-based studies).

Criterion	Ordered Categories						
	1st	2nd	3rd	4th	5th	6th	7th
Size (N) of Sample [a]	2-10 242, 35.8%	11-20 237, 35.1%	21-30 87, 12.9%	31-40 43, 6.4%	41-50 15, 2.2%	51-60 19, 2.8%	≥ 61 33, 4.9%
Randomization (of partic. to groups)	No 470, 69.5%	Yes 206, 30.5%					
Num. of Groups (of participants)	1 414, 61.2%	2 200, 29.6%	3 19, 2.8%	4 32, 4.7%	≥ 5 11, 1.6%		
Num. of Independ. Variables (IVs)	0 37, 5.5%	1 292, 43.2%	2 222, 32.8%	3 83, 12.3%	4 18, 2.7%	≥ 5 24, 3.6%	
Num. of Dependent Variables (DVs)	1-2 83, 12.3%	3-4 121, 17.9%	5-6 99, 14.6%	7-8 83, 12.3%	9-10 63, 9.3%	11-12 47, 7%	≥ 13 180, 26.6%
Num. of Trials (per participant)	1 125, 18.5%	2 74, 10.9%	3 163, 24.1%	4 46, 6.8%	5 106, 15.7%	6-8 60, 8.9%	≥ 9 102, 15.1%
Study Limitations	Not Reported 288 , 42.6%	Reported 388, 57.4%					
Recommendations (for future research)	Not Reported 244, 36.1%	Reported 432, 63.9%					

[a] Sampling Type: Non-random 674 (99.6%), Random 3 (0.4%).

Study Design Trends Over Time

All design variables met assumptions for polynomial trend analysis. Four of eight study design criteria in JAB research reports had statistically significant modest to large linear increases over the last 30 years (Figures 1 to 4). The number of trials tested per participant increased significantly by 3.7 trials across years (Figure 3, $R^2 = 0.36$, $P < 0.001$). Randomization (of participants to groups/treatments) increased significantly by 61% across years (Figure 1, $R^2 = 0.56$; $P < 0.001$). The number of studies reporting limitations and recommendations increased 66% (Figure 4, $R^2 = 0.44$, $P < 0.001$) and 36%, respectively over 30 years (Figure 4, linear $R^2 = 0.60$, $P < 0.001$). Sample size, number of groups of participants, number of independent variables (IVs), and the number of DVs showed no significant trend over time.

Table 2. Categories and counts (f, %) for the study statistics criteria (N=676 data-based studies).

Criterion	Ordered Categories				
	1st	2nd	3rd	4th	5th
Statistical Assumptions	Not reported 447, 66.1%	Reported 229, 33.9%			
Type of Statistics	Descriptive 59, 8.7%	Non-Param. 37, 5.5%	Parametric 580, 85.8%		
Statistical Method	Descriptive 59, 8.7%	Bivariate Correlation 59, 8.7%	Two-Group Comparison 150, 22.2%	ANOVA 335, 49.6%	Multivariate 73, 10.8%
Effect Size [a] (ES)	Not Reported 496, 88.9%	Reported 62, 11.1%			
Observed [b] Signif. (P)	Not Reported 325, 52.7%	Reported 292, 47.3%			

[a] based on the 558 studies that used other than descriptive analysis (8.7%) or bivariate correlation (8.7) as the main statistical method; [b] based on the 617 studies that used other than descriptive analysis (8.7%) as the main statistical method.

Table 3. Subcategories and counts (f, %) for the (main) statistical methods (N=676 data-based studies).

Stat. Method*	f, %	Subcategories						
Bivariate Correlation	59, 8.7%	Pearson 50, 84.7%	Simple Regress. 7, 11.9%	Spearman 2, 3.4%				
Two-group Comparison	150, 22.2%	t-test 126, 84%	Mann-Whitney 9, 6%	Wilcoxon 15, 12%				
Analysis of Variance [a]	335, 49.6%	ANOVA 325, 97%	Kruskall-Wallis 4, 1.2%	Friedman 6, 1.8%				
Multivariate Methods	73, 10.8%	Multiple Regress. 26, 35.1%	MANOVA 37, 50.1%	Factor Analysis 6, 8.1%	Discrim. Analysis 4, 5.4%			
Multiple Comparisons [a]	103, 15.2%	Scheffé 13, 12.6%	Tukey 39, 37.9%	Bonferroni 8, 7.8%	Newman-Keuls 9, 8.7%	Duncan /Dunnett 3, 2.9%	LSD 6, 5.8%	t-test 25, 24.3%

* From Table 2: [a] In order of rigor of control of the family-wise error rate: Scheffé, Tukey, Bonferroni (high control); Newman-Keuls, Duncan (moderate to low control); Dunnett, LSD, t-test (no control); Šidàk-Bonferonni, Holm, and Hochberg tests are lacking.

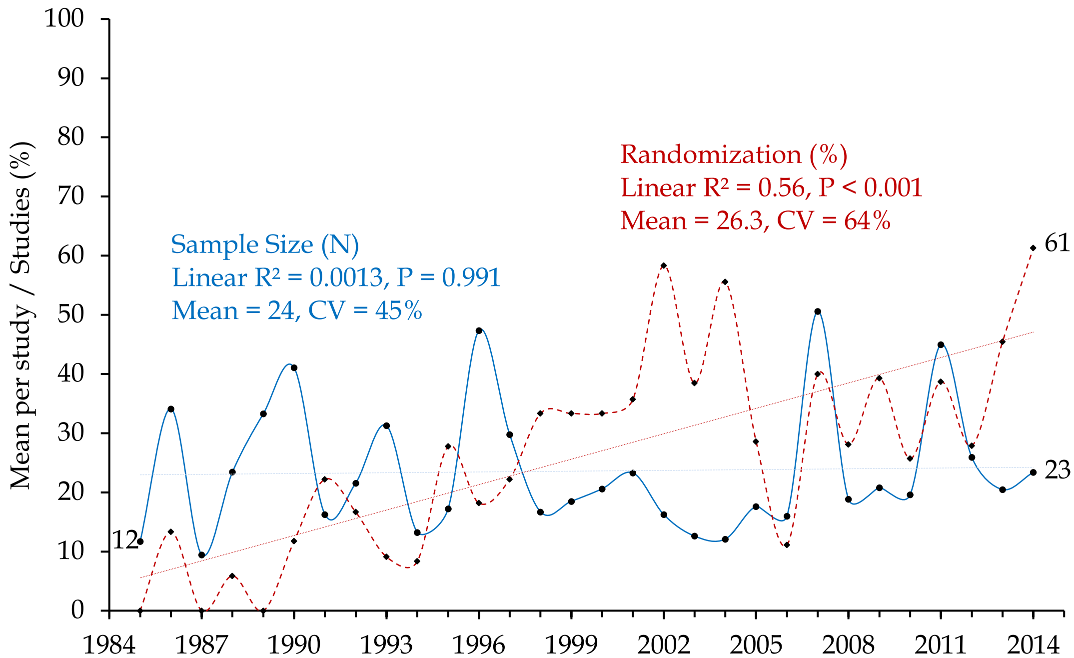

Figure 1. Variation and progress in the size of the samples and in the number of studies using randomization (of participants to groups/treatments).

