The evolution of coordination: A phylogenetic meta-analysis and systematic review

Liam Keeble¹, Joel C. Wallenberg² and Elizabeth E. Price³

¹Henry Wellcome Building, Medical School, Newcastle Upon Tyne,
NE2 4HH, United Kingdom

²Percy Building, School of English Literature, Language and
Linguistics, Newcastle University, Newcastle upon Tyne, NE1
7RU, United Kingdom

³School of Psychology, Newcastle University

Keywords

Cooperation, Problem-solving, Meta-analysis, Coordination, Cognition

Corresponding author

Liam Keeble: L.Keeble1@newcastle.ac.uk

Abstract

To solve many cooperative problems, humans must have evolved the ability to solve physical problems in their environment by coordinating their actions. There have been many studies conducted across multiple different species regarding coordinating abilities. These studies aim to provide data which will help illuminate the evolutionary origins of cooperative problem solving and coordination. However, it is impossible to make firm conclusions about the evolutionary origins of coordinating abilities without a thorough comparative analysis of the existing data. Furthermore, there may be certain aspects of the literature that make it very difficult to confidently address evolutionary and meta-analytic questions. This study aimed to rectify this by using meta-analysis, phylogenetic analysis, and systematic review to analyse the data already obtained across multiple studies, and to assess the reliability of this data. We found that many studies did not provide the information necessary for meta analysis, or were not comparable enough to other studies to be included in analyses, meaning meta analyses were underpowered or could not be conducted due to low samples of both studies and different species. Overall, we found that many studies reported small positive effects across studies, but the standard errors of these effects frequently traversed zero.

1 Introduction

Two major issues in the field of comparative cognition are: 1) that methods can vary across studies measuring the same variables with different species [1], and 2) that this makes it difficult to conduct thorough, quantitative comparisons between species in many cognitive tasks. The present study will attempt to demonstrate how these problems can be assessed with modern systematic review, and meta-analytic and phylogenetic methods, while also demonstrating potential issues and downfalls of these methods. We will attempt this by pooling results from studies assessing the abilities of different species to coordinate their actions, and assessing these studies in their reliability and rigour.

1.1 The string pull task

The ability to coordinate actions has been essential to the daily activity and survival of human beings throughout their evolutionary history [2, 3, 4, 5]. These abilities have also been discovered to be a part of many an animal's behavioural repertoire [6].

The cooperative string-pull task has been used frequently across a variety of different species to test what factors influence an animal's ability to solve complex coordination problems. To be successful in this task, two individuals must pull on two ends of a string simultaneously to retrieve rewards from a platform. If only one of the individuals pulls on the string without an other, the string will come loose and the task will be impossible to solve [7, 8]. Variations on this task have since been designed to test different species of animal in their abilities to coordinate their behaviour. For example, the simultaneous button pushing task used with dolphins [9].

One important iteration of this task is the delay test, where a partner is held back from the apparatus for a period of time. To be successful, participants must wait for their partner to arrive at the apparatus before pulling. The species that have been found to be successful in this iteration are chimpanzees [7], Asian elephants [10] and kea [11] among others. Unsuccessful species tested

30 using this paradigm include rooks [12], ravens [13], African grey parrots [14] and
31 domestic dogs [15]. In the simultaneous release iteration of the task (where both
32 participants are introduced to the apparatus at the same time), tolerance of a
33 conspecific being in close proximity is found as a good predictor of cooperative
34 success [7, 16, 13].

35 However, some researchers are skeptical as to whether tolerance is the best
36 predictor of success in this task [17]. Some have found that relationship quality
37 or affiliation between the participants is also a good predictor of success both
38 within and across species [18, 19, 20, 21, 12, 7], and other social factors likely
39 do play a fundamental role. But, researchers are yet to assess how these social
40 factors interact, and also whether there are any other social or non-social factors
41 also effecting coordination success. Such findings are yet to be synthesised and
42 compared in a standardised fashion. A meta analytic approach can standardise
43 measures of success both within and across species, and compare these measures
44 at the appropriate level and with the appropriate measures [22].

