Journal of Motor Learning and Development, 2023, 11, 15–28
https://doi.org/10.1123/jmld.2022-0042
© 2023 Human Kinetics, Inc.

Human Kinetics
REVIEW

Low Prevalence of A Priori Power Analyses in Motor Behavior Research

Brad McKay,[1] Abbey Corson,[2] Mary-Anne Vinh,[2] Gianna Jeyarajan,[3] Chitrini Tandon,[3] Hugh Brooks,[2] Julie Hubley,[2] and Michael J. Carter[1]

[1]Department of Kinesiology, McMaster University, Hamilton, ON, Canada; [2]School of Human Kinetics, University of Ottawa, Ottawa, ON, Canada; [3]School of Interdisciplinary Sciences, McMaster University, Hamilton, ON, Canada

A priori power analyses can ensure studies are unlikely to miss interesting effects. Recent metascience has suggested that kinesiology research may be underpowered and selectively reported. Here, we examined whether power analyses are being used to ensure informative studies in motor behavior. We reviewed every article published in three motor behavior journals between January 2019 and June 2021. Power analyses were reported in 13% of studies ($k = 636$) that tested a hypothesis. No study targeted the smallest effect size of interest. Most studies with a power analysis relied on estimates from previous experiments, pilot studies, or benchmarks to determine the effect size of interest. Studies without a power analysis reported support for their main hypothesis 85% of the time, while studies with a power analysis found support 76% of the time. The median sample sizes were $n = 17.5$ without a power analysis and $n = 16$ with a power analysis, suggesting the typical study design was underpowered for all but the largest plausible effect size. At present, power analyses are not being used to optimize the informativeness of motor behavior research. Adoption of this widely recommended practice may greatly enhance the credibility of the motor behavior literature.

Keywords: metascience, sample size planning, positivity rates, effect size

Motor behavior research frequently involves proposing hypotheses and subjecting them to statistical tests. The probability that a statistical test will correctly reject the null hypothesis, conditional on a true effect of a given size and an accepted rate of false-positive results, is called power (Cohen, 1962, 1988; Neyman, 1937, 1942). Power should be a central concern for statistical hypothesis testers with finite resources and the journals that publish their results. For researchers, power calculations are useful when designing studies to optimize the use of resources and especially for avoiding studies that have a low probability of producing informative results. For journals, the range of effects a study has the

McKay (bradmckay8@gmail.com, https://orcid.org/0000-0002-7408-2323) and Carter (cartem11@mcmaster.ca, https://orcid.org/0000-0002-0675-4271) are corresponding authors.

power to rule out is an indication of how potentially informative that study was a priori. Unfortunately, power analyses can also be misleading. Power can be seriously overestimated by the wrong parameters—many of which are entirely based on the researcher's judgment. To conduct a power analysis at least four parameters are required: the design of the study, the size of the assumed effect, the frequency of false positives, and the frequency of false negatives. Although each of these specifications should be justified (Lakens, 2022b; Lakens et al., 2018), researchers often rely on conventions. For example, false-positive and false-negative rates have conventionally been set at 5% and 20%, respectively (Cohen, 1988). Many researchers and journals may consider false-negative rates of 10% or 5% more appropriate, but this consideration should be made thoughtfully (see Lakens, 2022b, for a discussion).

Standardized effect sizes also have conventional benchmarks that researchers may rely on when designing studies. Recent metascience suggests doing so is likely to result in underpowered research designs in practice (Lovakov & Agadullina, 2021). Instead of relying on benchmarks, some researchers may base their effect size target on a previous study or the results of a pilot study. However, large multilab replication studies have revealed that original studies may overestimate the true effect of an independent variable by 100%–400% (Klein et al., 2018; Open Science Collaboration, 2015). Pilot studies are often even less helpful, as they tend to be smaller than published experiments so their estimates are even more imprecise (Albers & Lakens, 2018; Kraemer et al., 2006; Lakens & Evers, 2014). When available, meta-analyses provide an effect size estimate based on the aggregation of available data. However, selective reporting of results can distort meta-analytic estimates, and it can be difficult to correct for reporting bias (Carter et al., 2019; Thornton & Lee, 2000). Nevertheless, estimates that have been corrected for reporting bias are more accurate than naïve random effects estimates and should be used when available (Carter et al., 2019).