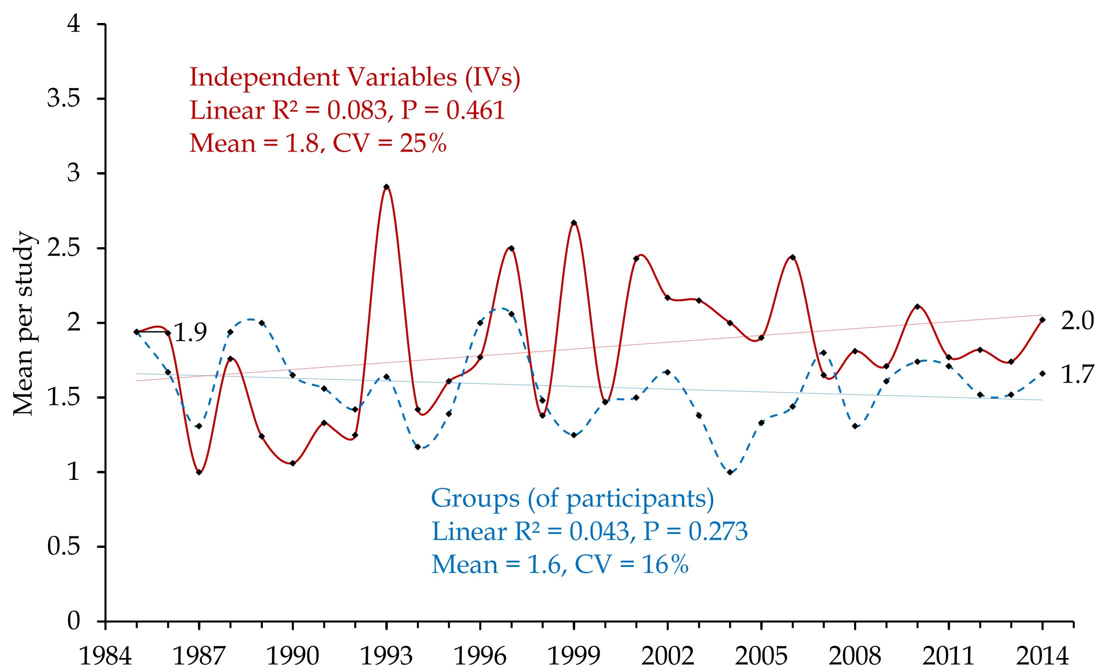

Figure 2. Variation and progress in the number of independent variables (IVs) and in the number of groups (of subjects).

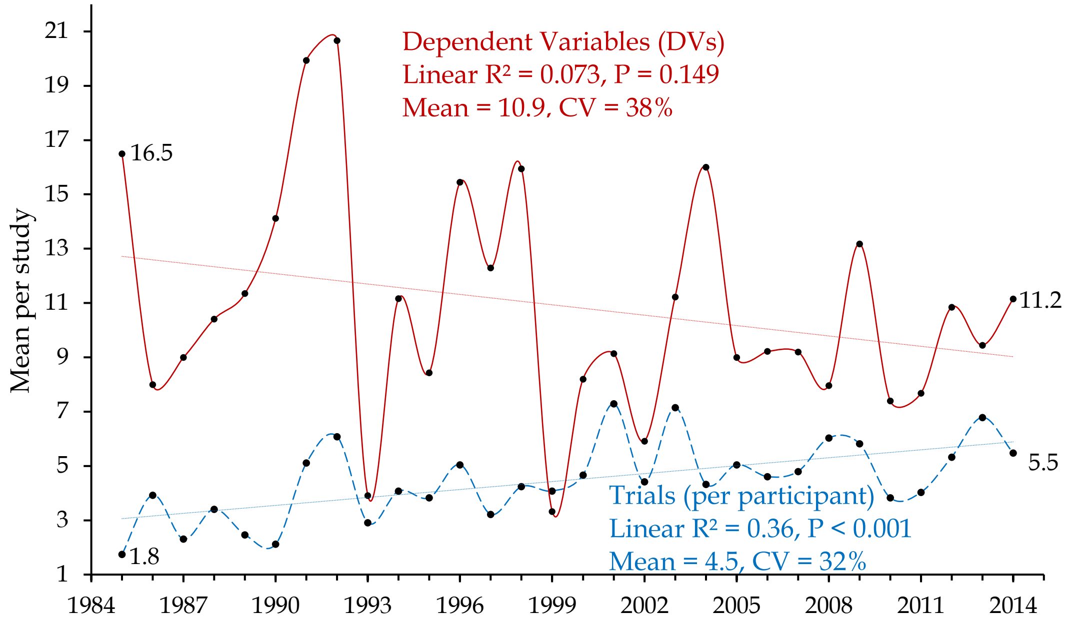

Figure 3. Variation and progress in the number of dependent variables (DVs) and in the number of (repeated) trials (per subject and condition).

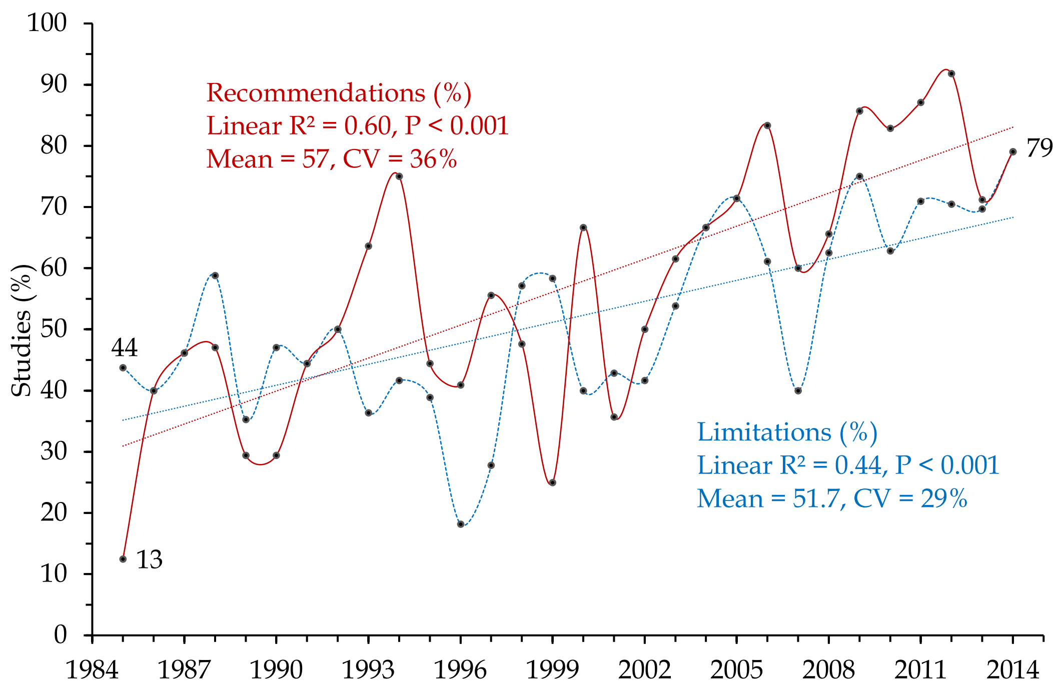

Figure 4. Variation and progress in the number of studies reporting limitations & recommendations.

