Meta-Psychology, 2023, vol 7, MP.2021.2938
https://doi.org/10.15626/MP.2021.2938
Article type: Commentary
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Nick Brown
Reviewed by: Borgert, N., Lawrence, J., Anaya, J.
Analysis reproduced by: Lucija Batinović
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/BJGAE

A critical re-analysis of six implicit learning papers

Brad McKay
McMaster University

Michael J. Carter
McMaster University

Abstract

We present a critical re-analysis of six implicit learning papers published by the same authors between 2010 and 2021. We calculated effect sizes for each pairwise comparison reported in the papers using the data published in each article. We further identified mathematically impossible data reported in multiple papers, either with deductive logic or by conducting a GRIMMER analysis of reported means and standard deviations. We found the pairwise effect sizes were implausible in all six articles in question, with Cohen's d values often exceeding 100 and sometimes exceeding 1000. In contrast, the largest effect size observed in a million simulated experiments with a true effect of $d = 3$ was $d = 6.6$. Impossible statistics were reported in four out of the six articles. Reported test statistics and η^2 values were also implausible, with several $\eta^2 = .99$ and even $\eta^2 = 1.0$ for between-subjects main effects. The results reported in the six articles in question are unreliable. Many of the problems we identified could be spotted without further analysis.

Keywords: Metascience, GRIMMER, Effect sizes, Perceptual motor learning

Introduction

Statistical reporting errors may commonly occur in psychology articles (Brown & Heathers, 2017; Nuijten et al., 2016) and such errors are often consistent with hypothesized results (Bakker & Wicherts, 2011). When the primary conclusions in research articles depend on reporting errors, replicability is unlikely and future research may be wasted if researchers attempt to build on the erroneously reported results (Munafò et al., 2017). In this paper, we scrutinize six papers published by the same two authors,[1] where the authors report a high number of erroneous or implausible data on which their primary conclusions depend. We first became aware of the Lola and Tzetzis (2021) paper when the paper was highlighted in a social media post (Gray, 2021). During an initial read through by one of us (BM), a number of reporting and statistical issues were noticed. The paper also referenced past research published by these authors. Given our concerns over the issues found in the Lola and Tzetzis (2021) paper, we deemed it necessary to examine these other papers. The data irregularities we found are similar across the target articles and at times even include repeated values (e.g., F-statistics) across multiple papers. Regardless of the conclusion one reaches with respect to the mechanism behind these errors, it is our contention that the results reported in these papers are unreliable and that the respective journals in which the papers are published should take cor-

[1] One of the six papers had a third author and one had a third and fourth author.

rective actions.[2] Below, we outline our causes for concern and the overarching issues we found across the six articles in question.

The articles in question

We reanalyzed six articles by Afroditi Lola, George Tzetzis, and their colleagues. In all experiments, the authors investigated the effects of implicit and explicit instructions on perceptual and motor learning. All experiments sampled young females who were enrolled in a volleyball camp (see Table 1). Our reanalysis of the target articles evaluated the plausibility of the reported means, standard deviations, and test statistics. We will refer to the six articles throughout this paper using the following numbering system based on reverse chronological order:

1. Lola, A.C., Giatsis, G., Pérez-Turpin, J.A., & Tzetzis, G.C. (2021). The influence of analogies on the development of selective attention in novices in normal or stressful conditions. *Journal of Human Sport and Exercise*. https://doi.org/10.14198/jhse.2023.181.12
2. Lola, A.C., & Tzetzis, G.C. (2021). The effect of explicit, implicit and analogy instruction on decision making skill for novices, under stress. *International Journal of Sport and Exercise Psychology*, 1-21. https://doi.org/10.1080/1612197X.2021.1877325
3. Lola, A.C., & Tzetzis, G.C. (2020). Analogy versus explicit and implicit learning of a volleyball skill for novices: The effect on motor performance and self-efficacy. *Journal of Physical Education and Sport*, 20(5), 2478-2486. https://doi.org/10.7752/jpes.2020.05339
4. Tzetzis, G.C., & Lola, A.C. (2015). The effect of analogy, implicit, and explicit learning on anticipation in volleyball serving. *International Journal of Sport Psychology*, 46(2), 152-166. https://doi.org/10.7352/IJSP.2015.46.152
5. Lola, A.C., Tzetzis, G.C., & Zetou, H. (2012). The effect of implicit and explicit practice in the development of decision making in volleyball serving. *Perceptual and Motor Skills*, 114(2), 665-678. https://doi.org/10.2466/05.23.25.PMS.114.2.665-678
6. Tzetzis, G.C., & Lola, C.A. (2010). The role of implicit, explicit instruction and their combination in learning anticipation skill, under normal and stress conditions. *International Journal of Sport Sciences and Physical Education*, 1, 54-59.[3]

Although there were some differences between the reported experiments in the target articles, there were many methodological commonalities that can be summarized. All six articles involved female children learning a volleyball skill as part of a volleyball camp. In each case, the participants were reported to have minimal experience (i.e., were described as novices) with the task at hand. The purpose of all six experiments was to evaluate perceptual or motor learning differences as a function of the type of instruction received during practice. Each experiment included a pre-test, an acquisition (i.e., practice) phase involving 12 sessions spaced over four weeks, and a post-test. A high stress test was also included in Articles 1, 2, and 6.