45 However, social factors may not play as fundamental a role in predicting
46 success in the cooperative string-pull as many researchers believe. Success could
47 in fact be dictated by physical problem solving cognition or inhibition alone, and
48 subjects could be using environmental cues to solve the problem of the task [23].
49 If data from less social species is missing from the comparative literature, then
50 it is impossible to assess the extent to which social factors, as opposed to non-
51 social factors, are responsible for success in the ability to coordinate actions.
52 It is therefore important to assess bias towards testing more social species in
53 the studies using this task. This would provide an empirical demonstration
54 that researchers should be testing a wider range of species from a variety of
55 different social environments, and a stronger foundation from which to direct
56 future research.

57 Recently, a review of this literature has suggested that there is a bias towards
58 testing social animals using this paradigm [24], and the lack of data concerning
59 the performance of less social species makes it difficult to ascertain the effects
60 that living in large groups and having tolerant relationships has on the evolution

of abilities to coordinate actions. This may also lead to (or, alternatively, be the result of) a publication bias towards successful species. Similar biases may also be found in the way different species are tested, the amount of trials in which they are trained on the task, and the number of trials in which they are tested on the task [24]. There is little consensus as to the evolutionary consequences of certain factors (social group size, diet, brain size) on the evolution of the ability to solve coordination problems. This is largely due to there being little quantitative comparison between species. Without such data, it is impossible to draw any conclusions regarding the selective pressures that contribute to the evolution of coordinating abilities.

As such, a systematic review and quantitative meta-analysis are necessary to standardise and rigorously estimate the effects of different methodological approaches across species, to review the methodological rigour and differences across studies, and to attempt to demonstrate how analyses may be conducted if reliable data is present.

1.2 Challenges of analysis

There are several issues in the comparative cognition literature that make drawing reliable conclusions from comparative and/or meta-analytic studies potentially difficult [1]. Differences in the level of analysis (group success or individual success) can mean studies are difficult to compare directly, and can vary the conclusions researchers draw from their results; different sample sizes can make results more or less reliable; and the methods (e.g. number of training and test trials, and apparatus) used by different researchers can make certain subjects, and thus species, more or less likely to be successful.

There is a lot of heterogeneity in the use of the cooperative string-pull task when testing the coordination abilities of different species [24], meaning results are difficult to compare, and studies are rarely of the same quality. Systematic review and, more specifically, quality assessment are useful tools for assessing the quality of studies, where quality is frequently defined as how well a study's

90 methods are matched to its goals [25]. Although quality assessment tools are
 91 most frequently used to assess whether results from studies testing for the same
 92 effect can be synthesised, they also provide a method for assessing whether
 93 studies are comparable, especially where the overall goals of a field, and thus
 94 the goals of individual studies in that field, are to make comparative assessments
 95 possible. These methods are therefore useful for assuring the comparability of
 96 results in the field of comparative cognition/psychology. Although comparisons
 97 between species in a single task will not provide a definitive answer to the
 98 evolution of coordination, tasks do provide a useful unit for comparison. Many
 99 questions in the field will eventually need to compare the performance of species
 100 in the same (or a very similar) task, using the same (or a very similar) design,
 101 and using a standardised measure of success in that task to attain answers,
 102 and future meta-analyses can synthesise evidence along different dimensions to
 103 bolster bodies of evidence further. But, doing so requires that individual studies
 104 ensure high levels of comparability in order to achieve high quality. However, the
 105 quality of existing literature has yet to be extensively assessed in these terms.

106 Due to large heterogeneity between study designs and a lack of comparable
 107 data, meta and phylogenetic analysis is, at present, going to be difficult for
 108 coordination tasks. Thus, it is essential that the sources of such heterogeneity
 109 and the effects this has on conducting meta and phylogenetic analyses need to
 110 be assessed. Quality assessment can identify sources of heterogeneity [25], and
 111 piloting meta analytic and phylogenetic analysis on available data can identify
 112 the effects such heterogeneity can have on analysis. Conducting such analyses
 113 rarely leads to completely firm conclusions. In fact, their primary usefulness
 114 is frequently derived from their ability to identify gaps and sources of bias in
 115 existing literature [26]. Without identifying these gaps and sources of bias, it
 116 remains difficult to identify what studies may need to be conducted in the future
 117 to make reliable comparative analyses possible, where existing studies may need
 118 to be modified and repeated, and what sources of heterogeneity in design need
 119 attention when designing those new and repeated studies.