Study Statistics

A minority (40%) of the reviewed studies reported data on statistical assumptions (Table 2). Most studies used parametric (85.8%) rather than non-parametric (5.5%) or descriptive (8.7%) statistics, and ANOVAs (49.6%) or two-group comparisons (22.2%), rather than bivariate correlations (8.7%) or multivariate methods (10.8%). The most common main statistical method was Pearson (84.7%) for bivariate correlations, t-test (84%) for two-group comparisons, ANOVA (97%) for Analysis of Variance, and MANOVA (50.1%) and Multiple Regression (35.1%) for multivariate methods (Table 3). In parametric ANOVA analyses, the most frequent post hoc tests conducted were Tukey's HSD (37.9%) and t-tests (24.3%). Only 43.2% of the reviewed studies reported observed P-values for test statistics, and only 9.2% reported effect sizes (ES).

Study Statistics Trends Over Time

All study statistics variables met assumptions for polynomial trend analysis. Eight of ten study statistics criteria in JAB research reports had statistically significant changes of different directions and shapes over the last 30 years (Figures 5 to 8). There were significant linear increases in studies statistical assumptions (Figure 5; $R^2 = 0.54$, $P < 0.001$) and use of parametric statistics (Figure 5, $R^2 = 0.40$; $P < 0.0164$). There was very small (3%) linear increase is use of non-parametric statistics (Figure 6, $R^2 = 0.20$, $P < 0.001$), however, there was a large (48%) quadratic decrease in the reporting of descriptive statistics (Figure 6, $R^2 = 0.81$, $P < 0.001$).

Regarding primary statistical tests there was no significant change in use of two group comparisons over time. Use of simple bivariate correlations has a significant (Figure 7a, $R^2 = 0.30$; $P = 0.00166$) 9% linear decrease. The use of multivariate statistics did not change over time, however use of ANOVA increased significantly (Figure 7b, $R^2 = 0.69$; $P < 0.001$) in a linear fashion. Reporting observed P-values for test statistics had a significant (Figure 8; $R^2 = 0.77$, $P < 0.001$) linear increase of 59%, while reporting effect sizes increased significantly (Figure 8; $R^2 = 0.73$, $P < 0.001$) in a quadratic fashion 29% beginning in the late 1990s.

DISCUSSION

The hypothesis of unchanging design and statistical analysis in applied research reports in JAB was not supported, with 12 of 18 research design and statistics criteria having significant linear or curvilinear changes over the last 30 years. While there was substantial evidence of changes in research design and statistical analysis in research reports published in JAB over time, several of the changing and stable criteria were not considered as evidence toward improvements in rigor or quality of applied biomechanics research reports. The importance of the present results to knowledge development in kinesiology from the applied sub-discipline of biomechanics are discussed in eight areas: sampling, number of trials analyzed, study variables, statistical assumptions, type of statistical analysis, statistical methods and multiplicity, reporting observed P-values and effect sizes and reporting of limitations and recommendations.

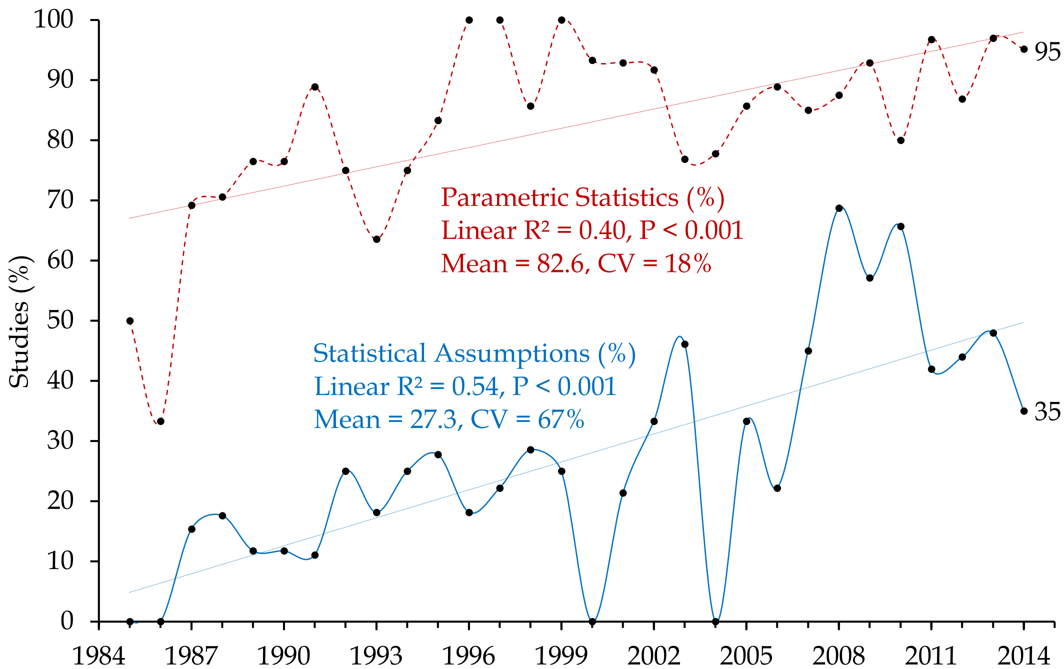

Figure 5. Variation and progress in the number of studies using parametric statistics and reporting statistical assumptions.

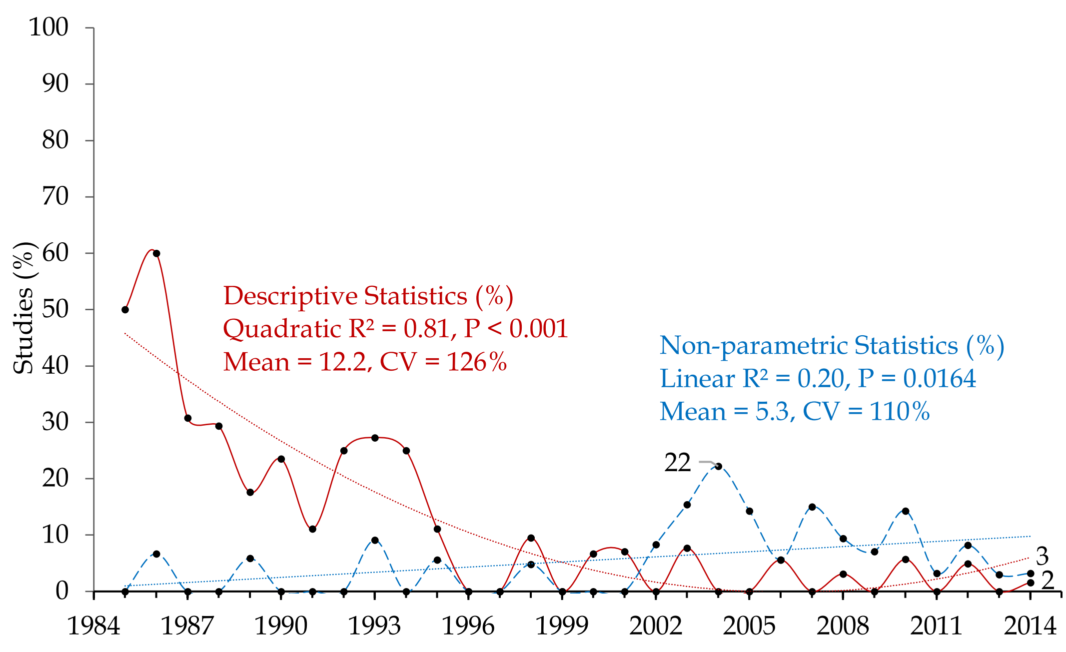

Figure 6. Variation and progress in the number of studies using descriptive or non-parametric statistics.