In Articles 1-4, the groups differed with respect to the type of instruction received: implicit, explicit, or analogy. In Articles 5 and 6, a sequential group (see below for description) replaced the analogy group. All six experiments also included a control group that did not practice the task. Implicit instruction did not contain any explicit information for how to perform the task and the learners were asked to perform a distracting task like counting backwards while practicing to prevent them from acquiring declarative rules for performance. In contrast, explicit instruction consisted of direct verbal instructions for performing the task. Analogy instruction was considered a type of implicit instruction wherein an analogy or metaphor was provided to the learner. For example, "Imagine that the opponents' surface is covered with water. Send the ball where there is more water and no opponents at the court." (Lola & Tzetzis, 2021, p. 9). Sequential instruction involved receiving explicit instruction for the first half of training followed by implicit instruction for the second half of training. Across experiments, the authors predicted that implicit forms of instruction—implicit, analogy, and sequential—would be more effective than explicit instruction for motor and perceptual learning. This advantage was also predicted to be greater when testing was conducted in a high stress situation. In Article 2 for instance, high stress was induced by falsely telling participants that the best performers would be selected for a draft to the national team. Further, it was predicted that analogy or sequential instruction would offer improvements relative to implicit instruction.

The primary outcome measures used in these experiments were reaction time (Articles 1, 2, 4, 5, and 6),

[2] We contacted the journal editors for Articles 2-6 on Sept 22 2021 and for Article 1 on Jan 21 2022. All editors indicated their intention to further investigate these issues and/or take corrective actions.

[3] This journal has been identified as a potential predatory journal. We were unable to find an online version of this article on the journal's webpage and interestingly, the earliest issue on the webpage is from 2016. We were only able to find an online version on ResearchGate (https://www.researchgate.net/profile/Angela-Calder/publication/234000504_The_scientific_basis_for_recovery_training_practices_in_sport/links/5428fff80cf26120b7b574ad/The-scientific-basis-for-recovery-training-practices-in-sport.pdf with the target article beginning on page 57).

Table 1

Participant demographics in each of the target articles.

Target article	Sample size and participant details
Article 1: Lola et al. (2021)	60 females, age range: 11 to 12 years (M_{age} and *SD* not reported)
Article 2: Lola & Tzetzis (2021)	60 females, age range: 10 to 11 years (M_{age} = 10.48, *SD* = 0.911)[a]
Article 3: Lola & Tzetzis (2020)	80 females, age range: 10 to 11 years (M_{age} = 10.48, *SD* = 0.911)[a]
Article 4: Tzetzis & Lola (2015)	60 females, age range: 9 to 12 years (M_{age} = 10.48, *SD* = 0.91)[a]
Article 5: Lola et al. (2012)	60 females, age range: 10 to 12 years (M_{age} = 11.2, *SD* = 0.3)
Article 6: Tzetzis & Lola (2010)	48 females, age range: 12 to 13 years (M_{age} = 12.38, *SD* = 0.34)

Note. [a]Articles 2-4 report identical means and standard deviations for the age of their participants despite a different sample size in Article 3 from Articles 2 and 4, and a different age range in Article 4 from Articles 2 and 3.

response accuracy (Articles 1, 2, 4, and 5), and motor performance measured on a 4-point scale (Article 3). In addition, Articles 2 and 6 included a measure of state anxiety, the Competitive State Anxiety Inventory-2 (Tsorbatzoudis et al., 1998), and Article 3 had a measure of self-efficacy using a Likert scale. The number of explicit rules recalled was assessed in Articles 2, 4, 5, and 6.

Methods

None of the six articles in question included a link to a public repository where the data could be accessed. We first wrote (email sent February 10, 2021) the corresponding author of Article 2 and asked if they would be willing to share the data for this experiment. The authors' response was that the data could not be shared as they were not finished with their analyses and were in the process of running different tests (A. Lola, personal communication, February 12 2021). We followed up this email (sent February 12 2021) by asking whether they would instead be willing to share the data from any of Articles 3 to 6 as these were less recent, and presumably all planned analyses had been completed. After a 2 week period with no response, we followed up with a third email (sent February 26 2021) and reiterated our interest in obtaining their data from any of these articles. The authors' response was that they were unable to share data from any of these articles because in some cases they no longer had the data and in other cases they had plans to conduct further analyses (A. Lola, personal communication, March 2 2021).

Our first two requests did not include any indication about our concerns regarding the data irregularities. Subsequently, in a fourth email (sent April 12 2021) we outlined our concerns for each article[4] and once again reiterated our request to the authors to share any available data for any of the target articles. These requests were once again refused. The authors did address some specific concerns regarding Article 2, but for the most part only provided more general responses to our concerns. The authors admitted that some of the values reported in the other target articles were incorrect, but did not identify which values or articles. Despite this, the authors maintained that the data irregularities—identified in our email and described in this paper—do not impact the veracity of their analyses or conclusions (A. Lola, personal communication, April 22 2021). We illustrate below that the data *and* analyses reported in each of the articles reviewed are unreliable. Our extracted data and analysis scripts can be accessed using either of the following links: https://osf.io/raz6q/ or https://www.github.com/cartermaclab/comm_lola-tzetzis-data-irregularities.