120 The most productive method for combating these issues is to begin con-

121 structing a dataset, and piloting possible analyses on existing data, whilst also
122 assessing the quality of that existing data. This will allow gaps in the literature,
123 unreliable studies, and plausible analyses to be identified. Thus, the aims of the
124 present study are as follows:

- 125 • To assess the quality of studies using animal coordination tasks.
- 126 • To assess the comparability of these same studies.
- 127 • To generate a comprehensive dataset of results from these studies.
- 128 • To identify any form of bias in these studies.
- 129 • And to determine the plausibility of the application of meta analysis and
130 phylogenetic analysis to data from these studies.

131 The present study will assess the quality of, and pool the results from, studies
132 using coordination tasks. Tests will be conducted to estimate the effects of
133 publication bias and methodological differences (e.g. level of analysis, sample
134 size, length of rope, and number of training and test trials) across studies, and
135 to address some of the following prescient questions regarding differences in
136 coordination success across species using meta and phylogenetic analysis:

- 137 • Does number of trials predict success in a simultaneous-release coordina-
138 tion task?
- 139 • Do social dynamics (affiliation and tolerance) predict success in a simultaneous-
140 release coordination task?
- 141 • Do brain size, social group size, or dietary factors predict success in a
142 delayed-release coordination task?

143 It is unlikely that analyses for the final question will be possible given avail-
144 able data. However, data will still be collected towards this question in the
145 hope that this will identify where future studies can be conducted to make such
146 analyses plausible for the field. Despite limited available data, some data will
147 allow the precision of such analyses to be assessed given available data.

Such a dataset of studies will be highly beneficial to the field, making it easier for researchers to direct their efforts towards gaps in the literature, and hopefully making it possible to assess evolutionary questions with reliable data in the future [26]. As such, the present study will extract data for some statistical analyses, but will also be used as an opportunity to extract some data that will not be used for analysis, but will help researchers to identify gaps in the literature and design studies that will be comparable to existing results for future comparative analysis.

2 Methods

2.1 Search protocol

All searches will be conducted using Google Scholar. Papers will be extracted from the first fifty pages returned from each search with the following truncated terms:

- Cooperat* string-pull
- Cooperat* problem solv*
- Cooperat* partner choice
- Tolera* and cooperat*
- Coordinat* task
- Tolerat* and coordinat*

Each of these phrases will also be used again, but with the instances of ‘cooperat*’ spelt ‘co-operat*’, and ‘coordinat*’ spelt ‘co-ordinat*’, so that the search protocol accounts for the alternative spellings of the terms. The search using the phrase ‘Cooperat* string-pull’ will be repeated once again but with ‘string-pull’ written as ‘string pull’, in case the term is used without the hyphen in any of the literature.

173 This search protocol will be repeated in the databases of the journals Animal
174 Behaviour, Animal Cognition, and Animal Behaviour and Cognition in case the
175 searches in Google Scholar do not return all relevant results.

176 All studies which cite Melis et al. [7], Hirata and Fuwa [8] or Crawford [27]
177 according to Google Scholar will also be included.

178 After these searches are complete, and following PRISMA guidelines [28],
179 all duplicates of studies will be removed, and then abstracts will be screened
180 for studies that are not relevant to the questions of the present review. These
181 non-relevant studies will also be removed from the dataset. The remaining
182 studies will be read, and those that do not meet the inclusion criteria will also
183 be removed.

184 Finally, the best method of assessing publication bias is to compare unpub-
185 lished results with published results [29]. The principal investigators (tradition-
186 ally the final author) of all studies found using the above search protocol will
187 be contacted in an attempt to gather unpublished results for comparison.

188 **2.2 Criteria for inclusion**

189 All forms of publication that searches return that fit the following criteria will be
190 included in the analysis (including theses and conference abstracts). However,
191 sources that are not published in academic journals will be excluded from most
192 analyses of publication bias.