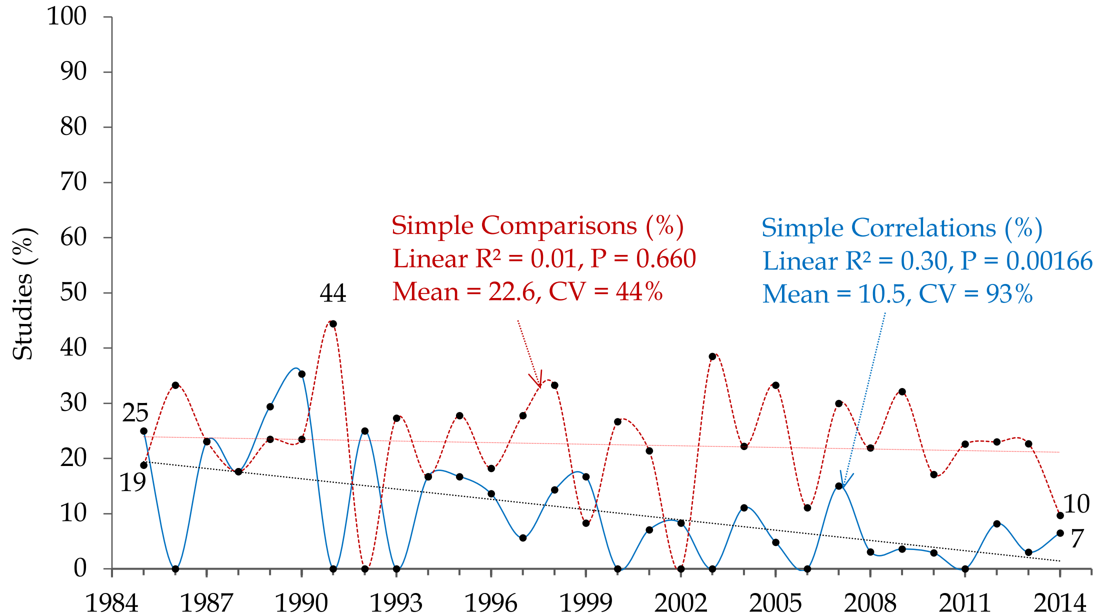

Figure 7a. Variation and progress in the number of studies using simple (two-group) comparison and simple (bivariate) correlation.

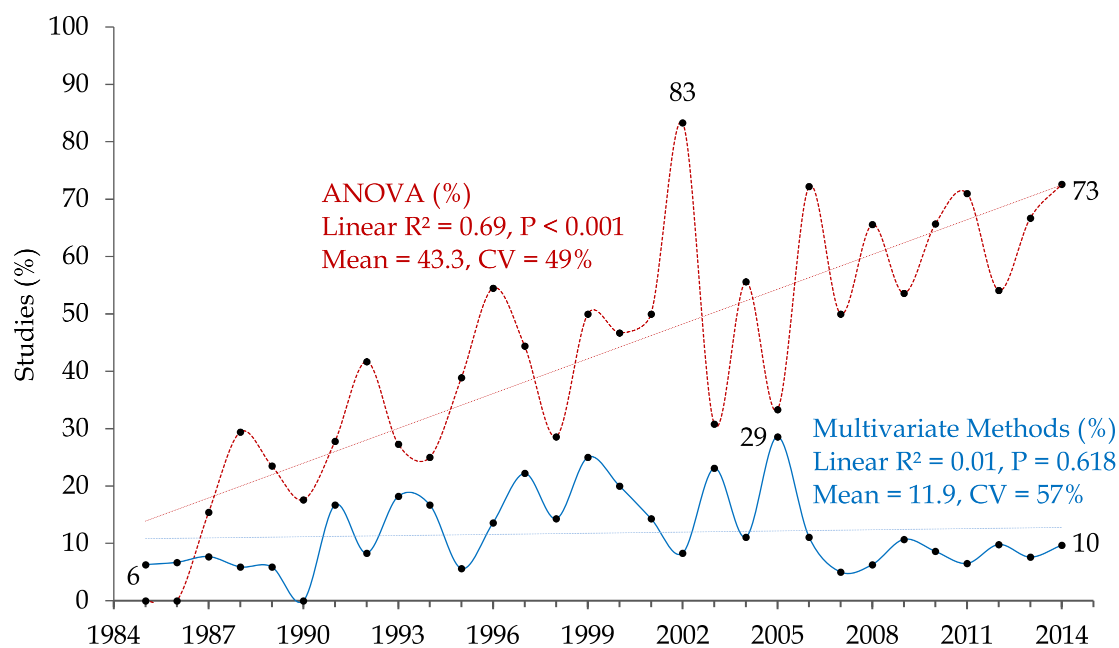

Figure 7b. Variation and progress in the number of studies using ANOVA and multivariate method.

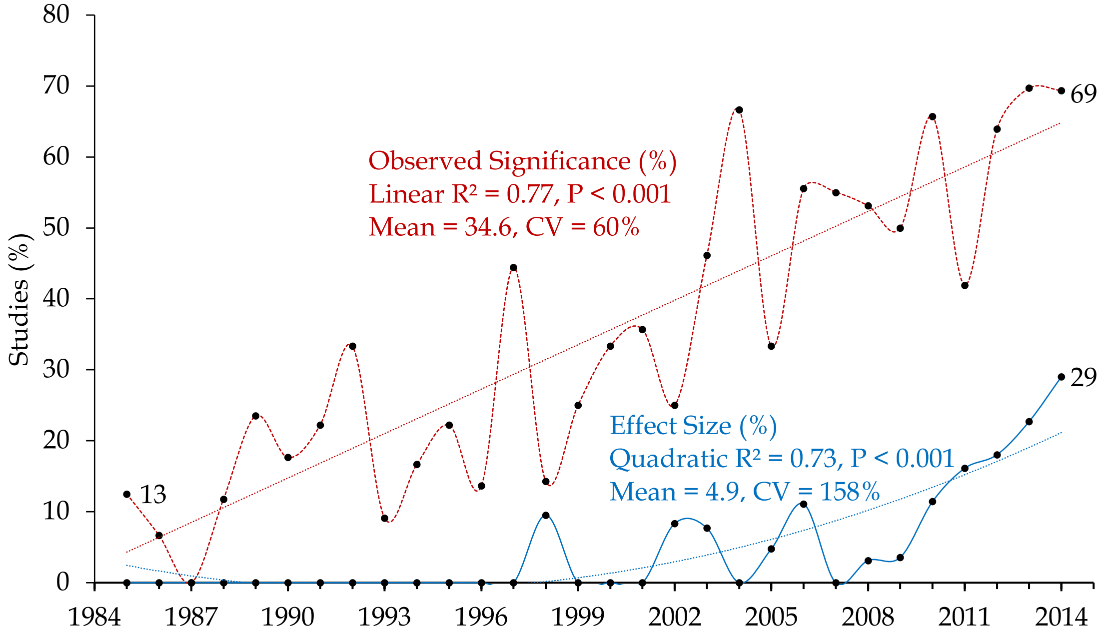

Figure 8. Variation and progress in the number of studies reporting observed statistical significance (P-value) and effect size (ES).

