STORK

SOCIETY FOR TRANSPARENCY OPENNESS AND REPLICATION IN KINESIOLOGY
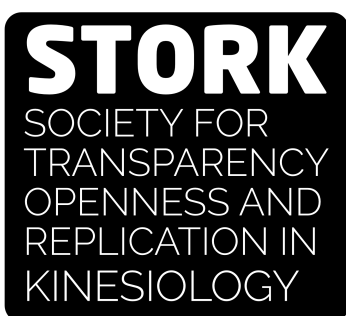
Cumulative evidence synthesis and consideration of 'research waste' using Bayesian methods

An example updating a previous meta-analysis of self-talk interventions for sport/motor performance

Hannah Corcoran[1] and James Steele[1,2,*]

[1]Department of Sport and Health, Solent University, UK
[2]Steele Research Ltd, UK
[*]Correspondence: james@steele-research.com

In the present paper we demonstrate the application of methods for cumulative evidence synthesis including Bayesian meta-analysis, and exploration of questionable research practices such as publication bias or p-hacking, in the sport and exercise sciences for the evaluation of experimental interventions. The use of such methods can aid in study planning and avoid "research waste". In demonstrating and discussing these methods we use the example of self-talk interventions and their effects upon sport/motor performance given a quantitative evidence synthesis has not been conducted on this topic, to the best of our knowledge, since 2011 when Hatzigeorgiadis et al. (2011) conducted their systematic review and meta-analysis. As such, this topic is ripe to use in demonstrating cumulative methods such as Bayesian updating. Therefore, our aim was to conduct an updated systematic review and Bayesian meta-analysis replicating the search, inclusion, and models of Hatzigeorgiadis et al. (2011) and demonstrate the application of cumulative evidence synthesis methods including; consideration of the initial probability that a new study of the effects of self-talk interventions would shift our prior belief in their effectiveness, the application of priors taken from the previous meta-analysis to be updated by new studies identified to a new posterior estimate of effect, and consideration of other possible sources of research waste from questionable research practices such as publication bias and p-hacking. Such methods as those demonstrated here, when used prospectively, can aid researchers in determining whether further research of a particular experimental intervention is in fact warranted. Considering the limited resources and time for conducting research we hope that highlighting the application of these methods might help researchers in the field to avoid research waste and more productively direct their research efforts.

Introduction

Cumulative evidence synthesis

Two questions that should be asked (though arguably are not asked often enough particularly in sport and and exercise science) by researchers when planning a study of an experimental intervention is *"what is the likelihood that the experimental intervention is superior to the control intervention given the evidence accumulated so far?"*, and *"what is the likelihood that a new trial, given some design parameters and previous evidence, will demonstrate the superiority of the experimental intervention?"*. The key here is to consider the *cumulative* nature of evidence provided by research and its synthesis. Indeed, to not do this could lead to redundancy or so called "research waste". Evidence synthesis methods are essential to determining whether or not there is justification for further research on a given topic, and the Cochrane-Collaboration and REWARD (Reduce Research Waste and Reward Diligence) Alliance have even established an award for efforts in the area of reducing "research waste" (Glasziou & Chalmers, 2018). However, across many domains there remains a high prevalence of redundancy and a low prevalence of attempts to minimise or reduce it (Lund et al., 2022).

Cumulative meta-analyses were proposed in the early 1990s and have since then been promoted as key tools to understand whether or not additional research is a worthwhile use of resources for addressing a particular question regarding experimental intervention (Clarke et al., 2014; Grainger et al., 2020). Further, Bayesian approaches are well positioned to tackle this (Biau et al., 2017). Within Bayesian statistical inference a prior probability distribution regarding the effect of interest is *updated* after the introduction of new evidence to a posterior probability distribution given Bayes theorem.

The trustworthiness of prior data should also be considered in evidence synthesis. Meta-analyses rely on the assumption that the sample of studies included is not based on a biased selection procedure either on the part of the systematic reviewer(s) or with regards to the studies present in the literature to sample from. For the latter, publication bias and p-hacking are the two most common phenomena that violate this assumption and can substantially influence cumulative evidence (Friese & Frankenbach, 2020). Publication bias is typically explored using methods such as selection models based on significance thresholds for p-values (McShane et al., 2016), funnel plot based regression methods (Stanley & Doucouliagos, 2014), or methods which combine these approaches and leverage the uncertainty in the underlying true data generating process such as robust Bayesian meta-analysis (RoBMA) with model averaging (Bartoš et al., 2022, 2023). For p-hacking, mixture models have recently been proposed (Moss & De Bin, 2023). The existence of evidence suggesting prior questionable research practices such as publication bias or p-hacking in previous literature might be cause for an evaluation of whether a research program regarding an experimental intervention is worth building upon, or starting afresh to determine if there really is an intervention effect whilst incorporating safe-guards for such issues e.g., pre-registration/registered-reports (Chambers & Tzavella, 2022).

In the present paper we demonstrate the application of methods for cumulative evidence synthesis including Bayesian meta-analysis, and exploration of questionable research practices such as publication bias or p-hacking, in the sport and exercise sciences for the evaluation of experimental interventions. We assume some prior knowledge of evidence synthesis and meta-analytic methods on the part of the reader, though for those unfamiliar suggest some recent introductory papers regarding their application in the field (Gunnell et al., 2020; M. Hagger, 2022; Steele et al., 2023). We use the example of self-talk interventions and their effects upon sport/motor performance given a quantitative evidence synthesis has not been conducted on this topic, to the best of our knowledge, since 2011 when Hatzigeorgiadis et al. (2011) conducted their systematic review and meta-analysis. As such, it is a ripe topic to use in demonstrating cumulative methods such as Bayesian updating. Note, we do not intend to present this paper as a comprehensive systematic review and meta-analysis of self-talk interventions in sport/motor performance nor a thorough theoretical review of the construct. We do however present a brief overview of the topic below for context of this as an example.

Self-talk interventions

Sport psychology as a broad field has focused on the theorising of psychological constructs that might impact upon performance, and the subsequent experimental testing of theoretically informed

interventions to address these constructs and subsequent performance. For example, a recent umbrella review identified thirty meta-analyses exploring the effects of different sport psychology constructs upon performance, thirteen of them examining the effects of interventions, finding an overall standardised mean difference (SMD) for positive constructs of 0.51 [95% confidence interval: 0.42, 0.58] (Lochbaum et al., 2022). One construct with a long history of philosophical, theoretical, and empirical work that has been the target of considerable investigation in this field has been self-talk (Brinthaupt & Morin, 2023; Geurts, 2018; Latinjak et al., 2023).

As a concept self-talk has been defined in various ways in previous work on the topic; though recent transdisciplinary review (Latinjak et al., 2023) has agreed upon a broad conceptualisation: *"verbalizations addressed to the self, overtly or covertly, characterized by interpretative elements associated to their content; and it [self-talk] either (a) reflects dynamic interplays between organic, spontaneous and goal-directed cognitive processes or (b) conveys messages to activate responses through the use of predetermined cues developed strategically, to achieve performance-related outcomes"* (Latinjak et al., 2019). Whilst there has been various narrative syntheses of research on self-talk in the past decade (Hardy et al., 2018; Latinjak et al., 2023; Van Raalte et al., 2016), only one systematic review and meta-analysis has explored the effects of self-talk interventions; Hatzigeorgiadis et al. (2011).

The meta-analysis by Hatzigeorgiadis et al. (2011) included a total of 32 studies and 62 effect size estimates revealing an overall SMD estimate of 0.48 [95% confidence interval: 0.38, 0.58] and also explored the effects of various theoretically driven moderators of the effectiveness of self-talk interventions. For example; characteristics of the tasks performed such as their novelty and whether they are fine or gross motor task, characteristics of the participants such as their level of experience with the task, the characteristics of the self-talk used including its content, whether it was self-selected or assigned, and whether it was used overtly or not, characteristics of the intervention and whether it included brief exposure or a training period, and testing the 'matching hypothesis' which posits that instructional self-talk should benefit fine tasks whereas motivational self-talk should benefit gross tasks to a greater degree.

Around the time that Hatzigeorgiadis et al. conducted their meta-analysis the quantitative synthesis of research findings using meta-analytic tools was still relatively new in the sport sciences (Hagger, 2006). However in the last decade, particularly in sport psychology, there has been an increasing reliance on meta-analyses (M. Hagger, 2022; Lochbaum et al., 2022). Despite the general proliferation of meta-analyses in the past decade, the effect of self-talk interventions has not been re-evaluated by means of such quantitative synthesis since 2011, when Hatzigeorgiadis et al. (2011) completed their work. During this period though, empirical research regarding self-talk interventions for sport and motor performance has burgeoned leading some to reflect on the field as "maturing" post-2011 (Hardy et al., 2018).

Whilst self-talk as a field may have matured in the post-2011 years with theoretical advancements in conceptualisation of the construct and proposed mediators of its effects on performance, efforts to improve operationalisation, and efforts to improve methodology used in studying self-talk (Brinthaupt & Morin, 2023; Geurts, 2018; Hardy et al., 2018; Latinjak et al., 2019; Latinjak et al., 2023; Van Raalte et al., 2016, 2019), it could be argued that understanding of the effectiveness of self-talk *interventions* (referred to in modern literature as 'strategic' self-talk; Latinjak et al. (2019)) was "mature" prior to 2011. The effect estimate from the meta-analysis of Hatzigeorgiadis et al. (2011) might be considered by some to have been already fairly precise only spanning 0.2 SMD, and also for many of the moderator estimates. Indeed, some authors had moved onto to attempting to explain *why* self-talk interventions are effective, exploring possible mechanisms with the starting assumption that these interventions had been proven as effective for enhancing performance (Galanis et al., 2016).

Despite this, many additional studies on self-talk interventions have been conducted since 2011. It may well be that such recent work has further improved our estimates of the effects of self-talk interventions and what moderates their effectiveness, or indeed contributed to other areas of understanding of the construct of self-talk. But, it is a reasonable question to ask, given the limited time and resource for conducting research in the field of sport science and what we might claim to have already known regarding these interventions, whether and to what extent these studies have advanced our understanding of their effects, or whether they have largely contributed to so called "research waste" (Glasziou & Chalmers, 2018; Grainger et al., 2020).

Aim of the present work

The aim of this present work is to demonstrate the application of methods for cumulative evidence synthesis including Bayesian meta-analysis, and exploration of questionable research practices such as publication bias or *p*-hacking, in the sport and exercise sciences for the evaluation of experimental interventions. Given that there has not been, to the best of our knowledge, a meta-analytic synthesis of the effects of self-talk interventions effects upon sport/motor performance since Hatzigeorgiadis et al. (2011) it represents a ripe topic to utilise as an example for these methods. Therefore, our aim was to conduct an updated systematic review and Bayesian meta-analysis replicating the search, inclusion, and models of Hatzigeorgiadis et al. (2011) in order to demonstrate the application of cumulative evidence synthesis methods including; consideration of the initial probability that a new study of the effects of self-talk interventions would shift our prior belief in their effectiveness, the application of priors taken from the previous meta-analysis to be updated by new studies identified to a new posterior estimate of effect, and consideration of other sources of research waste from questionable research practices such as possible publication bias and *p*-hacking.

Method

The method for this systematic review and meta-analysis was replicated with slight adaptation from Hatzigeorgiadis et al. (2011). We limited our searches to the date range of November 2011 to November 2023 to avoid double counting as we used the estimates from Hatzigeorgiadis et al. (2011) as informative priors in our meta-analyses which contain the information from studies prior to November 2011.