Effect size calculations and simulations

Means and standard deviations were extracted from each article for all measures and time points that were reported. Cohen's *d* was calculated for each pairwise comparison using the R package `compute.es`. Consistent with the group sizes reported in Article 3, which had the largest groups among the target articles, we simulated data from two groups of $n = 20$. We ran simulations with true effect sizes of $d = .8$ and $d = 3$ one million times each and report the range of effect sizes observed in those simulations.

Mathematically impossible data and granularity analysis

In two of the articles in question, it was clear that some of the reported results were not mathematically possible based on the scale of measurement that was used. When outcomes were single item integers

[4] Excluding Article 1 (Lola et al., 2021) because we were not aware of it at the time as it had not yet been accepted for publication.

(a granularity of 1), such as the number of explicit rules recalled, we used a web application (http://www.prepubmed.org/grimmer_sd/) to conduct a granularity analysis (GRIMMER) of reported means and standard deviations (Anaya, 2016). GRIMMER builds off the original Granularity-Related Inconsistency of Means (GRIM) analysis (Brown & Heathers, 2017), which leveraged the fact that the means of granular data are also granular. Given a data set of size N and granularity G, only means of granularity G/N are possible. Thus, all possible means for data of a given G and N can be enumerated, and only means that match these possibilities are considered GRIM-consistent. The GRIMMER analysis extends this test by also evaluating whether mean-standard deviation pairs are possible. First, the GRIM analysis is conducted to determine if the mean is GRIM-consistent. Next, lower and upper bounds of the standard deviation are calculated based on how many decimals D are reported $(SD \pm [\frac{0.5}{10^D}])^2$. Then all possible variances between these bounds are enumerated, converted back to standard deviations, and rounded to the nearest D decimals. The reported standard deviation is checked for a match with any of these values. Finally, the mean-variance pair is compared to possible mean-variance pairings (the GRIMMER test handles sample sizes between 5 and 99). Using GRIMMER, it is possible to determine if specific mean and standard deviation pairs are possible for data of a given sample size. To be conservative, we specified that we did not know whether the standard deviation was calculated for the sample or population, nor whether ambiguous values were rounded up or down. Mean and standard deviation pairs that are mathematically possible are considered GRIMMER consistent, while mean and standard deviation pairs that are not mathematically possible are GRIMMER inconsistent.

Eta-squared

Each of the articles reported only omnibus test statistics and then reported post-hoc analyses with symbols demarcating significant and non-significant differences. In response to our expression of concern, the authors suggested that many of the issues were due to misprints in the articles. Specifically, they indicated that the reported means and standard deviations in their tables were incorrect and the root of the errors had to be from them outsourcing the formatting of their tables. The authors then insisted that despite these typographic errors, their discussion of the results and corresponding conclusions were still accurate (A. Lola, personal communication, April 22, 2021). However, the test statistics reported for many analyses were implausibly large and the authors often reported η^2 values associated with the omnibus test. Our examination of the reported η^2 values revealed that, as with the reported pairwise comparisons, many were implausibly large.

Results

Implausible effect sizes

Cohen's d is used to describe the standardized mean difference of an effect and values can range between 0 and infinity in both the negative and positive direction. We calculated absolute values so that all effects were positive. Cohen's d_s (Cohen, 1988) is the observed difference between group means divided by their pooled standard deviation (see Lakens, 2013, for a detailed discussion). Conventional benchmarks for small, medium, and large effects are $d = .2$, .5, and .8, respectively (Cohen, 1962); however, this *mindless* approach to effect size interpretation has been heavily discouraged (Correll et al., 2020; Field, 2016; Lakens, 2013; Thompson, 2007). Recently, an analysis of 6447 Cohen's d statistics extracted from social psychology meta-analyses observed median and 75th percentile Cohen's d values of .36 and .65, respectively—suggesting the conventional benchmarks may overestimate typical effects (Lovakov & Agadullina, 2021). In the field of motor learning, recent meta-analyses have found average effect sizes in the published literature of $d = .19$ (McKay, Hussien, et al., 2022), $d = .54$ (McKay, Yantha, et al., 2022), and $d = .71$ (Lohse et al., 2016).

To evaluate the maximum plausible Cohen's d statistics one might encounter from experiments similar to those reported in the target articles, we conducted two simulations that each consisted of one million experiments (see Figure 1). We set the true effect size at $d = .8$, the conventional benchmark for a "large" treatment effect, in the first simulation. The largest effect size observed from the one million simulated experiments was $d = 2.97$. In the second simulation, we set the true effect size at $d = 3$, an unrealistically large effect size that might rarely be encountered in the psychology and/or motor learning literature. The maximum effect size observed in the one million simulated experiments was $d = 6.6$.

In the context of the maximum values observed in our simulations, all six articles in question reported implausibly large effect sizes. The original table of summary statistics in Article 1 indicated that the smallest post-intervention difference in reaction times was $d = .64$. However, all other effects were larger than $d = 8.7$ and the largest effect was $d = 41$. The accuracy data also reflected improbably large post-intervention differences, with two-thirds of all comparisons showing effects larger than $d = 5$ and a largest effect of $d = 13.35$.