Sampling

There was no significant change in sample sizes over the 30-years, with a large positive skew and most (71%) studies with small (2-20 participants) samples. This is comparable to previous findings showing that many sports biomechanics studies recruit less than 20 participants (40, 42, 44) and these small samples have not changed over the last 20 and 25-years (43, 44, 46). This lack of improvements in sample size over time is also apparent in other sub-disciplines of kinesiology (44).

Rigorous study designs require well justified sample size estimation before data collection (39). However, in the reviewed biomechanics reports the present study this design criterion was lacking. This shortcoming limits the generalizability of the findings in this sub-discipline (58). Thus, future applied biomechanics research would support sample size estimation and justification of how it affects the size of the effects reported, although there are important research questions in the field that cannot access sufficiently large samples (47).

Ideally, before estimating sample size, researchers should consider not only the importance and impact of the expected effect, but also the burden the study puts on the participants (2). Thus, a sample can be "unethically" small or large (e.g., underpowered or overpower study) depending on the pragmatic difference of the expected effect. Recent guidelines to prospective authors of sports medicine papers suggest that a small sample might be ethical if the participant burden is low and the study findings are practically or clinically important, irrespective of their statistical significance (29). Study importance should outweigh participant burden and risk, but samples must be large enough for the identification of practically useful effects with high probability

(56). The dangers of small samples common in biomechanics and other kinesiology sub-disciplines (44) to valid estimates of effects has recently been summarized by Knudson (45).

Almost all (99.6%) studies in JAB employed convenience or purposive sampling rather than random sampling, which is often unfeasible in human research (20). Assuming sample representativeness is an ideal state of statistical methods (50) and implies that samples in empirical research adequately reflect strata and traits of the target population of interest (39). Much biomechanics and kinesiology research involves convenience samples of participants of unknown compliance to the strata and the traits of the target populations, introducing the risk of bias when inferring from non-random samples (39). Nevertheless, when random sampling is unfeasible, as is the case in most kinesiology sub-disciplines, the potential for improving future research in this respect is rather limited. Future studies should acknowledge these limitations and limit attempts at generalization to the population. Kinesiology journals should also consider publishing more replication studies to improve the generalizability of important results and sizes of effects given the small, non-random samples in most published studies. It is not uncommon for authors of biomechanics and kinesiology in the discussion section of research reports to erroneously claim application and generalization of their results from small samples to unspecified populations (47).

Most (69.5%) JAB studies did not even randomize participants into experimental groups or conditions. Since overall about two thirds of the studies did not randomize group assignment, there is even greater risk of sampling bias beyond the small, nonrandomized convenience sampling noted above. Though randomization is not always possible, random assignment of participants into groups/treatments eliminates systematic error due to inter-participant variation (39). Randomization "prevents selection bias, produces the comparable groups, eliminates the source of bias in treatment assignments, and permits the use of probability theory to express the likelihood of chance as a source for the difference of end outcome" (69). Random assignment, however, did increase to a high level (61%) in 2014 (Figure 1), however this level and the very low levels of randomized control trials (RCT) needs to increase in the future. The high cost of RCTs means they are even at very low levels (10% or less) in better funded fields like sports medicine (10, 12).

Repeated Trials Per Participant

Much of the research undertaken in applied sports and exercise biomechanics will un-avoidably rely on rather small samples of participants. In these cases, biomechanics experts suggest that the results of small samples may be greatly enhanced by increasing the number of tested trials (5, 54), as, for example, in single-group designs (4). While the number of repeated trials of participants per group/condition in the reviewed studies increased in a linear fashion (Figure 3) in JAB, the 2014 mean level of 5.5 trails per condition are just now becoming adequate for good reliability in many biomechanical variables. Aside from the usual practice of recording more trials than those finally retained for analysis in biomechanics research (54), there a need for collecting more trials per participant/condition to obtain representative and reliable data. Bates et al. (5) commented on the need for more repeated trials in sports biomechanics and determined that approximately 10, 5, and 3 trials are preferable for samples consisting of 5, 10, and 20 participants, respectively, to achieve a statistical power of 90%. Similarly, Salo, Grimshaw, and Vitasalo (60) estimated that a relative reliability of 0.90 is attainable in many kinematics variables with at least eight trials. Thus, improvement in more accurate effects from small samples can be enhanced if the number of trials required to obtain valid results is determined before data collection (54).

Study Variables

A critical issue in scientific research is determining the proper variables to represent the phenomenon under study. Ideally, the DVs must be valid, reliable, and sensitive enough to respond proportionally to the underlying effects of the IVs. This study confirmed recent reports of statistical analysis of numerous DVs in applied biomechanics research (40, 41), with many (27%) studies statistically testing more than 13 DVs (27%). There was no significant trend in the use of DVs over time (Figure 3).

The complexity biomechanical models and advances in the biomechanical measurement systems allow researchers to analyze more variables than those needed to test a pre-specified theoretical or deterministic model (17). Except of purely descriptive studies (8-9%), the design and statistical flaws of multiple univariate statistical tests of numerous dependent variables inflate the experiment-wise type I error rates (41), biasing and complicating the interpretations of the results and size of effects (54). The mean values of 11 DVs tested from a sample size of 23 for the 2014 year does not argue for the accuracy of the results reported in JAB. If kinesiology re-searchers prospectively reduce DVs statistically tested based on theory or previous research result, future research will have improved accuracy of effects identified and interpretability of those findings.

There was also no significant change in the mean number of IVs examined over time in JAB (Figure 2). Most studies in the journal examined one to three IVs. Experimental economy and the advantage of examination of interactions in factorial designs are benefits of research designs with numerous independent variables (39). It appears that most research reports in JAB are not taking advantage of examination of multiple IVs. Apparently, the explicit justification for the choice of both the DVs and IVs s (34) is not always the case in this research field.

Statistical Assumptions

Reporting statistical assumptions increased in a linear fashion (Figure 5), however still a majority (65%) of 2014 reports did not address meeting assumptions of the statistical tests used. This was consistent with the low levels reported by Knudson (40) but was nominally better than very low levels (7-10%) of studies reporting assumptions in other kinesiology sub-disciplines (15, 16, 49). It appears there is need for improvement in this area of research reporting in kinesiology. Assumptions are important conditions under which statistical models give valid results (25). All these conditions must have been met before the hypothesized model is fitted to the variables under analysis, while investigators should mention any shortcomings regarding these assumptions (21). Assumptions can also be stated regarding experimental methodology (e.g. instruments, designs, experiments) and if not carefully addressed may affect research quality (28).