Criteria for including studies

Hatzigeorgiadis et al. (2011) did not explicitly state a process or strategy to formulating their research question and search methods. However, we assumed that the PICO (Participants, Intervention, Comparator and Outcome) framework was implicitly used and, with that assumption, we adopted the following inclusion criteria based on their description. Participants were healthy and of any performance level. The intervention was instruction to engage in positive self-talk[1]. The comparator was no self-talk or unrelated self-talk. Outcomes were sport or motor task performance. We included both between and within group experimental designs with either pre-post, or post-only measurements of performance as well as within group pre-post trials similarly to Hatzigeorgiadis et al. (2011).

Search strategy

Studies were obtained through electronic journal searches and review articles along with personal records and communication. The following databases – Sport Discus, PsycINFO, PsycARTICLES and Medline – were selected through the EBSCO database to search for the keywords. The SCOPUS database, used by Hatzigeorgiadis et al. (2011), was not used as it was not accessible through Solent University[2]. These keywords were searched in the format of, with the application of the Boolean commands, (self-talk OR self-instruction OR self-statements OR self-verbalizations OR verbal cues OR stimulus cueing OR thought content instructions) AND (sport OR performance OR motor performance OR task performance). The studies were all peer-reviewed, full text and published in English language journals. The search was limited to the date range of November 2011-November 2023. An initial search took place from October

---

[1] Hatzigeorgiadis et al. (2011) stated: *"As our purpose was to test the effectiveness of interventions aiming to improve performance, groups or conditions using negative … or inappropriate self-talk … were excluded. In addition, groups or conditions using assisted self-talk … were also excluded as assisted self-talk involves the use of external aids, such as headphones, and was not considered pure self-talk intervention."*

[2] Though we feel fairly confident, given the number of studies identified and the findings of our models reported below, that any missed studies would be unlikely to qualitatively impact the overall findings and conclusions of this work. We should also note that we originally intended to report a PRISMA flow diagram of our search and retrieval process. However, for transparency, we encountered an issue where we realised that some of the searches we had conducted were not correctly recorded by the tool being used to manage the process. We were, despite attempting to do so after realising this, unable to exactly replicate the searches (number of hits from initial search string in databases used) at the time we realised from attempting to reproduce the searches again. As such, we do not present a PRISMA diagram though as noted we do feel fairly confident we have not missed any key studies nor that minor omissions would affect the overall results of our analyses anyway. Further, as noted we have chosen this as an example to demonstrate the methods moreso than to conduct a comprehensive systematic review of the topic.

2022-November 2022 as this project was completed as part of the lead authors undergraduate thesis. We subsequently updated the search from November 2022-November 2023 prior to initially preparing this manuscript for publication.

Data extraction

The data extracted from the studies were for all positive self-talk intervention groups/conditions and for control comparison designs for the relevant comparator group/condition. Pre and/or post intervention and comparator, means, sample sizes and either standard deviation, standard errors, variances or confidence intervals were extracted in order to calculate the effect sizes. Also, in order to update the moderator analyses conducted by Hatzigeorgiadis et al. (2011) we also coded each effect size for motor demands (fine or gross), participant group (non-athletes[3] vs beginner athletes vs experienced athletes), self-talk content[4] (motivational vs instructional), the combination of motor demands and self-talk content to examine the matching hypothesis (motivational/gross vs motivational/fine vs instructional/gross vs instructional/fine), the task novelty (novel vs learned), cue selection and overtness selection (self-selected vs assigned), if the study was acute or involved a chronic training intervention (no-training vs training), and the study design[5] (pre/post - experimental/control vs pre/post - experimental vs post - experimental/control). The data extracted was imported into a spreadsheet in excel as a csv.

Statistical analysis

All code utilised for data preparation and analyses are available in either the Open Science Framework page for this project https://osf.io/dqwh5/ or the corresponding GitHub repository https://github.com/jamessteeleii/self_talk_meta_analysis_update. We cite all software and packages used in the analysis pipeline using the `grateful` package (Rodriguez-Sanchez et al., 2023) which can be seen here: https://osf.io/ftajc.

Examining the effects of a new trial upon belief in the effects of self-talk interventions

To begin with we examined through simulation what impact a single new trial might have had upon shifting belief in the prior estimate yielded by the meta-analysis of Hatzigeorgiadis et al. (2011). Sample size was simulated as 10, 20, 40, 80, 160, 320, 640, 1282, 2560, and 5120[6] with a 50:50 allocation to either self-talk intervention or control conditions, and we varied the sample effect size as an SMD of 0, 0.2, 0.4, 0.6, 0.8, and 1.0 reflecting a range from no effect of self-talk interventions to a large effect. In each combination of sample size and true effect size we set the sample parameters to the SMD and its corresponding sampling variance was calculated and then included as a single observation in a Bayesian random effects meta-analysis where the priors was set informatively for the intervention effects, and were set to be default weakly regularising for the heterogeneity (i.e., τ)[7]. Intervention effects were set with priors based on the effect estimates from Hatzigeorgiadis et al. (2011), reported in their Table 1, using a t-distribution ($t(k, \mu, \sigma)$) with $k - 2$ degrees of freedom (Higgins et al., 2009). We assumed k to be the number of effects included in the models reported by Hatzigeorgiadis et al. (2011). The prior for the intervention effect was set directly on the model intercept i.e., $t(60, 0.48, 0.05)$. The prior for heterogeneity was set as the default weakly regularising prior in `brms`; a half-t-distribution with $\mu = 0$, $\sigma = 2.5$, and $k = 3$. This constrained the prior to only allow positively signed values for τ though over a wide range of possible values. We fit each model using four Monte Carlo Markov Chains each with 2000 warmup and 6000 sampling iterations.

---

[3]Hatzigeorgiadis et al. (2011) referred to their non-athlete group as "students" presumably because in all studies they included it was the case that all non-athletes were from student populations. As this was not necessarily the case for studies included in our updated analyses we refer to them as "non-athletes".

[4]Some studies we included in our updated analysis used combined instructional and motivational, and also other forms of self-talk content e.g., rational. We coded these new categories also.

[5]Hatzigeorgiadis et al. (2011) included studies with multiple baseline measures but we did not identify any studies meeting this design in our updated analysis.

[6]Sample sizes were doubled up to roughly a similar sample size as the largest study post-2011 that we identified in our updated searches (Lane et al., 2016).

[7]Though Hatzigeorgiadis et al. (2011) report on τ it is not clear what level this applies to (and as noted it is not clear if they employed a hierachical model) and they do not report any interval estimate for this making it difficult to specify an informative prior distribution. As such, and given suggestions regarding heterogeneity priors (Röver et al., 2021; Williams et al., 2018), we opted for a weakly regularising distribution at all levels including the updated multilevel meta-analysis described further in these methods and in the random effects meta-analysis used in these simulations.

From each model we obtained draws from the posterior distributions for the intervention effects (i.e., the expectation of the value of the parameters posterior probability distribution) in order to present probability density functions visually. The same was done drawing samples from the prior distributions only in order to present both distributions visually for comparison of the prior to posterior updating. As a means of examining the extent to which the posterior distribution for the self-talk intervention effect estimate was shifted from the prior distribution as a result of introducing each new trial we calculated the proportion of the full posterior distribution within the 95% quantile interval of the prior distribution i.e., the range from 2.5% to 97.5% percentiles (equivalent to the 95% confidence interval of the estimate from Hatzigeorgiadis et al. (2011)), using the a Region of Practical Equivalence approach (Kruschke & Liddell, 2018). In essence, where the proportion of the posterior distribution that was within the 95% quantile interval from the prior distribution was ~95% then we would conclude that the new trial had little impact on shifting our prior belief in the intervention effect. This helps understand whether it might be worthwhile to conduct the kind of study that would need to be conducted in terms of sample size, and assuming what the true effect might be, or whether doing so might be a waste of resources given the precision of existing estimates of the intervention effect.

Updating the prior estimate from Hatzigeorgiadis et al. (2011) with newer studies

For all the groups/conditions in studies identified in our updated searches effect sizes were calculated as SMDs dependent on the design of the study. Firstly, all were signed such that a positive effect indicated that the self-talk intervention was favoured. For studies utilising a pretest-posttest-control comparison design we calculated the SMD between groups/conditions using the pooled pre-test standard deviation as per Morris (2008). For post-test only control comparison designs we calculated the SMD between groups/conditions based upon the pooled post-test standard deviation. Lastly, for single arm within group pre-post (or control-intervention) designs we calculated the SMD from pre- to post-intervention using the pre-test standard deviation.

Though it was not entirely clear from the reporting in the meta-analysis of Hatzigeorgiadis et al. (2011), they noted including a greater number of effect sizes than individual studies. As such, it was likely that their data had a hierarchical structure with effects nested within studies whether they explicitly applied a hierarchical model to it or not. The studies we identified and included also had hierarchical structure whereby we had effects nested within groups (for example when there were multiple self-talk interventions examined) nested within experiments (for example when a study reported on multiple experiments using different samples and/or designs) nested within studies. As such, we used multilevel mixed effects meta-analyses with nested random intercepts for effects, groups, experiments, and studies. Effects were all weighted by the inverse sampling variance. A main model was produced which included all effects and was intended to update the overall model from Hatzigeorgiadis et al. (2011), whereby their overall estimate reflected the fixed model intercept. In addition, we produced models for each of the aforementioned categorical moderators where we excluded the model intercept in order to set priors for each category directly based on the estimates and their precision reported (see footnote[7]) by Hatzigeorgiadis et al. (2011).

Priors for each model were again set informatively for the intervention effects, and were set to be weakly regularising for the heterogeneity (i.e., τ) at all levels of the model. Intervention effects were set with priors based on the effect estimates from Hatzigeorgiadis et al. (2011), reported in their Table 1, using a t-distribution ($t(k, \mu, \sigma)$) with $k - 2$ degrees of freedom (Higgins et al., 2009). We assumed k to be the number of effects included in the models reported by Hatzigeorgiadis et al. (2011). For the main model the prior for the intervention effect was set directly on the model intercept i.e., $t(60, 0.48, 0.05)$. For the moderator models, as noted, we removed the model intercept allowing us to set the priors directly on each category for each moderator based on the estimates from Hatzigeorgiadis et al. (2011) Table 1. In cases where moderators had new categories introduced in the newer studies, included in our analyses, we used $\mu = 0.48$ and $\sigma = 0.05$ taken from the overall estimate of Hatzigeorgiadis et al. (2011) and applied degrees of freedom $k = 3$ to be more conservative and allow greater mass in the tails of the prior distribution for these categories. In all models the heterogeneity priors at each level were set using the default weakly regularising prior in brms; a half-t-distribution with $\mu = 0$, $\sigma = 2.5$, and $k = 3$. This constrained the prior to only allow positively signed values for τ though over a wide range of possible values.

As we were interested in determining how much the new evidence produced since Hatzigeorgiadis et al. (2011) had updated our belief in the effects of self-talk interventions, we fit each model using four

Monte Carlo Markov Chains each with 4000 warmup and 40000 sampling iterations. This was in order to obtain precise Bayes Factors using the Savage-Dickey ratio (Gronau et al., 2020). Trace plots were produced along with $\hat{R}$ values to examine whether chains had converged, and posterior predictive checks for each model were also examined to understand the model implied distributions. These all showed good convergence with all $\hat{R}$ values close to 1 and posterior predictive checks seemed appropriate distributions for the observed data (all diagnostic plots can be seen in the supplementary materials: https://osf.io/ag6re).