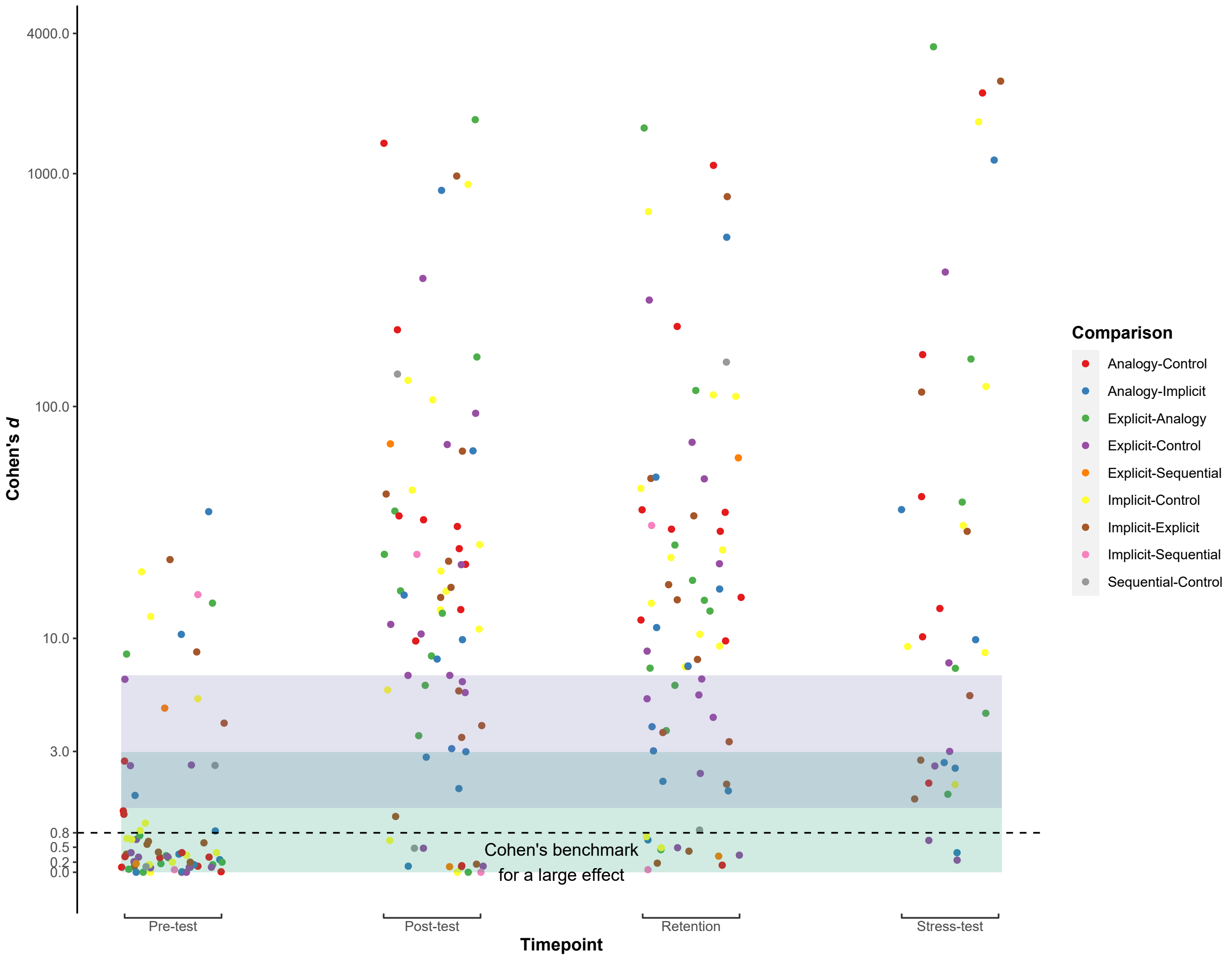

Figure 1. Absolute Cohen's *d* estimates from all articles except Article 6 plotted on a logarithmic scale. Only data from the original tables in Article 1 are included. All pairwise comparisons have been included for all dependent measures in each experiment. The range of observed values from a simulation of 1,000,000 experiments with a true effect of *d* = .8 is illustrated by shaded green and blue regions of the figure, reaching a maximum value of *d* = 2.97. The range of observed values from a simulation of 1,000,000 experiments with a true effect of *d* = 3 is illustrated by the shaded purple and blue regions of the figure, reaching a maximum value of *d* = 6.6.

However, a correction to the tables of summary statistics was published that included substantially smaller standard deviations than the original tables. While the updated data do imply smaller effect sizes, as we discuss below, they appear to be inconsistent with the reported analyses.

In Article 2, the smallest pre-test difference for reaction time was *d* = 1.29 and the largest pre-test difference was *d* = 35.32—although none of the groups were reported as significantly different in the article. The smallest post-intervention effect at any of the three time points was *d* = 286.42, while the largest effect was *d* = 3504.86. A similar picture emerges when analyzing the accuracy data. All the pre-test differences were improbably large (all *d*'s ≥ 2.52) despite being reported as not significantly different in the articles. Ten of the pairwise comparisons resulted in *d*'s ≥ 100 following treatment with the independent variables. The motor component data revealed post-treatment effect sizes ranging from *d* = 1.16 to *d* = 13.5.