A better strategy for choosing the effect size for an a priori power analysis does not rely on mean estimates, and instead the researcher specifies their smallest effect size of interest (Lakens, 2022b). If a researcher targeting 80% power estimates an effect is $d = 0.5$ but would still be interested if it was $d = 0.2$, they will miss their smallest effect size of interest 80% of the time. Instead of powering for the expected effect, researchers that power for their smallest effect size of interest guarantee their study design will not be underpowered for interesting effects. Researchers can extend this strategy to maximize the informativeness of their studies by making one-tailed predictions with 95% power. In this situation, null results are significantly smaller than the smallest effect size of interest. Studies designed this way may help prevent distortion from selection bias as both positive and negative results can be interpreted as significant.

Given the potential for power analyses to enhance the inferential value of studies and the myriad suboptimal strategies that may be employed, we chose to investigate the proportion of recent studies where motor behavior scientists reported a power analysis and their justification for their selected effect size. We focused on motor behavior research as recent meta-analyses have reported evidence of both underpowered research and substantial reporting bias in motor learning and sports science (Lohse et al., 2016; 2022a, b; Mesquida et al., 2022). For example, a meta-analysis of the self-controlled motor learning literature estimated the average power of all studies conducted was 6%, while 48% of studies reported significant results on the focal measure (McKay et al., 2022b). Other studies have estimated average power ranging from 20% (McKay et al., 2022a) to 50% (Mesquida et al., 2022), with significant indications of reporting bias. The combination of low power and significance-based selective reporting is pernicious to the accumulation of scientific evidence. Statistically significant results in studies with low power are likely to substantially overestimate the effect of the independent variable. When power dips below 10%, significant results in the wrong direction become increasingly likely (Gelman & Carlin, 2014).

If motor behavior research does not currently report power analyses—especially for the smallest effect size of interest—then future adoption of these best practices could potentially address issues of low power and selective reporting. Investigating this possibility, we examined the prevalence of a priori power analyses in three motor behavior journals, the justifications used for effect size assumptions, and their association with studies finding positive results. The goal of this study was descriptive. Our main purpose was simply to understand the current use of power analyses in the motor behavior literature. However, we did posit several exploratory hypotheses. We predicted that studies with a power analysis would have a different rate of positive results from studies without a power analysis. However, due to potential selection effects, we did not speculate about the direction of this difference a priori. We predicted that some justifications would differ in the frequency of positive results, with pilot studies being especially unsuccessful. We also predicted that differences in targeted power would be associated with different positivity rates given the primary function of a power analysis. Finally, we predicted that there would be a difference in the sample size obtained by studies that conducted a power analysis compared with those that did not, again without speculating about the direction.

Methods

Our design and analysis plan were preregistered after piloting our methods on a subsample of 40 articles. The preregistration, data, and code for this study are available using either of the following links: https://osf.io/wsdpv/ or https://github.com/cartermaclab/proj_power-motor-behaviour.

Power

Calculating a priori power for this study required estimating the final sample size and proportion of studies that would include a power analysis. Based on the pilot sample of articles, we estimated that the total number of studies we would extract would be approximately 500. The actual number was 636. We reasoned that if 10% of those studies included a power analysis, we would have 50 studies with power analyses and 450 studies without power analyses in our sample. The actual numbers were 13%, 85, and 551. Based on our rough estimates, we conducted simulations to estimate our power to detect differences in positive result rates of various plausible sizes. We based our expected positive results rate in studies without power analyses on estimates for psychology overall at 91.5% (Fanelli, 2010). We observed that, if our estimates were accurate, we would have 90% power to detect a difference of 16.5%, or a positive result rate of 75% in experiments with power analyses. Similarly, we estimated we would have 80% power to identify a positive result rate of 77.7% as significantly different. Unfortunately, if our estimated group sizes were accurate, we would have had low power (32%) to detect our smallest effect size of interest (6%). Given the actual sample sizes we observed, we had even greater power than planned to observe the effects we considered.