From each model we obtained draws from the posterior distributions for the intervention effects (i.e., the expectation of the value of the parameters posterior probability distribution) in order to present probability density functions visually, and also to calculate mean and 95% quantile intervals (i.e., 'credible' or 'compatibility' intervals) for each estimate. These gave us the most probable value of the parameter in addition to the range from 2.5% to 97.5% percentiles. The same was done drawing samples from the prior distributions only in order to present both distributions visually for comparison of the prior to posterior updating. For the main model draws were taken at the study level and an ordered forest plot produced showing each studies posterior distribution along with mean and 95% quantile intervals. We also calculated the 95% prediction intervals providing the range over which we can expect 95% of future effect estimates to fall and present each individual effect size on the forest plot.

To compliment the visual inspection of prior to posterior updating we also present log10 Bayes Factors (log10[BF]) calculated *against* 100 effects ranging from a SMD of 0 through to 1 and plot these log10(BF) curves for each model intervention effect estimate i.e., the Savage-Dickey ratio was calculated for each of 100 point effects in the interval (0,1) equally spaced. These were compared to Jeffreys (1998) scale regarding evidence *against* (i.e., 0 to 0.5 = weak evidence; 0.5 to 1 = substantial evidence; 1 to 1.5 = strong evidence; 1.5 to 2 = very strong evidence; 2 or greater = decisive evidence). Thus, a positive log10(BF) value indicated that, compared to the prior distribution (meaning the estimates of Hatzigeorgiadis et al. (2011)), there was now greater evidence *against* the SMD for which the log10(BF) was calculated. A loess smooth was then applied to these 100 values for visual presentation.

Lastly, as a supplemental analysis, we produced cumulative versions of our main model over each year since the publication of the meta-analysis from Hatzigeorgiadis et al. (2011). The first model started with the prior distribution noted above for our main model and only included effects from studies reported in 2011. Then we took the posterior distribution for the intervention effect from this model and used it as the prior for the next model which only included effects from studies reported in 2012. This was continued through each year up to the latest included studies. We then plotted the cumulative updating of the intervention effect based on the addition of each years newly reported studies. Note, for each of these models we employed four Monte Carlo Markov Chains each with 2000 warmup and 6000 sampling iterations given the focus was on presenting the updated estimates and to reduce the time required for cumulative models to be fit.

Examining the quality of the evidence and potential questionable research practices

Both simulating the impact of a new trial to determine if it is worth performing, or updating a prior meta-analysis estimate with new evidence from subsequent trials, entail the assumption that the previous estimate is not biased by questionable research practices such as publication bias or *p*-hacking (Friese & Frankenbach, 2020). The latter (i.e., updating a prior estimate) also relies on the assumption that the subsequent evidence to be included is also not biased by such influences. Hatzigeorgiadis et al. (2011) did employ the fail-safe *N* approach which determines the number of unpublished *null* studies that would reduce the meta-analytic effect estimate non-significant and concluded that publication bias was unlikely ($K_0 = 102$). However, fail-safe *N*, whilst previously widely used in meta-analyses around the time (Heene, 2010), was known to be flawed and other methods such as funnel plot based regression techniques instead recommended (Becker, 2005). Given this project was deliberately initiated as part of an undergraduate thesis with the intention of limiting the systematic review component to post 2011 due to time constraints and conduct an updated Bayesian meta-analysis, and that the purpose of the present manuscript is to demonstrate various cumulative evidence synthesis methods using the self-talk literature as an example, we did not ourselves acquire the data for studies from Hatzigeorgiadis et al. (2011) to enable us to examine the possible presence of questionable research practices that might impact the

prior estimate[8]. Instead we limit our examination to the subsequent post-2011 literature and make the reasonable assumption that, given the current replication crisis and subsequent methodological reform efforts kicked off proper in the early 2010s (Lakens, 2023), the presence of questionable research practices was likely as bad if not worse in the literature included in Hatzigeorgiadis et al. (2011).

Examining the presence of questionable research practices such as publication bias and p-hacking in the presence of data with a hierarchical structure such as we have in the present example is not simple. The methods noted in the introduction have primarily been developed for cases of fixed or random effects meta-analyses were each study contributes only a single effect in the model. However, some approaches have been extended to the hierarchical case such as funnel plot based regression methods (Rodgers & Pustejovsky, 2021) and robust Bayesian meta-analysis (RoBMA) with model averaging (Bartoš et al., 2023). However the latter, and in particular the selection methods incorporated, are very computationally intensive making them in most regards practically unfeasible.

As such, we utilised the multilevel precision-effect test (PET) and precision-effect estimate with standard errors (PEESE) to estimate the adjusted effect size accounting for small study effects such as publication bias (Rodgers & Pustejovsky, 2021). The PET-PEESE respectively model a linear and quadratic relationship between standard error and effect size, the latter assuming that studies with very small standard errors, and thus large samples, are likely to be reported regardless of results whereas small studies with large standard errors required increasingly larger effects to be selected for publication. This approach is a conditional two-step estimator of the two models whereby if the test of the adjusted effect size with an $\alpha = 0.10$ (for model selection only) is not significant (i.e., $p > \alpha$) then PET is reported whereas if it is significant (i.e., $p < \alpha$) then PEESE is reported. The adjusted estimate was compared to the estimate generated from a multilevel meta-analysis model of the included studies. Note, the adjusted PET-PEESE estimate of the intervention effects and the comparative estimate from the multilevel meta-analysis model where both conducted using frequentist models.

In addition to this we also examined both the random effects mixture model for p-hacking (Moss & De Bin, 2023) and RoBMA (Bartoš et al., 2023) (both with, and without, the inclusion of an informative prior[9] on the intervention effect from Hatzigeorgiadis et al. (2011), yet with default priors for all other parameters) though ignored the hierarchical structure to the data as these models respectively have not been extended to the multilevel case or are too computationally intractable at present. The random effects mixture model for for p-hacking provides an adjusted estimate of the intervention effect under the assumption that questionable research practices such as excluding observations, collecting new data *ex-post*, or selectively including covariates has occurred to produce a one-sided $\alpha \leq 0.05$ (to reflect the fact that typically it is effects in a particular direction that are of interest) or $\alpha \leq 0.025$ to account for reporting of two-sided tests. The adjusted estimate assuming the presence of p-hacking was then compared to a classical random effects model estimate. Both the p-hacking and classical models were fit using Bayesian estimation. The RoBMA fits a total of 36 different models with varying assumptions regarding the true data generating process underlying the included studies/effects and thus reflecting our uncertainty in it: selection models for publication varying sidedness of the tests (one- or two-sided) and the specific p-value cutoffs used (combinations of 0.025, 0.05, 0.10, and 0.50), the PET-PEESE regression based models, models assuming there is/is not an intervention effect, models assuming there is/is not heterogeneity, and models assuming there is/is not publication bias. These models are then combined using Bayesian model-averaging and weighted based on how well the model fit the data. Bayes Factors were then calculated to examine the evidence in favour of there being an effect, the presence of heterogeneity, and publication bias. Bayes factors were interpreted according to Lee and Wagenmakers (2014) adaptation of Jeffreys (1998) scale where between 1 and 3 (between 1 and 1/3) are regarded as anecdotal evidence, Bayes factors between 3 and 10 (between 1/3 and 1/10) are regarded as moderate

---

[8]We did however contact Hatzigeorgiadis et al. (2011) to ask if they could share the extracted data, given it was not openly available, so we could examine this. We received a responde and are awaiting confirmation as to whether they have the data available still and can share it with us to examine. This manuscript will be updated with this if we do receive the extracted data from pre-2011 studies.

[9]Note that for the mixture model of p-hacking the `publipha` package used does not allow for specification of a prior based upon the t distribution for that function. As such, instead of the $t(60, 0.48, 0.05)$ prior taken from Hatzigeorgiadis et al. (2011) for these models we used a normal prior with the same location and scale parameters. Given the the t distribution approaches the normal distribution with high degrees of freedom, and the prior for other models had a high number of degrees of freedom (i.e., $k = 60$), the choice of a normal distribution here very closely approximates the t distribution based prior anyway and if anything is conservative in examining an adjusted estimate for p-hacking.

evidence, and Bayes factors larger than 10 (smaller than 1/10) are regarded as strong evidence in favor of (against) a hypothesis.

Results

Since Hatzigeorgiadis et al. (2011) meta-analysis the number of studies conducted examining the effects of self-talk interventions on sport/motor performance has roughly doubled. We identified 35 new studies published from November 2011 up to November 2023 (Abdoli et al., 2018; Barwood et al., 2015; Beneka et al., 2013; Blanchfield et al., 2014; Cabral et al., 2023; Chang et al., 2014; de Matos et al., 2021; Galanis et al., 2018; Galanis, Hatzigeorgiadis, Charachousi, et al., 2022; Galanis, Hatzigeorgiadis, Comoutos, et al., 2022; Galanis et al., 2023; Gregersen et al., 2017; Hatzigeorgiadis et al., 2014, 2018; Hong et al., 2020; Kolovelonis et al., 2011; Lane et al., 2016; Latinjak et al., 2011; Liu et al., 2022; Marshall et al., 2016; McCormick et al., 2018; Naderirad et al., 2023; Osman et al., 2022; Panteli et al., 2013; Raalte, Cornelius, et al., 2018; Raalte, Wilson, et al., 2018; Sarig et al., 2023; Turner et al., 2018; Wallace et al., 2017; Walter et al., 2019; Weinberg et al., 2012; Young et al., 2023; Zetou et al., 2014; Zourbanos et al., 2013a, 2013b). These included 128 effects nested in 64 groups nested in 42 experiments. The included studies contained a total of 18761 participants (see Table 1). We included all but one study (Marshall et al., 2016), due to the sample size for groups in this study being too small to calculate SMDs (i.e., n = 2 to 3), in our analyses.

Table 1. Sample sizes for self-talk intervention and non-intervention control groups.

Group	Sample Size
Self-talk	
All ST	14895
Minumum ST	2
Median ST	17
Maximum ST	3442
Control	
All CON	3866
Minumum CON	2
Median CON	18
Maximum CON	3442

Note:

ST = self-talk

CON = non-intervention control

Examining the effects of a new trial upon belief in the effects of self-talk interventions