In Article 3, post-intervention motor performance effect sizes ranged from *d* = 3.1 to *d* = 20.95. Similarly, post-intervention self-efficacy effect sizes ranged from *d* = 1.79 to *d* = 44.46. Likewise, in Article 4 post-intervention reaction time effect sizes ranged from *d* = 2.28 to *d* = 35.97. Continuing this pattern, post-intervention response accuracy effect sizes ranged from *d* = 5.84 to *d* = 29.7.

In Article 5, many response accuracy effect sizes were implausibly large beginning at pre-test, wherein effects ranged from d = 2.53 to d = 15.50. Nevertheless, all pre-test comparisons were reported as non-significant. Following intervention, the effect sizes ranged from d = 23.13 to d = 155.08. Relative to other reported effect sizes, those reported for reaction time were not implausibly large at any time point, ranging from d = 0 to d = .86. However, the authors reported an implausibly large effect size, η^2 = .94, for the 4 (Group) x 3 (Time) ANOVA. Further, despite only one pairwise comparison being statistically significant, all post-intervention comparisons were reported as being significant in the article.

In Article 6, the authors did not report means and standard deviations for most of the analyses. However, η^2 effect sizes were reported and these ranged from η^2 = .52 to η^2 = .98. These effect sizes are discussed further below. All the post-intervention effects reviewed above were directionally consistent with the researchers' expectations. The sometimes implausibly large pre-test effects were not expected, but also were not reported as significant.

Impossible data and granularity analysis

In Article 2, the Competitive State Anxiety Inventory-2 was used to assess the level of cognitive and somatic stress experienced by participants. Responses were measured on a Likert scale ranging from 1 to 4 with the data appearing to represent the average response per item. At each of the three low-stress time points, the means reported for all four groups ranged from 1.02 to 1.09. During the high-stress time point, the means ranged from 3.95 to 4.09. The means for two groups were reported as greater than 4, which is not possible given the maximum score on the Competitive State Anxiety Inventory-2 is 4.

In Article 3, participants were asked to receive a served volleyball and pass it to a target consisting of three concentric circles. Motor performance was measured based on where the pass landed, with three points awarded for a pass to the central circle on the target, two points for the middle circle, one point for the outermost circle, and zero points for a pass that missed the target.[5] Results were presented as average performance per trial and the analogy group was reported to have a mean score of 3.00 at retention (a perfect score) but with a standard deviation of .09. The perfect score was not a rounding error because the same group was reported to have a mean score of 2.99 with a standard deviation of .11 on the post-test. These data are not possible.

In Articles 2, 5, and 6, the authors reported means and standard deviations for the number of explicit rules recalled by participants following the intervention phase. As a single item analysis of integers these results were suitable for a GRIMMER analysis. In Article 2, the mean and standard deviation pairs were GRIMMER inconsistent for all four groups (Implicit: M = .73, SD = .35; Analogy: M = 1.03, SD = .25; Explicit: M = 4.8, SD = .78; Control: M = .67, SD = .48; n = 15). In Article 5, the mean and standard deviation pair was GRIMMER consistent for the explicit rules group (M = 4.8, SD = 1.78) and the implicit group (M = 2.3, SD = 1.3). The mean and standard deviation pairs for the remaining two groups were GRIMMER inconsistent (Sequential: M = 4.2, SD = 1.07; Control: M = 1.8, SD = .3; n = 15). In Article 6, the mean and standard deviation pairs were GRIMMER consistent for three of the four groups if the standard deviations were calculated for the population rather than the sample (Sequential: M = 4.2, SD = 1.07; Implicit: M = 2.3, SD = 1.3; Control: M = 1.4, SD = .9; n = 12). For two of the groups, they were consistent regardless of which method of calculating the standard deviation was used. However, the results for the explicit group were GRIMMER inconsistent (M = 4.8, SD = 1.78).[6]

Eta-squared

Eta-squared (η^2) is calculated by dividing the sum of squares for the effect by the total sum of squares. It can be interpreted as analogous to R^2 as it represents the total variation in the dependent measure that can be explained by a given main effect or interaction in an ANOVA (Lakens, 2013). Benchmarks have been suggested for small, medium, and large effect sizes as η^2 = .01, .06, and .14, respectively (Cohen, 1988). Importantly, if the main effect of instruction-type results in η^2 = .99, as was commonly reported in the target articles, this suggests that 99% of the total variability in the outcome measure can be explained by group assignment alone. Such a result is implausible.

Article 2 did not report η^2 values but had the largest pairwise effects and F-statistics of the five articles in question. Article 3 reported η^2 = .994, η^2 = .996, and η^2 = .996 for the Time, Group, and Time x Group effects on motor performance, respectively. Similarly, variance

[5] Independent of the issues we have raised, this approach to measuring motor performance has been shown to be inappropriate and flawed for this type of task (Fischman, 2015; Hancock et al., 1995; Reeve et al., 1994).