Sample

All articles published in the *Journal of Motor Learning and Development*, *Human Movement Science*, and the *Journal of Motor Behavior* between January 2019 and June 2021 were uploaded to Covidence systematic review software and screened for inclusion (Figure 1). In total, 704 articles were reviewed. To be

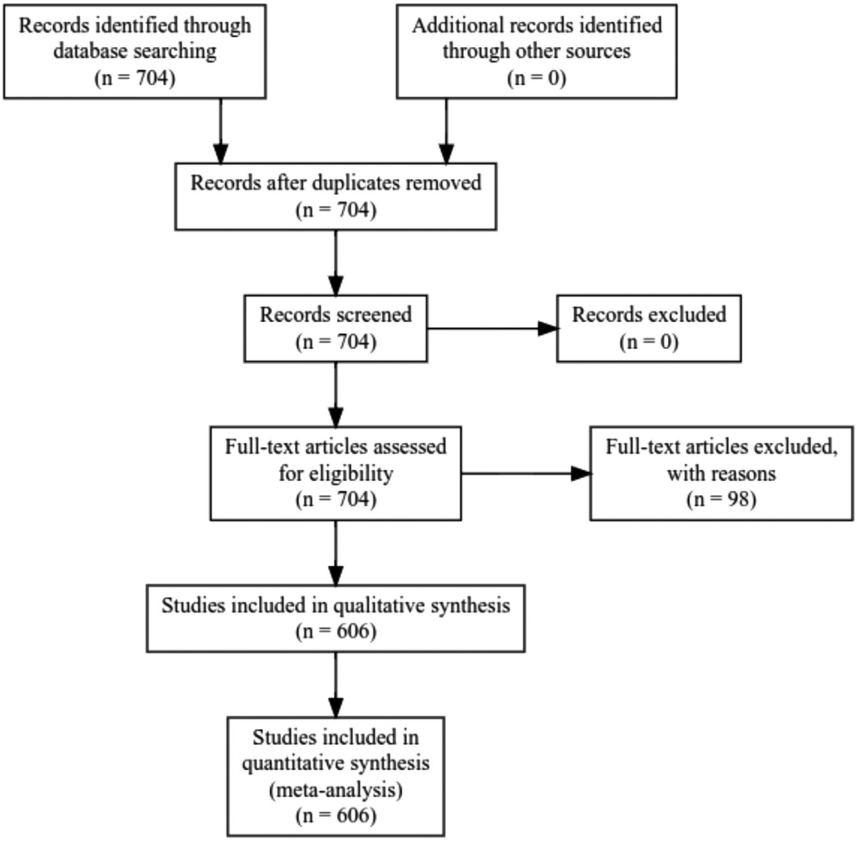
Records identified through database searching (n = 704)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 704)
Records screened (n = 704)
Records excluded (n = 0)
Full-text articles assessed for eligibility (n = 704)
Full-text articles excluded, with reasons (n = 98)
Studies included in qualitative synthesis (n = 606)
Studies included in quantitative synthesis (meta-analysis) (n = 606)

Figure 1 — PRISMA flow diagram.

included in the analysis, studies were required to meet the following criteria: (a) must be a primary study; (b) must test a hypothesis, including the null hypothesis; (c) there must be sufficient information available to adequately evaluate the criteria; and (d) we must have access to the full text. From the original 704 articles, 607 articles included at least one study that met the inclusion criteria and were included in the final analysis. Ninety-seven articles were excluded from the analysis for the following reasons: (a) The studies were not primarily quantitative (63 studies), and (b) the studies made no hypothesis (27 studies), or there was insufficient information or a faulty digital object identifier (DOI) to assess the paper (seven). The 607 included articles contributed a total of 636 eligible studies to the analysis.

Procedures

Data extraction was conducted by an extraction team of eight researchers. Two independent researchers evaluated each article in Covidence, with a third researcher resolving any conflicts. In situations where a member of the extraction team encountered a challenging item, the member flagged the study on Covidence for the items to be extracted and the consensus decision to be made by the first author ($N = 40$). We extracted data for 11 items, which are outlined in Table 1. For Item 5, determining whether the authors of a study concluded the results supported their hypothesis involved two steps. First, the primary hypothesis of a study was identified, either because the authors specified the hypothesis as primary or it was the first independent hypothesis reported. When hypotheses were listed with

Table 1 Elements of the Data Extraction Process and the Corresponding Action the Researchers Performed