193 To be included in the present meta-analysis, studies must use an apparatus
194 that requires two individuals to coordinate their behaviour in order to be suc-
195 cessful, and must also assure that individuals cannot receive a reward without
196 coordinating their behaviour (e.g. participate in an individual task for reward).
197 This is a necessary generalisation for the comparative element of the study [30],
198 given that a string-pull apparatus is difficult to replicate with certain species
199 that may survive in water and/or lack the anatomical necessities to operate such
200 an apparatus (e.g. dolphins). The apparatus used in a study must also have a
201 mechanism whereby if one individual attempts the task without another then

the task is failed, or researchers must have deemed one individual attempting the apparatus without an other as failure.

Studies must also include at least one of the following measures: a measure of the relationship between tolerance of individuals in a dyad and their success in a simultaneous-release coordination task; a measure of success in a delay-release coordination task; a measure of the relationship between relationship quality/affiliation of individuals in a dyad and success in a simultaneous-release coordination task; or a correlation measure of number of trials and success in a simultaneous-release coordination task.

When testing for a relationship between social dynamics and coordination success, tolerance and affiliation between a dyad can be measured either using observational or experimental methods, but cooperative success must be tested using a simultaneous release string-pull task, where both individuals are introduced to the apparatus at the same time. Tolerance is frequently measured by assessing how willing two individuals are to share food/eat in close proximity [7, 12, 13, 31]. Affiliation is frequently assessed using observational data (e.g. how frequently two individuals are near each other, or how frequently two individuals exhibit affiliative behaviours towards each other) [18, 19, 20, 21].

Studies providing measures of the relationship between trials and coordination success must have tested subjects in a simultaneous-release string-pull task, and the number of trials used must include both number of test and training trials received by an individual.

2.3 Quality assessment

Quality assessment of included studies will take place before data extraction. A quality assessment tool has been created that takes items from the checklist in appendix G of the *Methods for the development of NICE public health guidance* [32]. This checklist forms the basis for assessing the quality of non-randomized studies that report correlations and associations. It has been modified slightly to make it more relevant to the field and studies at hand. We hope that com-

231 parative researchers will use and adapt this tool for their own purposes in future
232 systematic reviews. The questions included are:

233 1. Population

- 234 (a) Is the source population well described?
- 235 (b) Do the selected participants or areas represent the eligible population
236 or area? (e.g. were both sexes represented fairly equally? Were a
237 variety of ages represented fairly equally?)

238 2. Methods

- 239 (a) Was the selection of explanatory variables based on a sound theoret-
240 ical basis?
- 241 (b) How well were likely confounding factors identified and controlled?
- 242 (c) Was the apparatus used appropriate for the species tested?*

243 3. Outcomes

- 244 (a) Were the outcome measures and procedures reliable? (subjective/objective
245 measures, inter-rater reliability)
- 246 (b) Were all the important outcomes assessed?

247 4. Analyses

- 248 (a) Was the study sufficiently powered to detect an intervention effect (if
249 one exists)? (Is power assessed/reported?)
- 250 (b) Were multiple explanatory variables considered in the analyses?
- 251 (c) Were the analytical methods appropriate?
- 252 (d) Was the precision of association given or calculable? Is association
253 meaningful? (Were estimates, standard error, p-values, confidence
254 intervals reported?)

255 5. Summary

- 256 (a) Are the study results internally valid (i.e. unbiased)?
- 257 (b) Are the findings generalisable to the source population (i.e. exter-
- 258 nally valid)?

259 Answers to the aforementioned questions can be coded on the following scale,
 260 taken from appendix G of the *Methods for the development of NICE public health*
 261 *guidance* [32]:

- 262 • ++: ‘...for that particular aspect of study design, the study has been
 263 designed or conducted in such a way as to minimise the risk of bias.’
- 264 • +: ‘...either the answer to the checklist question is not clear from the
 265 way the study is reported, or that the study may not have addressed all
 266 potential sources of bias for that particular aspect of study design.’
- 267 • -: ‘...for those aspects of the study design in which significant sources of
 268 bias may persist.’
- 269 • Not Reported (NR): ‘...for those aspects in which the study under review
 270 fails to report how they have (or might have) been considered.’
- 271 • Not applicable (NA): ‘...for those study design aspects that are not appli-
 272 cable given the study design under review (for example, allocation con-
 273 cealment would not be applicable for case-control studies).’