[6] You may have noticed that two of the same mean and standard deviation pairings (M = 4.8, SD = 1.78 and M = 4.2, SD = 1.07) were classified as GRIMMER inconsistent for one paper and consistent for the other. This is because of sample size differences (n = 15 and n = 12).

explained on the self-efficacy measure was η^2 = .995, η^2 = .994, η^2 = .997 for the Time, Group, and Time x Group effects, respectively. Article 4 also reported η^2 = .99 for all three effects on both response time and response accuracy measures.

Article 5 reported η^2 = 1.0 for the main effect of Time and the Time x Group interaction, as well as η^2 = .95 for the main effect of Group on the response time measure. Interestingly, the Time x Group interaction had the smallest reported significant F-statistic among the five articles in question. With respect to response accuracy, the reported effects were η^2 = .98, η^2 = .94, η^2 = .93 for the Time, Group, and Time x Group analyses, respectively. Article 6 reported η^2 = .66, η^2 = .52, η^2 = .72 for the Time, Group, and Time x Group analyses, respectively.

Other oddities

Although the means and standard deviations for the explicit rules analysis were only reported in three of the articles in question, analyses were reported in Articles 2, 4, 5, and 6. The reported test statistic in these four articles was F = 52.67, albeit with different degrees of freedom in Article 6 that reflected the different sample size in this experiment (48 versus 60 in the others). Articles 2-4 were published over a span of 6 years with reported samples sizes of 60 in Articles 2 and 4, and 80 in Article 3. Yet, the authors report identical means and standard deviations for the age of their participants in these three articles (see Table 1). We assumed that each article was based on different samples as none of the articles mentioned using any previously published data.

Article 1 was submitted to the *Journal of Human Sport and Exercise* following our correspondence with the authors and published online (Sept 3 2021) 11 days before we posted our preprint (Version 1). We were unaware of this article when posting the original preprint and it was not included in that version. However, when we became aware of this paper it was immediately apparent that data reported therein again reflected improbably large effect sizes. Further, in comparing the reaction time and accuracy means reported in Article 1 to those reported in Article 2, it appeared the data shared a remarkably similar pattern. To investigate this similarity further, we conducted a correlation analysis between the two data sets. The reaction time means for each group and time point were highly correlated between the two papers, r = .99, as were the accuracy means, r = .99.

New developments following September 28, 2021 update

After contacting the editors for Articles 2-5 on September 22, 2021 and updating our preprint on September 28th, at least two important developments transpired. First, there was a response from all four editors indicating an intention to investigate the issues we raised. The *International Journal of Sport and Exercise Psychology* is the only journal that has taken observable action to date, issuing an Expression of Concern regarding Article 2 on October 11th, 2021 (see https://doi.org/10.1080/1612197X.2021.1991102). The current journal editor where Article 4 was published included us in an email to the authors. We were also included in the authors' reply, wherein they again insisted that their analyses and conclusions remained valid. They offered an updated manuscript to the journal, but we are unaware of any decisions or further actions.

The second important development was that the tables in Article 1 were updated on October 8th, 2021. The new tables made changes to the standard deviation values for both the reaction time and accuracy measures. The standard deviations of the reaction time data were adjusted such that the decimal point was shifted one place to the right compared to the original version. For example, an original standard deviation of 10.25 is now 102.51. The adjustments to the accuracy data now show the original standard deviation values as standard errors and new standard deviations are reported. Although the adjustments to the tables reflect corrections to plausible mishaps in the publication process and correcting such errors should be applauded, the new data themselves are problematic when compared to the reported analyses.

To illustrate the disconnect between the new values and the test statistics in Article 1, we used the `R` package `faux` to simulate data from multivariate normal distributions with the same mean and standard deviation parameters as the original and updated tables. We then analyzed the data using the same 4 x 4 mixed ANOVA model reported by the authors and compared the test statistics we observed to those reported in Article 1. We tried various correlations between time points and chose the value that produced the closest agreement between our analyses and theirs (r = .8).

Our analysis of simulated reaction time data using the originally reported standard deviations produced F-statistics that were more similar to the values reported in Article 1 than our analysis based on the updated numbers. Using the originally reported figures, we observed an F = 18193 for the main effect of time. The authors reported F = 16055 for this analysis. Our analysis using the updated statistics resulted in F = 186. Similarly, the Time x Group interaction was $F = 4242$ in Article 1, $F = 6208$ in our simulation of the original parameters, and $F = 54$ using the updated numbers. The main effect of Group was $F = 7156$ in Article 1, $F = 3030$ in our simulation of the original parameters, and $F = 23$ using the updated numbers. As is evident, although the updated standard deviations lead to more plausible effect size calculations than the originally reported values, they are not consistent with the analyses reported in the paper.

We observed similarly discordant F-statistics with analyses of simulated accuracy data based on the updated statistics. Article 1 reported a main effect of Time of $F = 3278$, we observed $F = 2412$ when analyzing our simulation of the original parameters, and $F = 132$ when using the updated parameters. The authors reported a Group x Time interaction of $F = 657$, we found $F = 494$ with simulations of the original statistics, and $F = 27$ when using the updated values. Finally, the main effect of Group was reported as $F = 922$, we found $F = 254$ with simulations of the original statistics, and $F = 21$ when analyzing simulations of the updated data.