Item	Action
1. Did the study meet the inclusion criteria?	Yes or no, and provide reason
2. Did the authors report a power analysis?	Yes or no
3. Hypothesis quote	Copy pasted quote of the hypotheses
4. Results quote	Copy pasted quote of the results interpretation
5. Did the authors conclude support for any of the main hypotheses?	Yes or no
6. Sample size	Calculate average per group
7. Power analysis effect type	Select from a list
8. Power analysis effect estimate	Report the effect size used for the analysis
9. Power analysis effect converted to Cohen *d*	Perform conversion whenever possible
10. Effect size justification	Select from a list
11. Power estimate from the power analysis	Report value

multiple components, support for any component was considered support for the hypothesis. Any hypothesis explicitly labeled as secondary was not considered. Second, the interpretation of the results by the authors was examined. We coded support for hypotheses based on the interpretations in each paper, not based on our own criteria. Thus, if the authors predicted no effect of an independent variable, observed null results, and then concluded the results supported their hypothesis, we coded this as support for the hypothesis.

Statistical Analysis

To evaluate the overall prevalence of power analyses in the sampled literature, we calculated the percentage of all studies in our sample that conducted a power analysis:

$$\frac{\text{Studies with power analysis}}{\text{Studies with power analysis} + \text{studies without power analysis}} \times 100$$

We used a two-sided proportion test to assess whether the rate of positive results in studies with a power analysis was significantly different than in studies without a power analysis. We also tested whether the difference in positive result rates was statistically smaller than our smallest effect size of interest (6%) using an equivalence test for proportions.

We calculated the percentage of studies that conducted a power analysis with (a) each effect justification and (b) each power target. A two-sided, six sample proportion test was conducted to test whether at least two different effect size justifications in power analyses led to different rates of positive results. A two-sided, 11-sample proportion test was conducted to test whether at least two power targets resulted in a different rate of positive results. Last, we conducted a two-tailed Welch's *t* test to determine whether studies with power analyses had different sample sizes compared with studies without power analyses. Given the data were highly skewed, we also conducted a sensitivity analysis using a shift function (Rousselet et al., 2017; Rousselet & Wilcox, 2020; Wilcox, 2021).

Statistical tests were conducted using R (version 4.1.2; R Core Team, 2021) and the R packages *diagramme* (Iannone, 2016), *extrafont* (Version 0.18; Chang, 2022), *kableExtra* (version 1.3.4; Zhu, 2021), *papaja* (version 0.1.0.9999; Aust & Barth, 2020), *prisma* (Jack, 2019), *rcolorbrewer* (Neuwirth, 2022), *renv* (version 0.15.5; Ushey, 2022), *rogme* (version 0.2.1; Rousselet et al., 2017), *rsvg* (version 2.3.1; Ooms, 2022), *tidyverse* (version 1.3.1; Wickham et al., 2019), *tinylabels* (version 0.2.3; Barth, 2022), *toster* (Lakens, 2017), and *waffle* (version 1.0.1; Rudis & Gandy, 2019) were used in this project.

Results

Proportion of Studies With a Power Analysis

Out of 636 total studies, 85 included a power analysis and 551 did not. Therefore, 13% of all studies sampled reported the results of a power analysis.

Difference in Positivity Rates Between Studies With and Without a Power Analysis

As shown in Figure 2, studies that did not include a power analysis reported finding support for their primary hypothesis 85% of the time (95% confidence interval [82%, 88%]), while studies that included a power analysis found support 76% of the time (95% confidence interval [66%, 85%]). The difference in positivity rates was not statistically significant ($\chi^2 = 3.47$, $df = 1$, $p = .06$). The difference is positivity rates was not significantly smaller than our smallest effect size of interest ($Z = .546$, $p = .71$).

Justifications for Effect Sizes Used in Power Analyses

The most common justification reported in our sample was to base the expected effect size on a previous study ($n = 37$), accounting for 44% of all justifications. The second most common justification was to provide no justification at all ($n = 20$), which occurred in 24% of studies that included a power analysis. Cohen’s benchmarks for small, medium, and large effects ($n = 19$) were used in 22% of studies. Pilot studies ($n = 9$) were used as justification in 11% of the sample.