Given the precision of the prior distribution taken from Hatzigeorgiadis et al. (2011) it is highly unlikely that any further study would have shifted belief in the effect estimate. As can be seen in Figure 1, irrespective of the sample size or magnitude of effect and the extent to which it disagreed with the prior estimate, no single new study could shift the posterior distribution to an extent that it didn't still have ~95% of its mass within the 95% quantile interval of the prior distribution. As such, if we were to take the prior estimate from Hatzigeorgiadis et al. (2011) at face value in the process of deciding whether to conduct a new study of self-talk interventions, at least with respect to their main effects on sport/motor performance, we should likely conclude that it would be a waste of resources given the precision of this existing estimate of the intervention effect. Of course, this only considers the addition of a single study and as noted since 2011 the number of studies examining self-talk interventions has roughly doubled. As such, it is worth considering the extent to which this volume of additional evidence might have updated belief in the estimate of their effects.
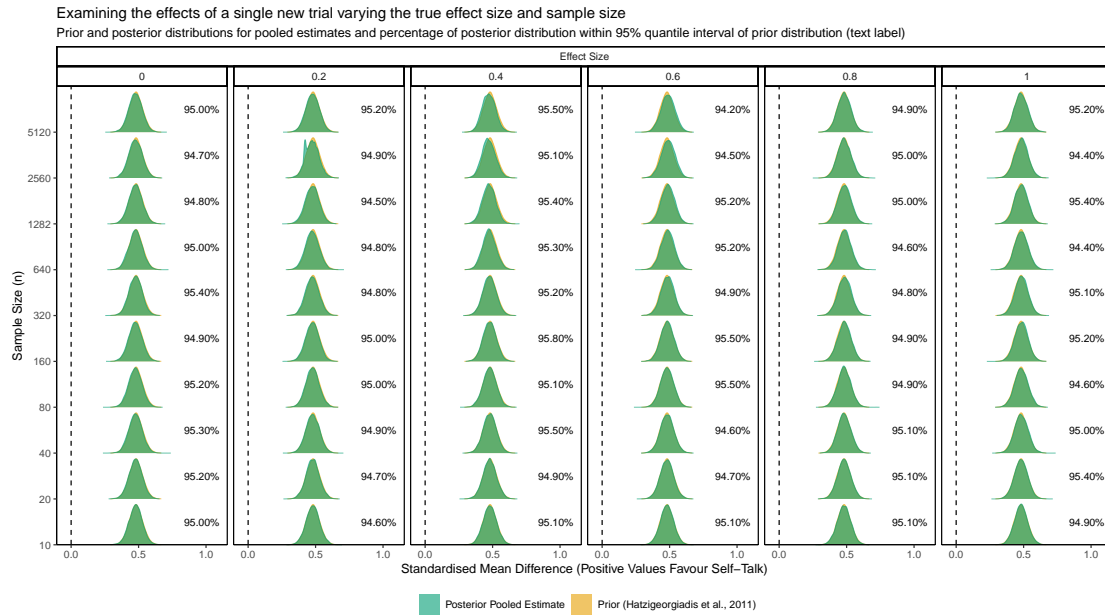
Figure 1. Prior and posterior distributions after the addition of a single new study yielding a single new effect varying in magnitude of standardised mean difference (null through to SMD=1 favouring self-talk interventions) and varying sample size of the new study (from n=10 to 5120).

Updating the prior estimate from Hatzigeorgiadis et al. (2011) with newer studies
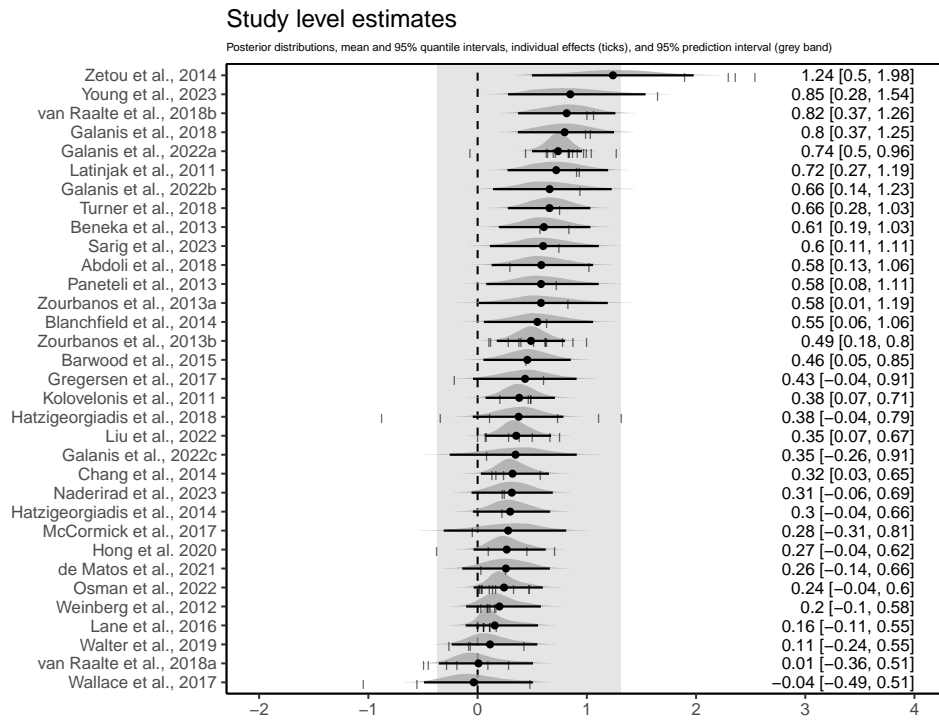
Main model

The overall mean and interval estimate for the SMD for self-talk interventions was 0.47 [95% quantile interval: 0.39, 0.56]. This was very similar to the estimate of overall effect in Hatzigeorgiadis et al. (2011) of 0.48 [95% confidence interval: 0.38, 0.58]. Heterogeneity (τ) at the study level was also similar to that reported by Hatzigeorgiadis et al. (2011), though as noted it is not clear what level theirs pertained to exactly. At the study level $\tau = 0.35$ [95% quantile interval: 0.11, 0.54], at the experiment level $\tau = 0.14$ [95% quantile interval: 0.01, 0.39], at the group level $\tau = 0.05$ [95% quantile interval: 0, 0.12], and at the effect level $\tau = 0.1$ [95% quantile interval: 0.03, 0.18]. An ordered forest plot of study level estimates is shown in Figure 2 panel (A), and the posterior pooled estimate for the overall SMD effect compared with the prior is shown in panel (B).

Considering the log10(BF) values calculated against the range of SMD effect sizes from 0 to 1 compared to Jeffreys scale (see Figure 2 panel [C]), the newly added evidence suggested that there was only "decisive" evidence updating the prior against effect sizes ranging from 0 to 0.08 and from 0.95 to 1. "Very strong" evidence was indicated against effect sizes ranging from 0.09 to 0.15 and from 0.84 to 0.94. "Strong" evidence was indicated against effect sizes ranging from 0.16 to 0.22 and from 0.72 to 0.83. "Substantial" evidence was indicated against effect sizes ranging from 0.23 to 0.3 and from 0.6 to 0.71. "Weak" or "negative" evidence was indicated against effect sizes ranging from 0.31 to 0.59. This suggested that the newly acquired evidence generally decreased our belief only in effect sizes that would likely already have been ruled out by the analysis of Hatzigeorgiadis et al. (2011). The supplementary cumulative models further supported this. They showed little change in either point or interval estimates for the SMD from year to year as a result of new studies during the period of November 2011 to November 2023 (see https://osf.io/9qrh5).

Moderators

For most of the moderators explored there was similarly little impact upon posterior estimates for the SMD from the introduction of new evidence accumulated since the analysis of Hatzigeorgiadis et al.
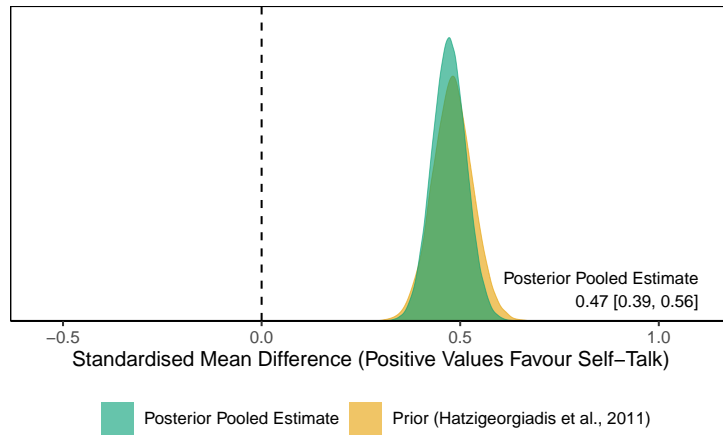
A

Study level estimates

Posterior distributions, mean and 95% quantile intervals, individual effects (ticks), and 95% prediction interval (grey band)

Zetou et al., 2014	1.24 [0.5, 1.98]
Young et al., 2023	0.85 [0.28, 1.54]
van Raalte et al., 2018b	0.82 [0.37, 1.26]
Galanis et al., 2018	0.8 [0.37, 1.25]
Galanis et al., 2022a	0.74 [0.5, 0.96]
Latinjak et al., 2011	0.72 [0.27, 1.19]
Galanis et al., 2022b	0.66 [0.14, 1.23]
Turner et al., 2018	0.66 [0.28, 1.03]
Beneka et al., 2013	0.61 [0.19, 1.03]
Sarig et al., 2023	0.6 [0.11, 1.11]
Abdoli et al., 2018	0.58 [0.13, 1.06]
Paneteli et al., 2013	0.58 [0.08, 1.11]
Zourbanos et al., 2013a	0.58 [0.01, 1.19]
Blanchfield et al., 2014	0.55 [0.06, 1.06]
Zourbanos et al., 2013b	0.49 [0.18, 0.8]
Barwood et al., 2015	0.46 [0.05, 0.85]
Gregersen et al., 2017	0.43 [−0.04, 0.91]
Kolovelonis et al., 2011	0.38 [0.07, 0.71]
Hatzigeorgiadis et al., 2018	0.38 [−0.04, 0.79]
Liu et al., 2022	0.35 [0.07, 0.67]
Galanis et al., 2022c	0.35 [−0.26, 0.91]
Chang et al., 2014	0.32 [0.03, 0.65]
Naderirad et al., 2023	0.31 [−0.06, 0.69]
Hatzigeorgiadis et al., 2014	0.3 [−0.04, 0.66]
McCormick et al., 2017	0.28 [−0.31, 0.81]
Hong et al. 2020	0.27 [−0.04, 0.62]
de Matos et al., 2021	0.26 [−0.14, 0.66]
Osman et al., 2022	0.24 [−0.04, 0.6]
Weinberg et al., 2012	0.2 [−0.1, 0.58]
Lane et al., 2016	0.16 [−0.11, 0.55]
Walter et al., 2019	0.11 [−0.24, 0.55]
van Raalte et al., 2018a	0.01 [−0.36, 0.51]
Wallace et al., 2017	−0.04 [−0.49, 0.51]

B

Updated Posterior Pooled Estimate

Prior and posterior distributions for pooled estimates, and mean and 95% quantile interval for posterior (text label)
Note: x–axis rescaled from panel (A) for easier comparison of prior and posterior distributions

Posterior Pooled Estimate
0.47 [0.39, 0.56]

Standardised Mean Difference (Positive Values Favour Self–Talk)

■ Posterior Pooled Estimate   ■ Prior (Hatzigeorgiadis et al., 2011)

C

Change in Evidence

Positive values indicate greater evidence against standardised mean difference after updating prior
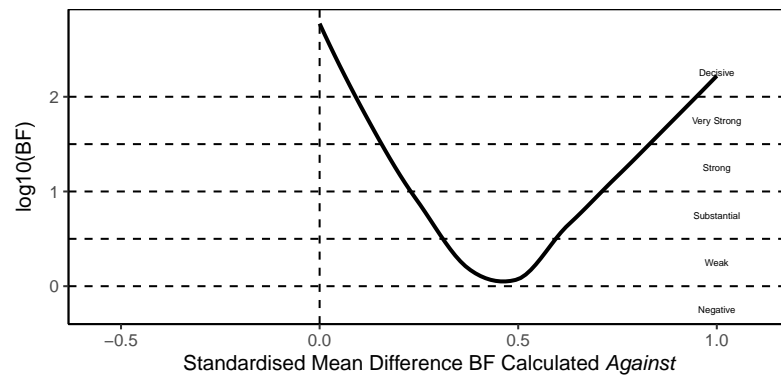Note, thresholds for evidence are indicated for Jeffreys scale

log10(BF)

Decisive
Very Strong
Strong
Substantial
Weak
Negative

Standardised Mean Difference BF Calculated *Against*

Figure 2. Panel (A) shows and ordered forest plot of study level effects. Panel (B) shows the prior and posterior distributions for the overall pooled estimates. Panel (C) shows log10(BF) calculated against each each point effect size from 0 to 1.