Discussion

We have reviewed concerning data irregularities spanning six articles investigating implicit motor and perceptual learning (Lola et al., 2021; Lola & Tzetzis, 2020, 2021; Lola et al., 2012; Tzetzis & Lola, 2010, 2015). These data irregularities include implausibly large effect sizes for pairwise comparisons and impossible descriptive statistics—both of which have been acknowledged by the authors as misprints due to an outsourcing of table formatting (A. Lola, personal communication, April 22, 2021). Further, the reported test statistics and associated η^2 values are also implausibly large, which is inconsistent with the authors' claim that the results and discussions remain valid despite these aforementioned typographic errors in the tables. We discovered that the data reported in Articles 1 and 2 are very highly correlated despite ostensibly coming from different experiments, samples, and situations (Lola et al., 2021; Lola & Tzetzis, 2021). Finally, we observed that recently updated tables of summary statistics in Article 1 were incompatible with the analyses reported in the paper, while the original summary statistics, which indicated implausible effect sizes, were more compatible with the analyses. Considering these findings, the conclusions from these articles are not reliable.

The data published in Article 1 are especially concerning. The article information indicates that it was submitted on June 17, 2021. Our email correspondence with the authors ended on April 22, 2021. Therefore, the article was submitted with data that reflect implausible effect sizes as large as $d = 41.0$ *after* we had shared our concerns about the previous five articles, and after the authors had suggested at least some of the implausible effect sizes were due to misprints. Despite this correspondence, the authors published an additional article with results that were not only implausible, but highly correlated with the results reported in a previous article. Subsequently, the authors published corrected tables with new standard deviation values. The means remained highly correlated with those reported in Article 2, but the new standard deviations were substantially larger than the original values. The updated summary statistics are not consistent with the test statistics reported in Article 1.

It is noteworthy that the results reported in each of these articles perfectly reflect the authors' expectations. Indeed, our attention was drawn to these articles after the Lola and Tzetzis (2021) paper was shared on Twitter (Gray, 2021); possibly because the results appeared to be exemplary. Although these errors seem unlikely to have aligned with expectations by chance alone, our exposure to them occurred after they had been selected for publication. We cannot rule out that these papers were selected for publication because of exemplary results and happened to have errors, and this selection caused those errors to correlate with the authors' expectations.

Other irregularities, such as a repeating F-statistic for all four analyses of explicit rules and the recurring age of participants potentially reflect sloppiness more than expectation. Indeed, the authors have already admitted that some values reported in their tables were in error, but failed to identify which values and which articles. Overall, it seems errors occurred in all of the articles we have reviewed. These errors were pervasive and appear to have substantially affected the conclusions of the articles in question. At a minimum, the consistent reporting errors across these six articles seem to reflect excessive carelessness throughout the publication process. Even if the authors offer additional corrections, which they have suggested they intend to do,[7] many in the research community may find it difficult to trust any of these results.

[7] There is no indication that such corrective actions were taken by the authors prior to us contacting the Editors on Sept 22 2021. However, as mentioned, the tables in Article 1 were updated and the authors offered some corrections in response to the editors of Article 5. We do not know if additional corrections have been submitted.

`R` packages used in this project

We used `R` (Version 4.0.4; R Core Team, 2021) and the `R` packages `compute.es` (Version 0.2.5; Re, 2013), `daff` (Version 0.3.5; Fitzpatrick et al., 2019), `faux` (Version 1.1.0; DeBruine, 2021), `gridGraphics` (Version 0.5.1; Murrell & Wen, 2020), `kableExtra` (Version 1.3.4; Zhu, 2021), `lemon` (Version 0.4.5; Edwards, 2020), `lsr` (Version 0.5; Navarro, 2015), `papaja` (Version 0.1.0.9997; Aust & Barth, 2020), `RColorBrewer` (Version 1.1.3; Neuwirth, 2014), `scales` (Version 1.2.0; Wickham & Seidel, 2020), and `tidyverse` (Version 1.3.0; Wickham et al., 2019).

Author Contact

Corresponding authors: Brad McKay (bradmckay8@gmail.com; mckayb9@mcmaster.ca) and Michael J. Carter (cartem11@mcmaster.ca)

Brad McKay 0000-0002-7408-2323
Michael J. Carter 0000-0002-0675-4271

Acknowledgements

We would like to thank Abbey Corson for her help with data extraction from the target articles.

Conflict of Interest and Funding

The authors declare no competing interests or funding for this project.

Author Contributions (CRediT Taxonomy)

Conceptualization: BM, MJC
Data curation: BM
Formal Analysis: BM
Methodology: BM, MJC
Project administration: BM, MJC
Software: BM
Supervision: MJC
Validation: BM, MJC
Visualization: BM
Writing – original draft: BM, MJC
Writing – review & editing: BM, MJC

Author order was determined by contribution.

Open Science Practices

This article earned the Open Data and the Open Materials badge for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is published in the online supplement.