Power Levels Targeted in Power Analyses

The most frequently targeted power was 80%, which was chosen in 65% of studies with a power analysis ($n = 55$). The next most common power target was 95%, accounting for 14% of all power targets ($n = 12$); followed by 90% power, occurring in 11% of power analyses ($n = 9$). Two studies did not state their targeted power, and several idiosyncratic power targets (96.7%, 96%, 95.33%, 85%, 75%, 70%, and 20%) were reported only once.

Difference in Positivity Rates as a Function of Effect Size Justification

Figure 3 illustrates the proportion of positive results for the four effect size justifications we found in our sample. Positivity rates were 100% for pilot study justification (9/9), 90% for studies with no justification (18/20), 68% for studies based on benchmarks (13/19), and 68% for studies based on previous studies (25/37). There was no significant difference between the positivity rates of any two justifications ($\chi^2 = 7.12$, $df = 3$, $p = .068$).

Difference in Positivity Rates as a Function of Target Power

Studies that targeted 80% power found support for their hypotheses 68% of the time (38/56). Studies that aimed for 90% power found support 100% of the time (8/8), and studies that aimed for 95% power found support 75% of the time (9/12). All studies that set an idiosyncratic power target or no target at all found support for their hypotheses (10/10). There was no significant difference between target power values ($\chi^2 = 7.22$, $df = 10$, $p = .70$).

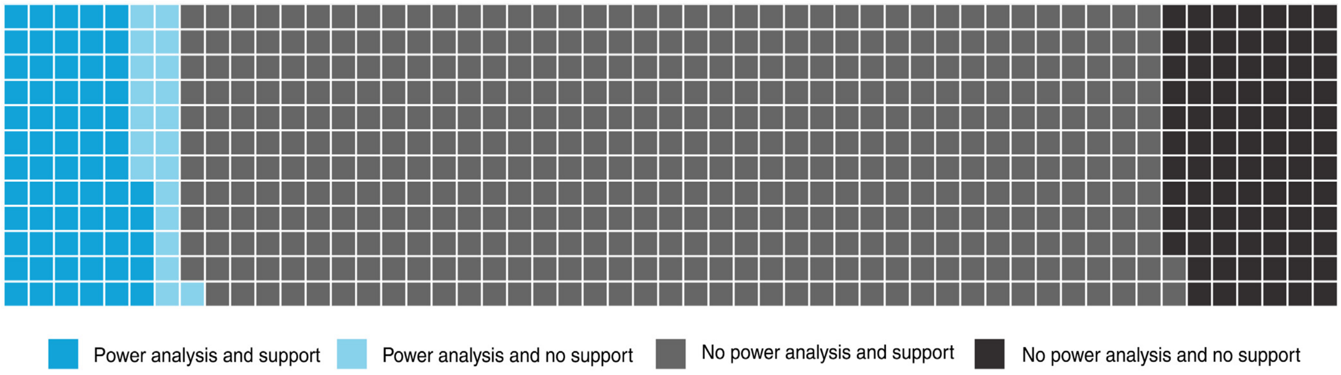

Figure 2 — Proportion of studies with (blue) and without (gray) power analyses and whether the authors concluded support for their primary hypotheses. Each square represents a single study in our sample. The majority of studies in our sample did not include a power analysis. The most common combination was “No Power-Analysis & Support” (light gray), while “Power-Analysis & No Support” (light blue) was the least common combination. For interpretation of the references to color in this figure, the reader is referred to the online version of this article.