(2011). Figure 3 shows each of the prior and posterior distributions for the moderators estimates and the log10(BF) results for each are available in the supplementary materials (see 'plots' folder https://osf.io/dqwh5/). Where there were more substantial changes from prior to posterior these typically revealed a shift in the magnitude of SMD estimate towards the overall pooled estimate from the main model e.g., for fine tasks in the motor demands model (Hatzigeorgiadis et al. (2011) = 0.67 [95% confidence interval: 0.53, 0.82]; posterior pooled estimate = 0.59 [95% quantile interval: 0.47, 0.71]), instructional in the self-talk content model (Hatzigeorgiadis et al. (2011) = 0.55 [95% confidence interval: 0.40, 0.70]; posterior pooled estimate = 0.45 [95% quantile interval: 0.34, 0.56]), instructional/fine in the matching hypothesis model (Hatzigeorgiadis et al. (2011) = 0.83 [95% confidence interval: 0.64, 1.02]; posterior pooled estimate = 0.56 [95% quantile interval: 0.4, 0.73]), for novel tasks in the task novelty model (Hatzigeorgiadis et al. (2011) = 0.73 [95% confidence interval: 0.47, 1.00]; posterior pooled estimate = 0.59 [95% quantile interval: 0.38, 0.8]), and for training interventions in the training model (Hatzigeorgiadis et al. (2011) = 0.80 [95% confidence interval: 0.57, 1.03]; posterior pooled estimate = 0.64 [95% quantile interval: 0.44, 0.83]).
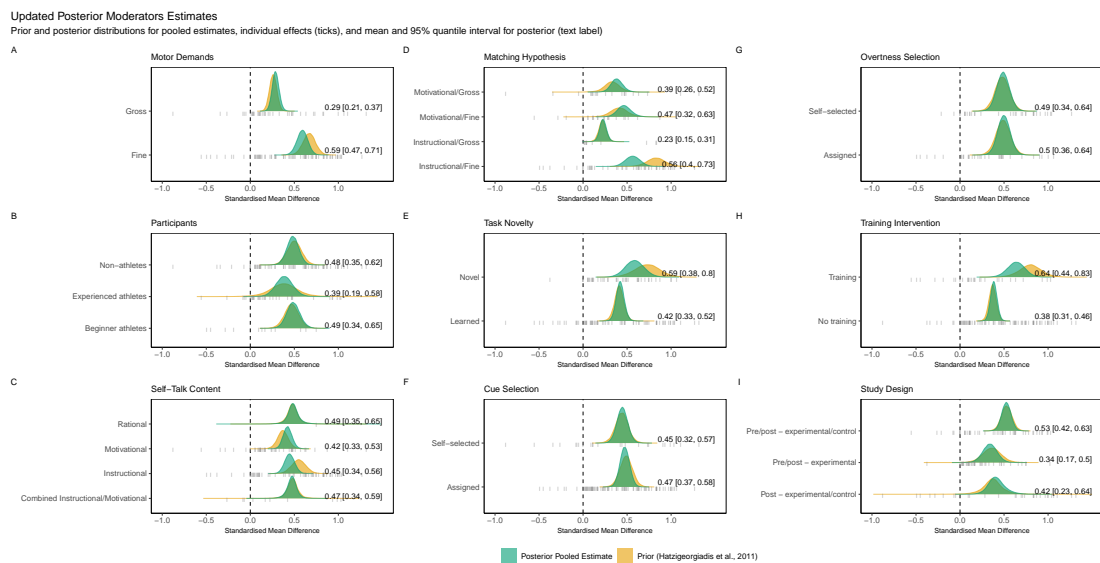

Figure 3. Prior and posterior distributions for each moderator explored; Panel (A) motor demands, (B) participant population, (C) self-talk content, (D) matching hypothesis, (E) task novelty, (F) cue selection, (G) overtness selection, (H) training intervention or not, and (I) study design.

Examining the quality of the evidence and potential questionable research practices

When considering only the new studies published since Hatzigeorgiadis et al. (2011) the frequentist multilevel model estimate (0.45 [95% confidence interval: 0.29, 0.6]) was very similar to both their prior estimate, and also the posterior Bayesian estimate from our updated model (see section above). For the conditional PET-PEESE estimator however the PET estimate was not statistically significant at the $\alpha = 0.10$ level and this adjusted estimate was compatible with a range of estimates from negative effects of similar magnitude to the previously reported positive effect in Hatzigeorgiadis et al. (2011), through to null and trivially positive effects (-0.14 [95% confidence interval: -0.42, 0.13]) suggesting that publication bias/small study effects may be present in this post-2011 literature. These estimates can be seen in the contour enhanced funnel plot in Figure 4.

The Bayesian mixture models assuming the presence of *p*-hacking (note, recall these models ignore the multilevel structure of the effects) also provided reduced adjusted effect estimates, though not to the same magnitude as seen in the PET estimate, with this reduction more prominent when only considering the post-2011 studies and not incorporating the prior from Hatzigeorgiadis et al. (2011). Without the
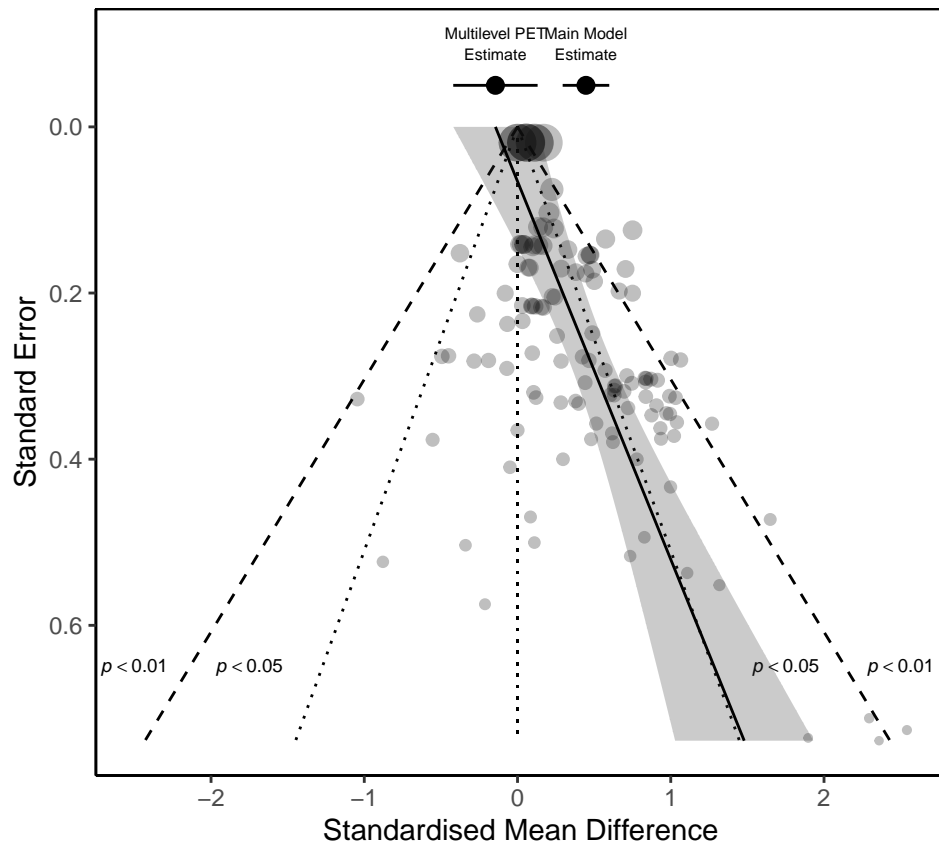
Figure 4. Countour enhanced funnel plot showing effect sizes (scaled in size by inverse sampling variance), thresholds for $p < 0.05$ and 0.01, and both PET-PEESE estimate (in this case it is the PET estimate from the conditional estimator; regression line and ribbon plus "Multilevel PET Estimate" labelled point and interval) and frequentist multilevel model estimate ("Main Model Estimate" labelled point and interval).

prior the estimate from a classic random effects model was 0.34 [95% quantile interval: 0.27, 0.42] and was reduced to 0.2 [95% quantile interval: 0.13, 0.28] assuming the presence of p-hacking. When incorporating the prior from Hatzigeorgiadis et al. (2011) the estimate from a classic random effects model was 0.39 [95% quantile interval: 0.33, 0.45] and was reduced to 0.32 [95% quantile interval: 0.25, 0.4] assuming the presence of p-hacking. These estimates and the posterior distributions can be seen in Figure 5.

When considering only the post-2011 studies and utilising default prior distributions, RoBMA found "strong" evidence against the effect, $BF_{10} = 0.096$ with mean model-averaged estimate of 0.01 [95% quantile interval: 0, 0.11], "strong" evidence in favor of the heterogeneity, $BF_{rf} = 5.872685 \times 10^{16}$ with mean model-averaged estimate $\tau = 0.21$ [95% quantile interval: 0.15, 0.3]. RoBMA without an informative prior also found "strong" evidence in favor of the publication bias, $BF_{pb} = 2.0625228 \times 10^6$. When including the prior from Hatzigeorgiadis et al. (2011) results were similar with RoBMA finding "strong" evidence against the effect, $BF_{10} = 0$ with mean model-averaged estimate of 0 [95% quantile interval: 0, 0], "strong" evidence in favor of the heterogeneity, $BF_{rf} = 1.3766468 \times 10^{23}$ with mean model-averaged estimate $\tau = 0.21$ [95% quantile interval: 0.15, 0.3]. RoBMA with an informative prior also found "strong" evidence in favor of the publication bias, $BF_{pb} = 1.9231176 \times 10^6$.
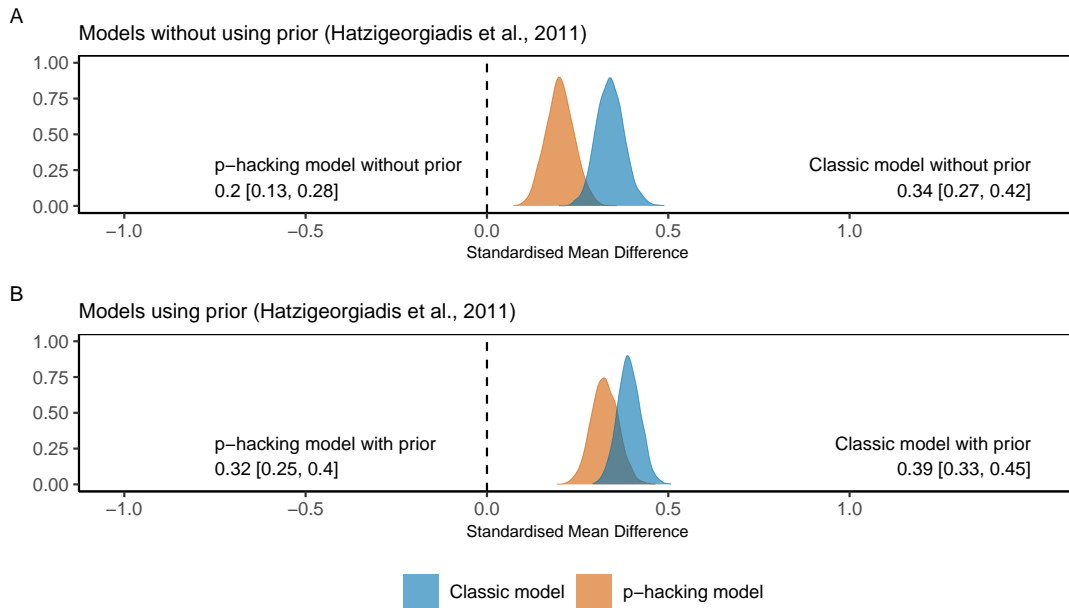
Figure 5. Posterior distributions for Bayesian random effects main model (classic model) and mixture model for *p*-hacking both with (B) and without (A) using the prior from Hatzigeorgiadis et al. (2011).