Type of Statistical Analysis

Over time there were large changes in types of statistical analyses in JAB, with a 48% quadratic decrease in descriptive analyses, a 45% linear increase of parametric statistics, and a 3% linear increase in use of nonparametric statistics. The increase in parametric statistical analysis is consistent with the quantitative nature and high levels of measurement of biomechanics data. The choice of parametric or nonparametric statistics in a study depends primarily on the type of raw data; whether it is nominal, ordinal, interval, or ratio (67). Nominal and ordinal data are qualitative and therefore appropriate only for nonparametric analysis. Interval and ratio data are quantitative, but their analysis via parametric statistics is conditional to their degree of departure from the basic distributional assumptions (62). The central limit theorem makes many parametric tests robust to moderate violations of normality, but when the underlying distribution is too asymmetric, and the samples are unequal, the actual type I error rates deviate excessively from their nominal values, and the tests of directional hypotheses become inaccurate. In these cases, the analysis is untrustworthy even after optimal data transformation (34), and non-parametric tests (i.e. Kruskal-Wallis, Friedman) become more powerful than their standard parametric counterparts (i.e. ANOVA; 39).

Since the results showed that 84% of the studies used convenience samples with 30 or fewer participants, the low levels of reporting assumptions may be due to the small sample sizes. When samples are small ($N < 30$) it is impractical to test distributional normality and we instead rely on the small sample theory which allows for an optimal choice of robust test statistics in cases of distributional non-normality (51). When data are qualitative or deviate from normality or homoscedasticity, nonparametric tests are preferable than their parametric counterparts (62). In JAB, nonparametric test statistics were uncommon (Figure 6), and, when chosen, were mostly two-sample comparisons (i.e. Mann-Whitney and Wilcoxon) as oppose to multi-sample comparisons (i.e. Kruskal-Wallis or Friedman ANOVA), respectively (Table 3). Together, these results indicate that many JAB studies may have incorrectly used parametric statistics.

Statistical Methods and Multiplicity

The most common (49.6%) statistical method used in JAB was some type of parametric ANOVA (Table 2) that tended to increase (73%) in a linear fashion. There was also a linear 9% decrease in the use of bivariate correlations. On the other hand, the studies using multivariate statistical methods had no change, remaining at low (10.8%) levels. Of the studies using MANOVA, most incorrectly applied numerous univariate post hoc comparisons, rather than follow-up discriminant analysis or the stricter step-down F-tests (70).

When numerous DVs show moderate to high inter-correlations, a MANOVA is more powerful than multiple univariate models, and provides the means of a more comprehensive interpretation of the results (70). Multivariate methods are complex but unveil meaningful combinations of DVs, thus enhancing the interpretation of multifactorial phenomena. The relative contributions of the variables involved in multivariate models can then be determined by step-down F-tests or discriminant function analysis in the context of dealing with redundancy and multidimensionality in multivariate structure (65).

In choosing the optimal number of DVs for theoretically sound statistical models, statistical parsimony along with internal validity are most important. Parsimony is considered during the planning of the research and helps to explain the problem under study with the fewer possible DVs (11). Less parsimonious models unavoidably measure interrelated variables, and this makes interpretation of the shared variance problematic (72). When the DVs possess low inter-correlations (i.e. < 0.50, depending sample size) then ANOVA is more powerful, but conducting as many ANOVAs as the DVs requires adjusting for the resulting inflation of type I error rate. A similar adjustment of the probability of type I error is also applied in all multiple comparison analyses associated with significant F ratios. In these cases, appropriate multiple comparison testing may involve either the rigorous post hoc tests of Scheffé or Tukey for between-subjects designs or some variation of the Bonferroni adjustment (i.e., Bonferroni, Šidàk-Bonferroni) for within-subjects designs (39). Multiplicity adjustments should be ac-companied by proper control of the experiment-wise error rate, especially when the results of multiple tests are connected and summarized in one conclusion (8, 59).

There are two simultaneous and two sequential multiplicity adjustment methods. Simultaneous adjustments assume no priority among the DVs (i.e., of equal importance). Oppositely, sequential adjustments require prior hierarchy of the DVs. The simultaneous adjustment methods are Bonferroni or Dunn (23) and Šidàk-Bonferroni (61), and the sequential adjustment methods are Holm (31) and Hochberg (30). With c comparisons and α_{FW} the familywise error rate the per comparison alpha level (α_{PC}) becomes α_{FW}/c with Bonferroni and $1-(1-\alpha_{FW})^{1/c}$ with Šidàk-Bonferroni. In sequential adjustments of k progressive comparisons from 1 to c, the per comparison alpha level at each of the k steps becomes $\alpha/(c-k+1)$ in Holm's step-down method and α/k in Hochberg's step-up method. There is also the false discovery rate (number of Type I errors divided by the number of significant tests), a tool to ensure a less stringent multiplicity adjustment (9). The Bonferroni and Šidàk-Bonferonni approaches are strict when the comparisons are many and non-orthogonal. The sequential methods of Holm's and Hochberg's are more powerful but always at the expense of type I error (1). Interestingly, the relatively easy Holm's and Hochberg's methods have not yet been adopted by contemporary researchers, although, for example, SPSS Statistics 23 Algorithms provide both simultaneous and sequential Bonferroni and Šidàk-Bonferroni adjustments, as well as a version of Hochberg's range (GT2).

On the other hand, multiple comparisons can be either planned or post hoc. The present data confirmed previous results that planned comparisons in applied biomechanics are lacking. Instead, multiple comparison tests comprised 68% real post hoc tests, 24% t tests, and 8% Bonferroni adjusted (Table 3). Tukey tests were most frequent (38%) followed Scheffé (13%), Newman-Keuls (9%), Bonferroni's (8%), and LSD (6%). Except for Scheffé, Tukey, and Bonferroni, which provide good protection for type I error rate inflation, all other tests (about 31%) are unadjusted. The common erroneous statistical analyses based on numerous univariate tests uncorrected for inflation of type I errors observed in this study was consistent with previous studies of applied biomechanics (40, 41) and physical education pedagogy (15). It is likely these problems exist in other sub-disciplines of kinesiology, posting a threat to the credibility of knowledge generated in the field.

Reporting Observed P values and Effect Sizes

Research reports in JAB had a 56% linear increase in reporting of observed P values for statistical tests and a 29% quadratic increase in reporting effect sizes beginning in the late 1990s. This is evidence of recent adoption of advances in statistical analysis and reporting in research reports in the journal and, perhaps, the field of applied biomechanics. How these trends influence research quality is unclear, given the weaknesses observer earlier in sample sizes and uncorrected testing of numerous DVs, as well as the debate between traditional statistical testing and magnitude-based statistical testing (14, 34, 6).

Setting this debate aside, observed P values of test statistics and effect sizes are both essential metrics for a comprehensive evaluation of statistical evidence for the observed experimental effects in the context of sample size, number of tests, and study design (59). The rejection of a null hypothesis based on a statistical test, however, does not alone indicate a substantial or meaningful effect (52). The need to focus on all issues of design and statistical testing has led to replacing the arbitrary levels of statistical significance (e.g. 0.05, 0.01) with observed P values (22) and likely sizes of effects. Cohen (18) concluded, "what it matters best is the importance of power analysis, and the determination of just how large (rather than how statistically significant) are the effects that we study."