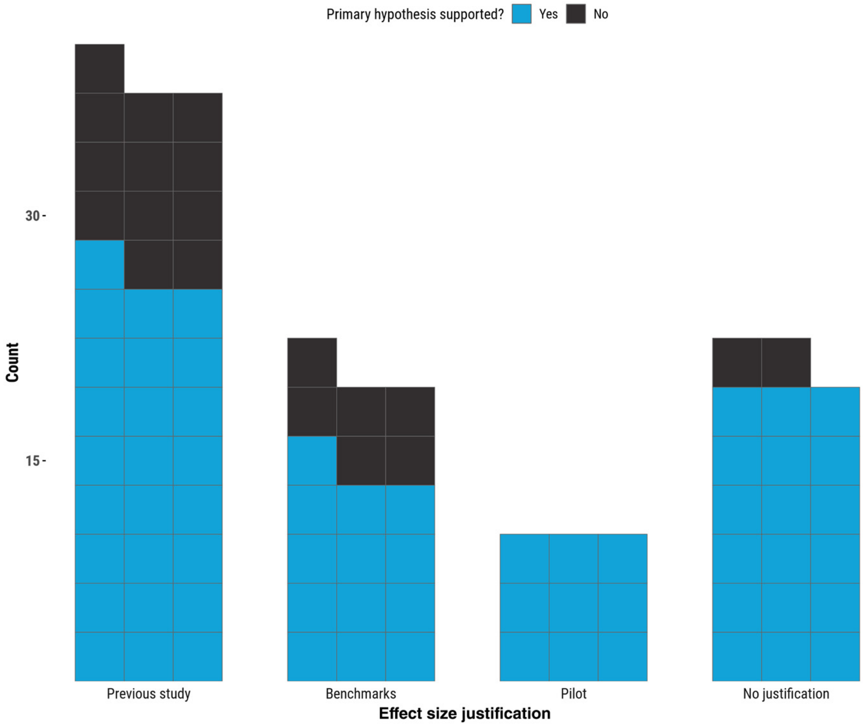

Figure 3 — Proportion of studies where authors concluded support (blue) or no support (black) for their primary hypotheses as a function of their effect size justification. Each square represents a single study. Of the list of possible effect size justifications, we only found data for four justifications. For interpretation of the references to color in this figure, the reader is referred to the online version of this article.

Difference in Sample Size Between Studies With and Without a Power Analysis

Studies that included a power analysis had significantly smaller mean sample sizes ($M = 21.91$) than studies that did not include a power analysis ($M = 40.98$), $t(624.59) = 3.43$, $p = .001$. However, sample sizes were highly skewed—especially among studies without a power analysis. The median sample for those studies ($Mdn = 17.5$) was similar to the sample sizes of studies with a power analysis ($Mdn = 16$). We conducted a shift function as a sensitivity analysis, and the results indicated no significant difference in sample size between studies with and without power analyses at any decile of their distributions.

Discussion

The purpose of this study was to investigate how frequently power analyses are reported in motor behavior articles, the justifications used for the effect size estimates, and the relationship between reporting power analyses and reporting positive results. We reviewed every article published in the *Journal of Motor Behavior*, the *Journal of Motor Learning and Development*, and *Human Movement Science* between January 2019 and June 2021 and identified 636 studies that tested a hypothesis. Of those 636 studies, 85 of them included a power analysis (13%). The rate of positive results was 85% overall and 76% when a power analysis was reported. The positive result rate was not significantly different between various effect size justifications or power targets.

Our results suggest that motor behavior research has not yet widely adopted power analyses to inform study design. When power analyses were reported, we observed a range of suboptimal effect size justifications. For example, 63% of studies that reported a power analysis based their effect size assumption on a previous study, a pilot study, or on effect size benchmarks. Another 24% of studies provided no justification at all. Each of these justifications (or lack thereof) is undesirable for different reasons. Previous studies—and especially pilot studies—are likely to provide exaggerated or noisy estimates of the unknown true effect. Effect size benchmarks may not match well the typical effect sizes one may find in their respective research area (Lovakov & Agadullina, 2021). Furthermore, Cohen's (1988) benchmarks differ depending on which analysis is used in a power analysis. A medium effect is over twice as large for a multiple regression analysis as compared with a *t* test (see Correll et al., 2020, for a discussion with additional examples). Not one study in the sample performed a power analysis based on their smallest effect size of interest. Power analyses can be an effective tool for researchers to ensure their studies are not underpowered, but to do so, the smallest effects of interest need to be targeted.

The rate of positive results observed in this study suggests that positive findings are overrepresented in the motor behavior literature. While the studies in our sample reported positive results 84% of the time, the median per group sample size was ~17, which would provide 84% power to detect $d = 1.05$ with an independent *t* test or $d = 0.76$ with a dependent *t* test. In comparison, the most optimistic estimates for well-known motor behavior phenomena are much smaller. Examples include the effect of feedback frequency on motor performance ($d = 0.19$; McKay et al., 2022a), self-controlled practice on retention performance ($d = 0.54$; McKay et al., 2022b), enhanced expectancies on retention performance ($d = 0.54$; Bacelar et al., 2021), and external focus of attention on retention performance ($d = 0.58$; Chua et al., 2021). Estimates for the true effects of these phenomena that have been corrected for reporting bias are markedly smaller, ranging from $d = 0$ to $d = 0.25$. Assuming the average effects investigated by the studies in our sample were similar to the optimistic estimates for other motor behavior effects, it is likely this literature was underpowered on average and potentially heavily censored.