Discussion

The aim of this work was to demonstrate the application of cumulative evidence synthesis methods including; consideration of the initial probability that a new study of the effects of self-talk interventions would shift our prior belief in their effectiveness, the application of priors taken from the previous meta-analysis to be updated by new studies identified to a new posterior estimate of effect, and consideration of other sources of research waste from questionable research practices such as possible publication bias and *p*-hacking. Such methods, when used prospectively, can aid researchers in determining whether further research of a particular experimental intervention is in fact warranted; and when used in retrospect may reveal where research has been a waste. Given it has been over a decade since Hatzigeorgiadis et al. (2011) published their meta-analysis of self-talk interventions, we used this as an example to demonstrate these methods. We now discuss the implications of our results regarding this example and make suggestions for researchers in the sport and exercise sciences to aid them in planning of research.

Would a new study of a self-talk intervention have changed our prior belief in their effects based on the results of Hatzigeorgiadis et al. (2011)?

As noted, it could be argued that the estimate of the effectiveness of self-talk interventions was sufficiently precise based on the research conducted prior to 2011. The effect estimate from the meta-analysis of Hatzigeorgiadis et al. (2011) has a width spanning 0.2 SMD, and similarly so for many of the moderator estimates. Given this, a researcher considering whether or not to conduct a study of self-talk interventions effects on sports/motor performance could consider whether this is worthwhile by using Bayesian updating with simulated studies to determine both how large an effect, and how large a sample, would be needed to meaningfully shift the prior to posterior distribution. If an unrealistically large effect (whether positive or negative) or an impractically large sample size, or both, would be needed for a new studies effect estimate to shift belief in the effect estimate then it might be considered that it is either implausible or simply not worth the resources to conduct such a study. In such a case it might

be considered that the current estimate of the effect is sufficiently precise. Indeed, this appears to be the case with the effect estimate from Hatzigeorgiadis et al. (2011).

As can be seen in Figure 1, given the precision of the prior distribution from Hatzigeorgiadis et al. (2011), irrespective of the sample size or magnitude of effect and the extent to which it disagreed with the prior estimate, no single new study could shift the posterior distribution to an extent that it didn't still have ~95% of its mass within the 95% quantile interval of the prior distribution. As such, taking the estimate from Hatzigeorgiadis et al. (2011) at face value, we would likely conclude that performing a new study would be a waste of resources. The number of studies since 2011 has roughly doubled and yet no single study *a priori* would have been able to shift belief in the estimate of the effects of self-talk interventions. Of course, it may be that this sheer volume of additional evidence might have updated belief in the estimate of their effects (indeed, this is something that could have also been simulated i.e., how many additional studies with particular effect sizes and sample sizes would be needed to shift belief?). But, had this been considered during study planning, it might have saved researchers in this field considerable resources that could have been directed towards other research questions or programmes.

To what extent have studies published since Hatzigeorgiadis et al. (2011) updated belief in the effects of self-talk interventions?

Although *a priori* the use of simulation is valuable to see whether a new study is worth adding to the body of literature because it will meaningfully change our beliefs in an effect, in the example of self-talk interventions there has already been a considerable addition of evidence to the corpus. In such a situation it's worthwhile to consider the extent to which this new evidence has shifted beliefs. The use of Bayesian updating for meta-analysis, where there is already a previous meta-analytic estimate on which to base a prior distribution, can also be a more efficient means of conducting evidence synthesis as searches and inclusion of studies can be limited to dates *after* the publication of that previous estimate. The extent to which new evidence has shifted beliefs can be quantified and the value of such work can be reflected on.

So, we also updated using Bayesian methods the results of Hatzigeorgiadis et al. (2011) with studies published since 2011. Since that time a further 35 new studies had been published which were identified in our searches and we could include all but one in our analysis. Our findings suggested that the cumulative impact of this research over the last decade and more has done little to further our understanding of the effects of self-talk interventions. The results showed that the overall pooled estimate from the meta-analysis was a SMD of 0.47 [95% quantile interval: 0.39, 0.56]. This was very similar to the previous estimate of overall effect in Hatzigeorgiadis et al. (2011) of 0.48 [95% confidence interval: 0.38, 0.58]. The log10(BF) calculated indicated that the included studies largely reflected "weak", or even very mildly "negative", evidence against effects ranging from 0.31 to 0.59, and only provided "decisive" evidence updating the prior against *a priori* implausible effect sizes ranging from 0 to 0.08 and from 0.95 to 1.

In one sense, the findings of the updated Bayesian meta-analyses do reiterate the positive effect of self-talk interventions on sport/motor performance on average reported by Hatzigeorgiadis et al. (2011). Indeed, the estimate reflects the typical effect of other psychological interventions (0.51 [95% confidence interval: 0.42, 0.58]) as identified by Lochbaum et al. (2022) in their umbrella review; though notably they also reported a wide range of overall effect estimates between positively directed interventions/strategies (0.15 to 1.35). Reflecting this, Hatzigeorgiadis et al. (2011) reported $\tau = 0.27$ (though as noted in footnote[7] it is not clear which level this pertains to) for the self-talk intervention studies they explored, and our study level estimate was not dissimilar to this ($\tau = 0.35$ [95% quantile interval: 0.11, 0.54]). Indeed, our prediction interval ranged from -0.37 to 1.32. Considering this heterogeneity in effects, Hatzigeorgiadis et al. (2011) previously examined varied theoretically plausible moderators of the effectiveness of self-talk interventions which we were also able to update with new evidence. It may be the case that since 2011 the majority of research has focused on understanding exactly how best to employ self-talk interventions.

Hatzigeorgiadis et al. (2011) suggested that, considering theoretically driven moderators, there could possibly be differential effects of self-talk interventions under certain task conditions, based on the nature of the intervention, or for different participant populations. Their results suggested that, whilst self-talk interventions in general were effective, greater effects were seen for fine motor tasks and the performance

of novel tasks. They also found results supportive of the matching hypothesis (i.e., that instructional self-talk as more effective for fine motor tasks than motivational, and that instructional was more effective for fine compared with gross motor tasks). By and large, our moderator analyses reiterated the findings of Hatzigeorgiadis et al. (2011). Posterior estimates were broadly similar for all moderators which were updated, and log10(BF) suggested that the newer evidence provided relatively weak evidence against the prior effects reported by Hatzigeorgiadis et al. (2011) or evidence suggesting slightly smaller effects for certain moderators (though qualitative conclusions remained the same). Where there were clearer shifts in the posterior distributions these were typically towards the overall pooled effect size within the main model possibly suggesting that some factors were not as strong in moderating effects as previously thought.

Despite supporting the effectiveness of self-talk interventions and the factors that moderate this, our results suggest that cumulatively the past decade and more of research has done little to further our understanding of these effects. Of course, as we noted in the introduction, this lack of change to our beliefs despite the cumulative evidence could be interpreted as evidence that the production of such research could be considered wasteful. The present updated Bayesian meta-analyses, and indeed the *a priori* simulations too, makes this quite clear; considering the limited resources and time for conducting research, it may be worth moving onto to other more pertinent questions or research programmes. But, both simulating the impact of a new trial to determine if it is worth performing, or updating a prior meta-analysis estimate with new evidence from subsequent trials, entail the assumption that the previous estimate is not itself biased by questionable research practices (Friese & Frankenbach, 2020). Further, in updating a prior estimate there is the assumption that the newer evidence is also free from such questionable research practices.

Is there evidence of questionable research practices in the self-talk interventions literature?

As explained, we limit our examination of the presence of questionable research practices to the subsequent post-2011 literature and make the reasonable assumption that, given the replication crisis and subsequent methodological reform efforts kicked off proper in the early 2010s (Lakens, 2023), the presence of questionable research practices was likely as bad if not worse in the literature included in Hatzigeorgiadis et al. (2011). The existence of evidence suggesting prior questionable research practices such as publication bias or *p*-hacking might be cause for an evaluation of whether prior research is in fact worth building upon, or whether instead it may be worth starting afresh to determine if there really is an intervention effect whilst incorporating safe-guards for such issues e.g., pre-registration/registered-reports (Chambers & Tzavella, 2022).

The results of our exploration of possible questionable research practices sadly paint a troubling picture for self-talk interventions. There is strong evidence that both publication bias and *p*-hacking seem to be present. The adjusted estimates in all of the methods examined were reduced compared to the unadjusted estimated effect of self-talk interventions with this being most evident considering models of publication bias and to an extent that it was compatible with a null effect (i.e., PET-PEESE and RoBMA model averaging). Granted, in our present example we have only considered the newly produced evidence since 2011 and it might be argued that the studies prior to this may not be subject to such questionable research practices. However, as noted we think it is a reasonable assumption that things were probably as bad, if not worse, prior to 2011. Further, in the Bayesian mixture model for *p*-hacking, and the RoBMA, we were also able to explore adjusted estimates whilst accounting for the prior estimate from Hatzigeorgiadis et al. (2011). Even when considering this we still find strong evidence of questionable research practices influencing this literature.

Has research conducted since Hatzigeorgiadis et al. (2011) been a waste?

Given the results presented we find it hard to conclude anything other than that there has been considerable "research waste". Some might see this conclusion as being quite uncharitable but we think that it is justified regarding the specific context explored here: namely, the literature on self-talk *interventions* and their effects on sports/motor performance. Considering the general lack of impact seen upon the prior estimate from Hatzigeorgiadis et al. (2011) from simulating a new study, and the

relative lack of impact from the updating of this prior estimate with newly conducted studies post-2011, in so far as the effects of self-talk interventions are concerned we have not learned anything new.

An objection to concluding that these post-2011 studies are "research waste" though might be that they may have refined understanding of more specific questions, such as the impact of different moderators on the self-talk interventions effectiveness, which may have been the goals of these new studies (not merely whether self-talk interventions in general improve performance). As mentioned, some had concluded that post-2011 the field had "matured" and, under the assumption that self-talk interventions do indeed improve performance, moved towards theorising and exploring the possible mechanisms for *why* it was effective (Galanis et al., 2016; Hardy et al., 2018). However, as we also saw, many of the theoretically informed moderators that were explored also updated little and where they did they tended to shift closer to the overall posterior pooled effect estimate. Further, considering the results of our exploration of questionable research practices, it is not particularly clear whether the already proven positive effects of self-talk interventions is a reasonable assumption to work from any more, and research addressing the question of *why* the interventions are effective might be built upon a house of cards. Based upon the results presented here it is not clear whether self-talk interventions have any impact on sports/motor performance.

The amount of research since Hatzigeorgiadis et al. (2011) roughly doubled. Had approaches such as those presented here been employed *a priori* by researchers in planning these studies to consider the impact that such evidence might have had on updating prior estimates of self-talk interventions effects, or indeed the quality of those prior estimates, a lot of waste might have been avoided. Granted, some of the methods presented here were not necessarily readily available to understand and utilise by researchers at the time they were conducting their work. But, the example of self-talk interventions can act as a warning to researchers in the sport and exercise sciences about what kinds of questions they should ask themselves prior to beginning their work.