The numerous recommendations to increase reporting of effect sizes in kinesiology and biomechanics (33, 71, 41) may have begun to increase reporting of these important statistics in JAB in the late 1990s, similar to the increase in sport and exercise biomechanics beginning in 1991 noted by Mullineaux et al. (54). Despite these positive trends, however, still 71% of JAB reports published in 2014 did not report any size of effects. Not reporting sizes of effects is also frequent in research in other sub-disciplines of kinesiology (49, 15), in psychology (26), and in medicine (68). There is still a clear need of improvement in reporting sizes of effects in applied biomechanics using standardized effect sizes like Cohen's d and Glass' Δ or also variance accounted for effect sizes (41).

Limitations and Recommendations

Reporting of limitations and recommendations increased in a linear fashion 66% and 36%, respectively. By 2014, 70 to 80 percent of studies in JAB criteria. reported this important considerably (Figure 4). Limitations are methodological weaknesses that can reduce both the validity of the study and the credibility of the conclusions; they should be clearly stated beyond estimating the magnitude and direction of random and systematic errors (36). Limitations in applied biomechanics are not only design specific, but also model, algorithm, procedure, and instrument specific (73). There is room for improvement in JAB author's ethical responsibility to report study limitations and recommendations to improve future research (36). Recommendation may include important next steps in seeking higher levels of evidence to refute, confirm, or extend the current consensus of research on a study topic (47).

Limitations of the Present Study

The present study was limited to the 676 empirical studies in JAB and did not analyze the subfields of biomechanics identified by the journal beginning in 1993. Second, there could be investigator errors in reviewing and identifying all studies with inflated type I error rate, as often studies using numerous ANOVAs or t tests did not report enough information to confirm use or lack of control for alpha inflation. The results are influenced by the hierarchal research design and statistical criteria used and the inability to include all issues that affect research rigor and potential bias. Some research questions and logistical factors do not allow for the most rigorous quantitative designs and statistical analyses. Given these limitations of the data we did not employ logistic or non-linear regression for testing the association between year of publication and design or statistics criteria.

Conclusions

It was concluded there was evidence of small improvements in research design and statistics in JAB over the last 30 years, however there is still room for improvement to meet higher levels of research rigor and current recommendations on statistical analysis and reporting. Even with improvements, a large portion of the research reported in JAB by 2014 was not free of problems in several design and statistical analysis criteria. Most important to study rigor, there continues to be problems with small sample sizes, a small number of repeated trials tested per participant and condition, and errors in statistical analysis, particularly uncorrected univariate testing of numerous DVs. Remediation of these problems in future research are critical to the accuracy of research results. Given the evidence of continued weaknesses in peer review in kinesiology (48) and the slow nature of self-correction in science in general (37), this should be a call to action for all biomechanics and kinesiology researchers, reviewers, and journal editors to hold each other to contemporary standards of research design and statistical analysis.

Recommendations

This action for future sports and exercise biomechanics research, and in research in other quantitative sub-disciplines of kinesiology, should involve directly addressing the major methodological shortcomings noted in previous articles and supported by this study of the JAB. Design improvements include use of larger samples sizes, with sample size justification prior to data collection, more randomization of participants to treatments/groups, and more repeated trials per participant.

Given most studies employ multiple ANOVAs or other univariate statistical tests of numerous DVs that inflate the experiment-wise type I error rate, there should be justification of the biomechanical variables chosen for analysis and statistical analyses addressing multiplicity adjustments (8). Authors should also report the interrelations between the DVs under analysis, along with a summary of statistical diagnostics that address the assumptions of the statistical model used. In addition, authors should report how inflation of type I error is addressed when using multiple univariate statistical analyses. When studies focus on a combination of numerous DVs, authors should provide a theory based multivariate hypothesis and perform a suitable multivariate analysis. There should also be adequate justification for the use of parametric or nonparametric statistical tests. Reporting of observed P values of statistical tests is common, however there needs to be greater reporting of the practical importance through effect sizes of the findings.

Future research in biomechanics could focus on patterns of specific biomechanics research methods (e.g., force plate, motion analysis, muscle testing), as well as more study of misuse of research design statistical analyses and their likely effect on erroneous results in the field. Additional longitudinal studies on research design and statistical testing should also be conducted on journal reports from other sub-disciplines of kinesiology.

REFERENCES

1. Abdi H. The Bonferonni and Šidák corrections for multiple comparisons. Encycl. Meas. Stat 3: 103-107, 2007.
2. Bacchetti P, Wolf LE, Segal MR, McCulloch CE. Ethics and sample size. Am J Epidemiol 161(2): 105-110, 2005.
3. Bates BT. Comment on "the influence of running velocity and midsole hardness on external impact forces in heel-toe running". J Biomech 22: 963-965, 1989.
4. Bates BT. Single-subject methodology: an alternative approach. Med Sci Sport Exer 28(5): 631-638, 1996.
5. Bates BT, Dufek JS, Davis HP. The effect of trial size on statistical power. Med Sci Sport Exer 24(9): 1059-1065, 1992.
6. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. Int J Sport Physiol Perform 1(1): 50-57, 2006.
7. Begley CG, Ioannidis JP. Reproducibility in science. Circ Res 116(1): 116-126, 2015.
8. Bender R, Lange S. Adjusting for multiple testing - when and how?. J Clin Epidem 54(4): 343-349, 2001.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B (Methodological), 289-300, 1995.
10. Bleakley C, MacAuley D. The quality of research in sports journals. Brit J Sport Med 36(2): 124-125, 2002.
11. Box GE. Science and statistics. J Am Stat Ass 71: 791-799, 1976.
12. Brophy RH, Gardner MJ, Saleem O, Marx RG. An assessment of the methodological quality of research published in The American Journal of Sports Medicine. Am J Sport Med 33(12): 1812-1815, 2005.
13. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neu-rosci 14(5): 365-376, 2013.
14. Carver R. The case against statistical significance testing. Harvard Educ Rev 48(3): 378-399, 1878.
15. Chatoupis C, Vagenas G. An analysis of published process-product research on physical education teaching methods. Int J Appl Sport Sci 23(1): 271-289, 2011.
16. Chen A, Zhu W. Revisiting the assumptions for inferential statistical analyses: A conceptual guide. Quest 53(4): 418-439, 2001.
17. Chow JW, Knudson DV. Use of deterministic models in sports and exercise biomechanics research. Sport Biomech 10(3): 219-233, 2011.
18. Cohen J. Things I have learned (so far). Am Psychol 45(12): 1304-1312, 1990.
19. Cohen J. A power primer. Psychol Bull 112(1): 155-159, 1992.
20. Collins KM, Onwuegbuzie AJ, Jiao QG. Prevalence of mixed-methods sampling designs in social science research. Eval Res Educ 19(2): 83-101, 2006.
21. Dowdy S, Wearden S, Chilko D. Statistics for Research (Vol. 512). John Wiley & Sons; 2011.

22. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomized and non-randomized studies of health care interventions. J Epidemiol Commun H 52(6): 377-384, 1998.

23. Dunn OJ. Multiple comparisons among means. J Am Stat Ass 56: 52-64, 1961.

24. Elms AC. The crisis of confidence in social psychology. Am Psychol 30(10): 967-976, 1975.

25. Everitt B. The Cambridge dictionary of statistics. New York: Cambridge University Press; 2006.

26. Ferguson CJ. An effect size primer: A guide for clinicians and researchers. Prof Psychol: Res Pract 40: 532–538, 2009.

27. Gill DL. Measurement, statistics, and research design issues in sport and exercise psychology. Meas Phys Educ Exer Sci 1(1): 39-53, 1997.

28. Hagger MS, Chatzisarantis NL. Assumptions in research in sport and exercise psychology. Psychol Sport Exer 10(5): 511-519, 2009.

29. Harriss DJ, Atkinson G. Ethical standards in sport and exercise science research: 2016 update. Int J Sport Med 36(14): 1121-1124, 2015.

30. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75(4): 800-802, 1988.

31. Holm SA. A simple sequentially rejective multiple test procedure. Scand J Stat 6(2): 65-70, 1979.

32. Hoekstra R, Morey RD, Rouder JN, Wagenmakers EJ. Robust misinterpretation of confidence intervals. Psychon Bull Rev 21(5): 1157-1164, 2014.

33. Hopkins WG. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a P value. Sportsci 11: 16–20, 2007.

34. Hopkins W, Marshal SW, Batterham AM, Hanin J. Progressive statistics for studies in sports medicine and exercise science. Med Sci Sport Exerc 41(1): 3-12, 2009.

35. Ioannidis JP. Why most published research findings are false. PLoS Med 2(8): 696-701, 2005.

36. Ioannidis JP. Limitations are not properly acknowledged in the scientific literature. J Clin Epidemiol 60(4): 324-329, 2007.

37. Ioannidis JP. Why science is not necessarily self-correcting. Perspect Psychol Sci 7(6): 645-654, 2012.

38. James CR, Bates BT. Experimental and statistical design issues in human movement re-search. Meas Phys Educ Exerc Sci, 1(1): 55-69, 1997.

39. Keppel G, Wickens TD. Design & Analysis. A researcher's Handbook (4th ed.). New Jersey: Pearson Prentice Hall; 2004.

40. Knudson D. Statistical and reporting errors in applied biomechanics research. In Q. Wang (Ed.), Proc XXIII Int Symp Biom Sport 2: 811-814. Beijing: China Institute of Sport Science; 2005.

41. Knudson D. Significant and meaningful effects in sports biomechanics research. Sport Biomech 8(1): 96-104, 2009.

42. Knudson D. Authorship and sampling practice in selected biomechanics and sports science journals. Perceptual and Motor Skills 112(3): 838-844, 2011.

43. Knudson D. Twenty-year trends of authorship and sampling in applied biomechanics research. Percept Motor Skill 114(1): 16-20, 2012.

44. Knudson D. Twenty years of authorship, sampling, and references in kinesiology research reports. Inter J Kinesiol High Educ, 1-9, 2017a.

45. Knudson D. Confidence crisis of results in biomechanics research. Sport Biomech, 1-9, in press.

46. Knudson D, Bahamonde R. Twenty-five year trends of authorship and sampling in ISBS proceedings. In ISBS-conference Proc Arch 16(1): 381-384, 2012.

47. Knudson D, Elliott B, Hamill J. Proposing application of results in sport and exercise research reports. Sport Biomech 13(3): 195-203, 2014.

48. Knudson D, Morrow Jr JR, Thomas JR. Advancing kinesiology through improved peer review. Res Quart Exer Sport 85(2): 127-135, 2014.

49. Kouvelioti R, Vagenas G. Methodological and statistical quality in research evaluating nutritional attitudes in sports. Int J Sport Nutr Exer Metab 25(6): 624-635, 2015.

50. Kruskal W, Mosteller F. Representative sampling, III: The current statistical literature. Inter Stat Rev, 245-265, 1979.

51. Lehmann EL. "Student" and Small-Sample Theory. Stat Sci 14(4): 418- 426, 1999.

52. Mayo DG, Spanos A. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. Brit J Philos Sci 57(2): 323-357, 2006.

53. Morris HH. Statistics and biomechanics: Selected considerations. In J. M. Cooper, & B. Haven (Eds.), Proc CIC Symp: Biomech (pp. 216-225). Bloomington, IN: Indiana State Board of Health; 1981.

54. Mullineaux DR, Bartlett RM, Bennett S. Research design and statistics in biomechanics and motor control. J Sport Sci 19(10): 739-760, 2001.

55. Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science a crisis of confidence? Perspect Psychol Sci 7(6): 528-530, 2012.

56. Prentice R. Invited commentary: ethics and sample size - another view. Am J Epidemiol 161(2): 111-112, 2005.

57. Reid G, Prupas A. A documentary analysis of research priorities in disability sport. Adapt Phys Activ Q 15: 168–178, 1998.

58. Rushton L. Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. Occup Envir Med 57(1): 1-9, 2000.

59. Sainani KL. The problem of multiple testing. Phys Med Rehab 1(12): 1098-1103, 2009.

60. Salo A, Grimshaw PN, Viitasalo JT. Reliability of variables in the kinematic analysis of sprint hurdles. Med Sci Sport Exer 29(3): 383-389, 1997.

61. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62(318): 626-633, 1967.

62. Siegel S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Company; 1956.

63. Silverman S, Solmon M. The unit of analysis in field research: Issues and approaches to design and data analysis. J Teach Phys Educ 17: 270-284, 1998.

64. Simmons JP, Nelson LD, Simonsohn U. False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22: 1359–1366, 2011.

65. Spector PE. Redundancy and dimensionality as determinants of data analytic strategies in multivariate analysis of variance. J Appl Psychol 65(2): 237, 1980.

66. Stergiou N, Scott, MM. Baseline measures are altered in biomechanical studies. J Biomech 38(1): 175-178, 2005.

67. Stevens S. On the theory of scales of measurements. Science, New Series, 103: 677-680, 1946.

68. Sullivan GM, Feinn R. Using effect size - or why the p value is not enough. J Grad Med Educ 4(3): 279-282, 2012.

69. Suresh K. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. J Hum Reprod Sci 4(1): 8-11, 2011.

70. Tabachnick BG, Fidell LS. Using Multivariate Statistics (5th ed.). Boston: Pearson Education, Allyn & Bacon; 2007.

71. Thomas JR, Salazar W, Landers DM. What is missing in p in < .05? Effect size. Res Q Exer Sport 62: 344-348, 1991.

72. Tonidandel S, LeBreton JM. Beyond step-down analysis: A new test for decomposing the importance of dependent variables in MANOVA. J Appl Psychol 98(3): 469, 2013.

73. Yeadon MR, Challis JH. The future of performance-related sports biomechanics re-search. J Sport Sci 12(1): 3-32, 1994.

74. Zhang J, DeLisle L, Chen S. Analysis of AAHPERD research abstracts published under special populations from 1968 to 2004. Adapt Phys Act Q 23(2): 203-217, 2006.