All three journals that we sampled from either explicitly mention power in their instructions for authors (*Human Movement Science*), or reference Journal Article Reporting Standards (JARS) (*Journal of Motor Learning and Development*) or Consolidated Standards of Reporting Trials (CONSORT) (*Journal of Motor Behavior)* reporting standards, both of which include power analyses. Therefore, adoption of power analyses targeting interesting effects does not require a policy shift, simply the enforcement of current guidelines. Since no studies in this sample powered for the smallest effect size of interest, if *Human Movement*

Science, the *Journal of Motor Behavior*, and/or the *Journal of Motor Learning and Development* enforce their existing guidelines, then their future publications will look dramatically different. We believe this is a promising path forward to increase the reliability of motor behavior research and the evidence-based recommendations for coaching and rehabilitation.

Limitations

The specific designs and test statistics from the studies in our sample were not extracted, so we cannot calculate the estimated average power of the sample. This also complicates the interpretation of sample size differences among studies with and without power analyses. For example, if studies that used within-subjects designs were also more likely to conduct a power analysis, it would make sense that those studies would have smaller samples overall. Within-subjects designs are substantially more powerful than between-subjects, so all other things being equal, studies with within-subjects designs require less participants to be adequately powered.

We do not differentiate between partial and full support for hypotheses, nor did we code for whether the hypothesis was directional, nondirectional, or if the null hypothesis was framed as the primary hypothesis in the study. As such, we must be cautious not to regard positivity rate as a direct analog for implied power. There were studies that predicted no difference between experimental conditions, failed to reject the null hypothesis, and then interpreted the result as supporting their primary hypothesis. While this approach to hypothesis testing is problematic, our goal with this study was to describe the proportions of positive results and power analyses, not to critique the specific methods employed in each study.

Conclusion

Our results suggest that power analyses targeting the smallest effect size of interest (Lakens, 2022a) have the potential to change the state of the motor behavior literature. Hypothesis tests are the norm in this space, yet power calculations targeting interesting effects are not. It is logical for researchers to plan studies with a high probability of producing informative results, and it is consistent with current reporting standards (Appelbaum et al., 2018). Given the recent concern about the reliability of established motor behavior phenomena Mesquida et al. (2022), we believe power analyses have an important role to play in increasing the credibility of our field.

Acknowledgments

All authors thank Aaron Schwarz and Abigail Morgan for their help with data extraction. Open science practices: The preregistration, data, and scripts can be accessed using either of the following links: https://osf.io/wsdpv/ or https://github.com/cartermaclab/proj_power-motor-behaviour. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (RGPIN-2018-05589; MJC) and McMaster University (MJC).

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74,* 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Appelbaum, M., Cooper, H., Kline, R.B., Mayo-Wilson, E., Nezu, A.M., & Rao, S.M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3. https://doi.org/10.1037/amp0000191

Aust, F., & Barth, M. (2020). papaja: Prepare reproducible APA journal articles with R Markdown. https://github.com/crsh/papaja

Bacelar, M.F.B., Parma, J.O., Murrah, W.M., & Miller, M.W. (2021). Meta-analyzing enhanced expectancies on motor learning: Positive effects but methodological concerns. *International Review of Sport and Exercise Psychology,* Advanced online publication. https://doi.org/10.1080/1750984X.2022.2042839

Barth, M. (2022). tinylabels: Lightweight variable labels. https://cran.r-project.org/package=tinylabels

Carter, E.C., Schönbrodt, F.D., Gervais, W.M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chang, W. (2022). Extrafont: Tools for using fonts. https://CRAN.R-project.org/package=extrafont

Chua, L.-K., Jimenez-Diaz, J., Lewthwaite, R., Kim, T., & Wulf, G. (2021). Superiority of external attentional focus for motor performance and learning: Systematic reviews and meta-analyses. *Psychological Bulletin, 147*(6), 618–645. https://doi.org/10.1037/bul0000335