Although the literature on self-talk *interventions* and their effects on sports/motor performance specifically could be considered wasteful, as explained, the post-2011 years have seen theoretical advancements in conceptualisation and operationalisation of the construct of self-talk (Brinthaupt & Morin, 2023; Geurts, 2018; Hardy et al., 2018; Latinjak et al., 2019; Latinjak et al., 2023; Van Raalte et al., 2016, 2019). Alongside the discussion within the broader psychological literature regarding replication concerns there have also been calls to improve theorising (Eronen & Bringmann, 2021) and measurement practices (Flake & Fried, 2020). These are important steps that researchers should consider before they move to testing hypotheses such as whether intervening on a psychological construct produces particular causal outcomes (Scheel et al., 2021). As such, it would be unfair to suggest that work aimed in these specific theoretical/operationalisation directions has been wasteful and attempts at cumulative evidence synthesis which draw such conclusions might also consider qualitative appraisal of the broader literature saving conclusions about "research waste" for the specific context explored. But, whilst it has been argued that cumulative evidence synthesis methods such as meta-analysis to test theoretically derived hypotheses (Hagger & Hamilton, 2024), for self-talk interventions specifically it may be worth starting afresh to determine if there really is an intervention effect whilst incorporating safe-guards for such issues as the questionable research practices explored here e.g., pre-registration/registered-reports (Chambers & Tzavella, 2022).

Conclusion

This work has demonstrated the application of cumulative evidence synthesis methods including; consideration of the initial probability that a new study of the effects of self-talk interventions would shift our prior belief in their effectiveness, the application of priors taken from the previous meta-analysis to be updated by new studies identified to a new posterior estimate of effect, and consideration of other sources of research waste from questionable research practices such as possible publication bias and *p*-hacking. We presented the application of these methods in the example of self-talk interventions for sports/motor performance given it has been over a decade since Hatzigeorgiadis et al. (2011) published their meta-analysis of self-talk interventions. The application of these methods highlight that much of the literature on this topic could be considered "research waste"; the impact of any new study upon the prior estimate from Hatzigeorgiadis et al. (2011) would have had added little evidence to change prior beliefs, the doubling of research conducted since 2011 also added little evidence to update the prior estimate, and

it is likely that much of the research post-2011 (and likely prior to 2011) has been subject to questionable research practices making it unclear whether self-talk interventions affect sports/motor performance or not. This example can act as a warning to researchers in the sport and exercise sciences about what kinds of questions they should ask themselves prior to beginning their work. Such methods as those demonstrated here, when used prospectively, can aid researchers in determining whether further research of a particular experimental intervention is in fact warranted; and when used in retrospect may reveal where research has been a waste. Considering the limited resources and time for conducting research their application may suggest it to be more worthwhile moving onto other more pertinent research questions or programmes. In the case of self-talk specifically, future research might be best placed to continue with the intention of building strong theories regarding the conceptualisation and operationalisation of self-talk and, following this, test if there really is an intervention effect as hypothesised whilst incorporating safeguards for questionable research practices e.g., pre-registration/registered-reports (Chambers & Tzavella, 2022).

Contributions

All authors contributed substantially to conception and design, acquisition of data, analysis and interpretation of data, drafting the article or revising it critically for important intellectual content, and provided final approval of the version to be published.

Peer Review

This manuscript was reviewed and recommended through Peer Community in Health and Movement Sciences (Wolff & Gaveau, 2024).

Funding Information

No funding was received for this project.

Data and Supplementary Material Accessibility

All extracted data and code utilised for data preparation and analyses are available in either the Open Science Framework page for this project https://osf.io/dqwh5/ or the corresponding GitHub repository https://github.com/jamessteeleii/self_talk_meta_analysis_update. Other supplementary analyses and plots are also available there.

References

Abdoli, B., Hardy, J., Riyahi, J. F., & Farsi, A. (2018). A Closer Look at How Self-Talk Influences Skilled Basketball Performance. *The Sport Psychologist*, *32*(1), 9–15. https://doi.org/10.1123/tsp.2016-0162

Bartoš, F., Maier, M., Quintana, D. S., & Wagenmakers, E.-J. (2022). Adjusting for Publication Bias in JASP and R: Selection Models, PET-PEESE, and Robust Bayesian Meta-Analysis. *Advances in Methods and Practices in Psychological Science*, *5*(3), 25152459221109259. https://doi.org/10.1177/25152459221109259

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, *14*(1), 99–116. https://doi.org/10.1002/jrsm.1594

Barwood, M. J., Corbett, J., Wagstaff, C. R. D., McVeigh, D., & Thelwell, R. C. (2015). Improvement of 10-km time-trial cycling with motivational self-talk compared with neutral self-talk. *International Journal of Sports Physiology and Performance*, *10*(2), 166–171. https://doi.org/10.1123/ijspp.2014-0059

Becker, B. J. (2005). Failsafe N or File-Drawer Number. In *Publication Bias in Meta-Analysis* (pp. 111–125). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870168.ch7

Beneka, A., Malliou, P., Gioftsidou, A., Kofotolis, N., Rokka, S., Mavromoustakos, S., & Godolias, G. (2013). Effects of instructional and motivational self-talk on balance performance in knee injured. *European Journal of Physiotherapy*, *15*(2), 56–63. https://doi.org/10.3109/21679169.2013.776109

Biau, D. J., Boulezaz, S., Casabianca, L., Hamadouche, M., Anract, P., & Chevret, S. (2017). Using Bayesian statistics to estimate the likelihood a new trial will demonstrate the efficacy of a new treatment. *BMC Medical Research Methodology*, *17*(1), 128. https://doi.org/10.1186/s12874-017-0401-x

Blanchfield, A. W., Hardy, J., De Morree, H. M., Staiano, W., & Marcora, S. M. (2014). Talking yourself out of exhaustion: The effects of self-talk on endurance performance. *Medicine and Science in Sports and Exercise*, *46*(5), 998–1007. https://doi.org/10.1249/MSS.0000000000000184

Brinthaupt, T. M., & Morin, A. (2023). Self-talk: Research challenges and opportunities. *Frontiers in Psychology*, *14*.

Cabral, L. L., da Silva, C. K., Delisle-Rodriguez, D., Lima-Silva, A. E., Galanis, E., Bertollo, M., Hatzigeorgiadis, A., Villarejo-Mayor, J. J., & Pereira, G. (2023). Motivational self-talk mitigates the harmful impact of mental fatigue on endurance performance. *Journal of Applied Sport Psychology*, *0*(0), 1–19. https://doi.org/10.1080/10413200.2023.2208643

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Chang, Y.-K., Ho, L.-A., Lu, F. J.-H., Ou, C.-C., Song, T.-F., & Gill, D. L. (2014). Self-talk and softball performance: The role of self-talk nature, motor task characteristics, and self-efficacy in novice softball players. *Psychology of Sport and Exercise*, *15*(1), 139–145. https://doi.org/10.1016/j.psychsport.2013.10.004

Clarke, M., Brice, A., & Chalmers, I. (2014). Accumulating Research: A Systematic Account of How Cumulative Meta-Analyses Would Have Provided Knowledge, Improved Health, Reduced Harm and Saved Resources. *PLoS ONE*, *9*(7), e102670. https://doi.org/10.1371/journal.pone.0102670

de Matos, L. F., Bertollo, M., Stefanello, J. M. F., Pires, F. O., da Silva, C. K., Nakamura, F. Y., & Pereira, G. (2021). Motivational self-talk improves time-trial swimming endurance performance in amateur triathletes. *International Journal of Sport and Exercise Psychology*, *19*(3), 446–459. https://doi.org/10.1080/1612197X.2020.1717576

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Friese, M., & Frankenbach, J. (2020). P-Hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456–471. https://doi.org/10.1037/met0000246

Galanis, E., Hatzigeorgiadis, A., Charachousi, F., Latinjak, A. T., Comoutos, N., & Theodorakis, Y. (2022). Strategic Self-Talk Assists Basketball Free Throw Performance Under Conditions of Physical Exertion. *Frontiers in Sports and Active Living*, *4*.

Galanis, E., Hatzigeorgiadis, A., Comoutos, N., Charachousi, F., & Sanchez, X. (2018). From the Lab to the Field: Effects of Self-Talk on Task Performance Under Distracting Conditions. *The Sport Psychologist*, *32*(1), 26–32. https://doi.org/10.1123/tsp.2017-0017

Galanis, E., Hatzigeorgiadis, A., Comoutos, N., Papaioannou, A., Morres, I. D., & Theodorakis, Y. (2022). Effects of a strategic self-talk intervention on attention functions. *International Journal of Sport and Exercise Psychology*, *20*(5), 1368–1382. https://doi.org/10.1080/1612197X.2021.1963304

Galanis, E., Hatzigeorgiadis, A., Zourbanos, N., & Theodorakis, Y. (2016). Chapter 8 - Why Self-Talk Is Effective? Perspectives on Self-Talk Mechanisms in Sport. In M. Raab, P. Wylleman, R. Seiler, A.-M. Elbe, & A. Hatzigeorgiadis (Eds.), *Sport and Exercise Psychology Research* (pp. 181–200). Academic Press. https://doi.org/10.1016/B978-0-12-803634-1.00008-X

Galanis, E., Papagiannis, E., Nurkse, L., Theodorakis, Y., & Hatzigeorgiadis, A. (2023). The effects of strategic self-talk on divided attention following physical exhaustion. *International Journal of Sport and Exercise Psychology*, *21*(5), 883–893. https://doi.org/10.1080/1612197X.2022.2090989

Geurts, B. (2018). Making Sense of Self Talk. *Review of Philosophy and Psychology*, *9*(2), 271–285. https://doi.org/10.1007/s13164-017-0375-y

Glasziou, P., & Chalmers, I. (2018). Research waste is still a scandal—an essay by Paul Glasziou and Iain Chalmers. *BMJ*, *363*, k4645. https://doi.org/10.1136/bmj.k4645

Grainger, M. J., Bolam, F. C., Stewart, G. B., & Nilsen, E. B. (2020). Evidence synthesis for tackling research waste. *Nature Ecology & Evolution*, *4*(4), 495–497. https://doi.org/10.1038/s41559-020-1141-6

Gregersen, J., Hatzigeorgiadis, A., Galanis, E., Comoutos, N., & Papaioannou, A. (2017). Countering the Consequences of Ego Depletion: The Effects of Self-Talk on Selective Attention. *Journal of Sport and Exercise Psychology*, *39*(3), 161–171. https://doi.org/10.1123/jsep.2016-0265

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, *92*, 1–29. https://doi.org/10.18637/jss.v092.i10

Gunnell, K., Poitras, V. J., & Tod, D. (2020). Questions and answers about conducting systematic reviews in sport and exercise psychology. *International Review of Sport and Exercise Psychology*, *13*(1), 297–318. https://doi.org/10.1080/1750984X.2019.1695141

Hagger, M. (2022). Meta-analysis. *International Review of Sport and Exercise Psychology*, *15*(1), 120–151. https://doi.org/10.1080/1750984X.2021.1966824

Hagger, M. S. (2006). Meta-analysis in sport and exercise research: Review, recent developments, and recommendations. *European Journal of Sport Science*, *6*(2), 103–115. https://doi.org/10.1080/17461390500528527

Hagger, M. S., & Hamilton, K. (2024). Testing theory predictions and intervention mechanisms of action in sport and exercise psychology using meta-analysis. *Sport, Exercise, and Performance Psychology*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/spy0000352

Hardy, J., Comoutos, N., & Hatzigeorgiadis, A. (2018). Reflections on the Maturing Research Literature of Self-Talk in Sport: Contextualizing the Special Issue. *The Sport Psychologist*, *32*(1), 1–8. https://doi.org/10.1123/tsp.2017-0141

Hatzigeorgiadis, A., Bartura, K., Argiropoulos, C., Comoutos, N., Galanis, E., & D. Flouris, A. (2018). Beat the Heat: Effects of a Motivational Self-Talk Intervention on Endurance Performance. *Journal of Applied Sport Psychology*, *30*(4), 388–401. https://doi.org/10.1080/10413200.2017.1395930

Hatzigeorgiadis, A., Galanis, E., Zourbanos, N., & Theodorakis, Y. (2014). Self-talk and Competitive Sport Performance. *Journal of Applied Sport Psychology*, *26*(1), 82–95. https://doi.org/10.1080/10413200.2013.790095

Hatzigeorgiadis, A., Zourbanos, N., Galanis, E., & Theodorakis, Y. (2011). Self-Talk and Sports Performance: A Meta-Analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *6*(4), 348–356. https://doi.org/10.1177/1745691611413136

Heene, M. (2010). *A brief history of the Fail Safe Number in Applied Research* (No. arXiv:1010.2326). arXiv. https://doi.org/10.48550/arXiv.1010.2326

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. https://doi.org/10.1111/j.1467-985X.2008.00552.x

Hong, X., Liao, Y., Shi, Y., Qi, C., Zhao, M., & Raalte, J. L. V. (2020). An Empirical Test of the Self-Talk Dissonance Hypothesis: The Effects of Self-Talk Overtness and Personality on Performance. *The Sport Psychologist*, *34*(3), 173–176. https://doi.org/10.1123/tsp.2019-0134

Jeffreys, H. (1998). *The Theory of Probability*. OUP Oxford.