References

Anaya, J. (2016). *The grimmer test: A method for testing the validity of reported measures of variability* (tech. rep.). https://doi.org/10.7287/peerj.preprints.2400v1

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown* [R package version 0.1.0.9997]. https://github.com/crsh/papaja

Bakker, M., & Wicherts, J. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Brown, N., & Heathers, J. (2017). The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*(4), 363–369. https://doi.org/10.1177/1948550616673876

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Correll, J., Mellinger, C., McClelland, G., & Judd, C. (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

DeBruine, L. (2021). *Faux: Simulation for factorial designs* [R package version 1.1.0]. Zenodo. https://doi.org/10.5281/zenodo.2669586

Edwards, S. M. (2020). *Lemon: Freshing up your 'ggplot2' plots* [R package version 0.4.5]. https://CRAN.R-project.org/package=lemon

Field, A. (2016). *An adventure in statistics: the reality enigma*.

Fischman, M. (2015). On the continuing problem of inappropriate learning measures: Comment on wulf et al. (2014) and wulf et al. (2015). *Human Movement Science*, *42*, 225–231. https://doi.org/10.1016/j.humov.2015.05.011

Fitzpatrick, P., de Jonge, E., & Warnes, G. R. (2019). *Daff: Diff, patch and merge for data.frames* [R package version 0.3.5]. https://CRAN.R-project.org/package=daff

Gray, R. [(2021). The effect of explicit, implicit and analogy instruction on decision making skill for novices, under stress.

Hancock, G., Butler, M., & Fischman, M. (1995). On the problem of two-dimensional error scores: Measures and analyses of accuracy, bias, and consistency. *Journal of Motor Behavior*, *27*(3), 241–250. https://doi.org/10.1080/00222895.1995.9941714

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

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

Lola, A., Giatis, G., Pérez-Turpin, J., & Tzetzis, G. (2021). The influence of analogies on the development of selective attention in novices in normal or stressful conditions. *Journal of Human Sport and Exercise*, *2023*, 139–152. https://doi.org/10.14198/jhse.2023.181.12

Lola, A., & Tzetzis, G. (2020). Analogy versus explicit and implicit learning of a volleyball skill for novices: the effect on motor performance and self-efficacy. *Journal of Physical Education and Sport*, *20*(5), 2478–2486. https://www.cabdirect.org/cabdirect/abstract/20203562097

Lola, A., & Tzetzis, G. (2021). The effect of explicit, implicit and analogy instruction on decision making skill for novices, under stress. *International Journal of Sport and Exercise Psychology*, *0*(0), 1–21. https://doi.org/10.1080/1612197X.2021.1877325

Lola, A., Tzetzis, G., & Zetou, H. (2012). The effect of implicit and explicit practice in the development of decision making in volleyball serving. *Perceptual and Motor Skills*, *114*(2), 665–678. https://doi.org/10.2466/05.23.25.PMS.114.2.665-678

Lovakov, A., & Agadullina, E. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *00*, 1–20. https://doi.org/10.1002/ejsp.2752

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

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

Munafò, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie du Sert, N., Simonsohn, U., Wagenmakers, E., Ware, J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Murrell, P., & Wen, Z. (2020). *Gridgraphics: Redraw base graphics using 'grid' graphics* [R package version 0.5-1]. https://CRAN.R-project.org/package=gridGraphics

Navarro, D. (2015). *Learning statistics with r: A tutorial for psychology students and other beginners. (version 0.5)* [R package version 0.5]. University of Adelaide. Adelaide, Australia. http://ua.edu.au/ccs/teaching/lsr

Neuwirth, E. (2014). *Rcolorbrewer: Colorbrewer palettes* [R package version 1.1-2]. https://CRAN.R-project.org/package=RColorBrewer

Nuijten, M., Hartgerink, C., van Assen, M., Epskamp, S., & Wicherts, J. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

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

Re, A. C. D. (2013). *Compute.es: Compute effect sizes*. https://cran.r-project.org/package=compute.es

Reeve, T., Fischman, M., Christina, R., & Cauraugh, J. (1994). Using one-dimensional task error measures to assess performance on two-dimensional tasks: Comment on 'attentional control, distractors, and motor performance'. *Human Performance*, *7*(4), 315–319. https://doi.org/10.1207/s15327043hup0704_6

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/pits.20234

Tsorbatzoudis, H., Barkoukis, V., Kaissidis-Rodafinos, A., & Grouios, G. (1998). A test of the reliability and factorial validity of the greek version of the csai-2. *Research Quarterly for Exercise and Sport*, *69*(4), 416–419. https://doi.org/10.1080/02701367.1998.10607717

Tzetzis, G., & Lola, A. (2010). The role of implicit, explicit instruction and their combination in learning anticipation skill under normal and stress conditions. *International Journal of Sport Sciences and Physical Education*, *1*, 54–59.

Tzetzis, G., & Lola, A. (2015). The effect of analogy, implicit, and explicit learning on anticipation in volleyball serving. *International Journal of Sport Psychology*, *46*(2), 152–166. https://doi.org/10.7352/IJSP.2015.46.152

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), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Seidel, D. (2020). *Scales: Scale functions for visualization* [R package version 1.1.1]. https://CRAN.R-project.org/package=scales

Zhu, H. (2021). *Kableextra: Construct complex table with 'kable' and pipe syntax* [R package version 1.3.4]. https://CRAN.R-project.org/package=kableExtra