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Correll, J., Mellinger, C., McClelland, G.H., & Judd, C.M. (2020). Avoid Cohen's "small," "medium," and "large" for power analysis. *Trends in Cognitive Sciences, 24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*(4), Article e10068. https://doi.org/10.1371/journal.pone.0010068

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Iannone, R. (2016). DiagrammeRsvg: Export DiagrammeR graphviz graphs as SVG. https://CRAN.R-project.org/package=DiagrammeRsvg

Jack, O.W. (2019). PRISMAstatement: Plot flow charts according to the "PRISMA" statement. https://CRAN.R-project.org/package=PRISMAstatement

Klein, R.A., Vianello, M., Hasselman, F., Adams, B.G., Adams, R.B., Alper, S., Aveyard, M., Axt, J.R., Babalola, M.T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M.J., Berry, D.R., Bialobrzeska, O., Binan, E.D., Bocian, K., Brandt, M.J., Busching, R., . . . Nosek, B.A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kraemer, H.C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J.A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals.

Archives of General Psychiatry, 63(5), 484–489. https://doi.org/10.1001/archpsyc.63.5.484

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 1,* 1–8. https://doi.org/10.1177/1948550617697177

Lakens, D. (2022a). Improving your statistical inferences. https://lakens.github.io/statistical_inferences/

Lakens, D. (2022b). Sample size justification. *Collabra: Psychology, 8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., et al. (2018). Justify your alpha. *Nature Human Behaviour, 2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., & Evers, E.R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278–292. https://doi.org/10.1177/1745691614528520

Lohse, K., Buchanan, T., & Miller, M. (2016). Underpowered and overworked: Problems with data analysis in motor learning studies. *Journal of Motor Learning and Development, 4*(1), 37–58. https://doi.org/10.1123/jmld.2015-0010

Lovakov, A., & Agadullina, E.R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology, 51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

McKay, B., Hussien, J., Vinh, M.-A., Mir-Orefice, A., Brooks, H., & Ste-Marie, D.M. (2022a). Meta-analysis of the reduced relative feedback frequency effect on motor learning and performance. *Psychology of Sport and Exercise, 61,* Article 102165. https://doi.org/10.1016/j.psychsport.2022.102165

McKay, B., Yantha, Z., Hussien, J., Carter, M.J., & Ste-Marie, D. (2022b). Meta-analytic findings in the self-controlled motor learning literature: Underpowered, biased, and lacking evidential value. *Meta-Psychology, 6,* Article MP.2021.2803. https://doi.org/10.15626/MP.2021.2803

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports science: A narrative review of selected methodological issues in the field. *SportRxiv*. https://sportrxiv.org/index.php/server/preprint/view/127

Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes. https://CRAN.R-project.org/package=RColorBrewer

Neyman, J. (1937). "Smooth test" for goodness of fit. *Scandinavian Actuarial Journal, 1937*(3–4), 149–199. https://doi.org/10.1080/03461238.1937.10404821

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society, 105*(4), 292–327. https://doi.org/10.2307/2980436

Ooms, J. (2022). Rsvg: Render SVG images into PDF, PNG, (encapsulated) PostScript, or bitmap arrays. https://CRAN.R-project.org/package=rsvg

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rousselet, G.A., Pernet, C.R., & Wilcox, R.R. (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience, 46*(2), 1738–1748. https://doi.org/10.1111/ejn.13610

Rousselet, G.A., & Wilcox, R.R. (2020). Reaction times and other skewed distributions: Problems with the mean and the median. *Meta-Psychology, 4,* 1–39. https://doi.org/10.1101/383935

Rudis, B., & Gandy, D. (2019). Waffle: Create waffle chart visualizations. https://gitlab.com/hrbrmstr/waffle

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology, 53*(2), 207–216. https://doi.org/10.1016/S0895-4356(99)00161-4

Ushey, K. (2022). Renv: Project environments. https://CRAN.R-project.org/package=renv

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Wilcox, R.R. (2021). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic press.

Zhu, H. (2021). kableExtra: Construct complex table with ‘kable’ and pipe syntax. https://CRAN.R-project.org/package=kableExtra