Kolovelonis, A., Goudas, M., & Dermitzaki, I. (2011). The effects of instructional and motivational self-talk on students' motor task performance in physical education. *Psychology of Sport and Exercise*, *12*(2), 153–158. https://doi.org/10.1016/j.psychsport.2010.09.002

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Lakens, D. (2023). *Concerns about Replicability, Theorizing, Applicability, Generalizability, and Methodology across Two Crises in Social Psychology*. OSF. https://doi.org/10.31234/osf.io/dtvs7

Lane, A. M., Totterdell, P., MacDonald, I., Devonport, T. J., Friesen, A. P., Beedie, C. J., Stanley, D., & Nevill, A. (2016). Brief Online Training Enhances Competitive Performance: Findings of the BBC Lab UK Psychological Skills Intervention Study. *Frontiers in Psychology*, *7*.

Latinjak, A. T., Hatzigeorgiadis, A., Comoutos, N., & Hardy, J. (2019). Speaking clearly . . . 10 years on: The case for an integrative perspective of self-talk in sport. *Sport, Exercise, and Performance Psychology*, *8*(4), 353–367. https://doi.org/10.1037/spy0000160

Latinjak, A. T., Morin, A., Brinthaupt, T. M., Hardy, J., Hatzigeorgiadis, A., Kendall, P. C., Neck, C., Oliver, E. J., Puchalska-Wasyl, M. M., Tovares, A. V., & Winsler, A. (2023). Self-Talk: An Interdisciplinary Review and Transdisciplinary Model. *Review of General Psychology*, 10892680231170263. https://doi.org/10.1177/10892680231170263

Latinjak, A. T., Torregrosa, M., & Renom, J. (2011). Combining Self Talk and Performance Feedback: Their Effectiveness With Adult Tennis Players. *The Sport Psychologist*, *25*(1), 18–31. https://doi.org/10.1123/tsp.25.1.18

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Liu, H.-Y., Lu, F. J. H., Hsu, Y., Gill, D. L., Chiu, Y.-H., & Peng, Y.-H. (2022). Interactive Effects on Motor Performance of Mindfulness, Performance Under Pressure, Self-Talk, and Motor Task Characteristics. *Perceptual and Motor Skills*, *129*(2), 307–327. https://doi.org/10.1177/00315125211069042

Lochbaum, M., Stoner, E., Hefner, T., Cooper, S., Lane, A. M., & Terry, P. C. (2022). Sport psychology and performance meta-analyses: A systematic review of the literature. *PloS One*, *17*(2), e0263408. https://doi.org/10.1371/journal.pone.0263408

Lund, H., Robinson, K. A., Gjerland, A., Nykvist, H., Drachen, T. M., Christensen, R., Juhl, C. B., Jamtvedt, G., Nortvedt, M., Bjerrum, M., Westmore, M., Yost, J., Brunnhuber, K., & on behalf of the Evidence-Based Research Network. (2022). Meta-research evaluating redundancy and use of systematic reviews when planning new studies in health research: A scoping review. *Systematic Reviews*, *11*(1), 241. https://doi.org/10.1186/s13643-022-02096-y

Marshall, D. V., Hanrahan, S. J., & Comoutos, N. (2016). The Effects of Self-Talk Cues on the Putting Performance of Golfers Susceptible to Detrimental Putting Performances Under High Pressure Settings. *International Journal of Golf Science*, *5*(2), 116–134.

McCormick, A., Meijen, C., & Marcora, S. (2018). Effects of a Motivational Self-Talk Intervention for Endurance Athletes Completing an Ultramarathon. *The Sport Psychologist*, *32*(1), 42–50. https://doi.org/10.1123/tsp.2017-0018

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Morris, S. B. (2008). Estimating Effect Sizes From Pretest-Posttest-Control Group Designs. *Organizational Research Methods*, *11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Moss, J., & De Bin, R. (2023). Modelling publication bias and *p* -hacking. *Biometrics*, *79*(1), 319–331. https://doi.org/10.1111/biom.13560

Naderirad, N., Abdoli, B., Farsi, A., & Hassanlouei, H. (2023). The effect of instructional and motivational self-talk on accuracy and electromyography of active and passive muscles in elbow joint position sense test. *International Journal of Sport and Exercise Psychology*, *21*(4), 600–615. https://doi.org/10.1080/1612197X.2022.2078854

Osman, N., Lee, E. L. Y., Malek, N. F. A., Asmi, A. M. N., Abdullah, N. F., & Nadzalan, A. M. (2022). The effects of instructional and motivational self-talk on axe kick performance in taekwondo. *Journal of Martial Arts Anthropology*, *22*(5), 42–47.

Panteli, F., Tsolakis, C., Efthimiou, D., & Smirniotou, A. (2013). Acquisition of the Long Jump Skill, Using Different Learning Techniques. *The Sport Psychologist*, *27*(1), 40–52. https://doi.org/10.1123/tsp.27.1.40

Raalte, J. L. V., Cornelius, A. E., Mullin, E. M., Brewer, B. W., Dyke, E. D. V., Johnson, A. J., & Iwatsuki, T. (2018). I Will Use Declarative Self-Talk . . . Or Will I? Replication, Extension, and Meta-Analyses. *The Sport Psychologist*, *32*(1), 16–25. https://doi.org/10.1123/tsp.2016-0088

Raalte, J. L. V., Wilson, L., Cornelius, A., & Brewer, B. W. (2018). Self-Talk in a SCUBA Diving Context. *The Sport Psychologist*, *32*(3), 244–247. https://doi.org/10.1123/tsp.2017-0091

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

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

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., Weber, S., & Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, *12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

Sarig, Y., Ruiz, M. C., Hatzigeorgiadis, A., & Tenenbaum, G. (2023). The Effects of Instructional Self-Talk on Quiet-Eye Duration and Golf-Putting Performance. *The Sport Psychologist*, *37*(3), 201–209. https://doi.org/10.1123/tsp.2023-0023

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755. https://doi.org/10.1177/1745691620966795

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Steele, J., Fisher, J. P., Smith, D., Schoenfeld, B. J., Yang, Y., & Nakagawa, S. (2023). Meta-analysis of variation in sport and exercise science: Examples of application within resistance training research. *Journal of Sports Sciences*, *41*(17), 1617–1634. https://doi.org/10.1080/02640414.2023.2286748

Turner, M. J., Kirkham, L., & Wood, A. G. (2018). Teeing up for success: The effects of rational and irrational self-talk on the putting performance of amateur golfers. *Psychology of Sport and Exercise*, *38*, 148–153. https://doi.org/10.1016/j.psychsport.2018.06.012

Van Raalte, J. L., Vincent, A., & Brewer, B. W. (2016). Self-talk: Review and sport-specific model. *Psychology of Sport and Exercise*, *22*, 139–148. https://doi.org/10.1016/j.psychsport.2015.08.004

Van Raalte, J. L., Vincent, A., Dickens, Y. L., & Brewer, B. W. (2019). Toward a common language, categorization, and better assessment in self-talk research: Commentary on "Speaking clearly . . . 10 years on." *Sport, Exercise, and Performance Psychology*, *8*(4), 368–378. https://doi.org/10.1037/spy0000172

Wallace, P. J., Mckinlay, B. J., Coletta, N. A., Vlaar, J. I., Taber, M. J., Wilson, P. M., & Cheung, S. S. (2017). Effects of Motivational Self-Talk on Endurance and Cognitive Performance in the Heat. *Medicine & Science in Sports & Exercise*, *49*(1), 191. https://doi.org/10.1249/MSS.0000000000001087

Walter, N., Nikoleizig, L., & Alfermann, D. (2019). Effects of Self-Talk Training on Competitive Anxiety, Self-Efficacy, Volitional Skills, and Performance: An Intervention Study with Junior Sub-Elite Athletes. *Sports*, *7*(6), 148. https://doi.org/10.3390/sports7060148

Weinberg, R., Miller, A., & Horn, T. (2012). The influence of a self-talk intervention on collegiate cross-country runners. *International Journal of Sport and Exercise Psychology*, *10*(2), 123–134. https://doi.org/10.1080/1612197X.2012.645135

Williams, D. R., Rast, P., & Bürkner, P.-C. (2018). *Bayesian Meta-Analysis with Weakly Informative Prior Distributions*. PsyArXiv. https://doi.org/10.31234/osf.io/7tbrm

Wolff, W., & Gaveau, J. (2024). Bayesian cumulative evidence synthesis and identification of questionable research practices in health & movement science. *Peer Community in Health and Movement Sciences*, *1*, 100044. https://doi.org/10.24072/pci.healthmovsci.100044

Young, A. D., Hollander, D. B., Baiamonte, B. A., Bowers, A., Hebert, E. P., & Kraemer, R. R. (2023). How Much Does Self- Talk Influence Fatigue? A Comparison of Performance, Perceived Exertion, and Neuromuscular Patterns during High-Intensity Power Cleans: *International Journal of Strength and Conditioning*, *3*(1). https://doi.org/10.47206/ijsc.v3i1.255

Zetou, E., Vernadakis, N., Bebetsos, E., & Liadakis, N. (2014). The effect of self-talk on Tae-kwon-do skills' learning of novice athletes and perceived use of it. *Journal of Human Sport and Exercise*, *9*(1), 124–135. https://doi.org/10.4100/jhse.2014.91.13

Zourbanos, N., Hatzigeorgiadis, A., Bardas, D., & Theodorakis, Y. (2013a). The Effects of Self-Talk on Dominant and Nondominant Arm Performance on a Handball Task in Primary Physical Education Students. *The Sport Psychologist*, *27*(2), 171–176. https://doi.org/10.1123/tsp.27.2.171

Zourbanos, N., Hatzigeorgiadis, A., Bardas, D., & Theodorakis, Y. (2013b). The effects of a self-talk intervention on elementary students' motor task performance. *Early Child Development and Care*, *183*(7), 924–930. https://doi.org/10.1080/03004430.2012.693487