274 Items followed by a ‘*’ are items that are not originally from the *Methods*
 275 *for the development of NICE public health guidance* [32], but have been added
 276 by the authors of the present study.

277 Entries in the checklist will also include study ID, a description of the study
 278 design, and the name of the quality assessor.

279 Studies will not be excluded from the meta analysis if they are found to
 280 have sources of bias during quality assessment. Doing so would likely make
 281 meta analysis impossible due to lack of data. However, leaving these studies
 282 in the meta analysis despite them being judged relatively unreliable will also

serve our purpose of illustrating the bias and heterogeneity such studies can introduce into a meta analysis, and so they will be flagged in the visualisations and discussions of results.

2.4 Data extraction

Data extraction will be performed by the lead author. Another author will extract data from 10% of the total studies included in the present study, and inter-extractor agreement will be assessed. Data extraction will also be conducted by the lead author a second time one week after initial extraction. If inter-extractor agreement is low (Cohen's Kappa < 0.8) for the first 10%, 30% of all papers will be double coded [33].

The data extracted will include all the relevant measures that allowed a study to be included in the present analysis, alongside the average number of test and training trials per individual, whether or not the study employed repeated measures, the level of analysis (group/individual), the length of the rope used in the task, the sample size, the delay time used in delay-release coordination tasks, the average number of training trials until success, whether the string-pull used is delay-release or simultaneous-release, a description of the apparatus (e.g. string-pull, button-push), whether subjects in a study have participated in a string-pull task in another study, and the conclusion drawn in a delay-release string-pull (were species successful or not?).

If studies provide a measure of success, or a measure of relationship regarding the above variables, but not as a correlation coefficient (where relationships between variables are concerned), or as an odds ratio (where success in the delay task is concerned), then these measures will be converted using functions from the compute.es package [34]. Provided, that is, that the relevant information for doing so is made available by authors. If the primary investigators of studies who are contacted provide unpublished data where it is possible to attain effects for the above measures from that data, then the appropriate analysis will be carried out to do so.

312 2.5 Generating the phylogeny

313 The ape R package will be used to analyse phylogenetic trees [35] constructed
 314 using the phangorn r package [36]. Genetic data for mtDNA of species will be
 315 downloaded from GenBank [37] using the ‘read.GenBank’ ape function. Muscle
 316 software will be used to align genetic sequences [38] via the phyloch R pack-
 317 age [39]. A rooted phylogenetic tree will be constructed (‘upgma’ phangorn
 318 function), and two unrooted phylogenetic trees will be constructed using two
 319 different neighbour joining (using the ‘NJ’ and ‘BIONJ’ functions respectively)
 320 algorithms [40] for comparison.

321 The phangorn r package will also be used to assess the reliability of trees.
 322 Maximum likelihood methods will be used to assess tree reliability, and trees
 323 will then be compared using AIC [36, 41, 42]. The tree with the lowest AIC
 324 will be used to generate lambda correlation structures [43, 44] to be used in
 325 generalised least squares models for testing evolutionary predictors of success
 326 in a coordination task (see ‘Statistical analyses’ subsection). It is important
 327 to note that this is not the best fitting tree with reality, but the tree that fits
 328 the data best of those trees compared. Two different trees will be used in final
 329 analyses: one which includes all species in the meta-analysis, and one which
 330 includes all primates included in the meta-analysis. Since a variety of species of
 331 primate are tested using similar paradigms, the data available likely allows for
 332 more reliable quantitative comparisons at the within-family level.

333 Data for brain size, social group size and dietary breadth will be taken from
 334 several sources [45, 46, 47, 48].

335 2.6 Statistical analysis

336 Multilevel random-effects meta-analysis models will be conducted using the
 337 metafor r package[49]. The ‘rma.mv’ function will be used to conduct the meta-
 338 analysis of results assessing whether the number of trials correlated with coord-
 339 ination success across studies, whether tolerance correlated with coordination
 340 success across studies, whether affiliation correlated with coordination success

across studies, and whether there was a similar effect of success in a delay-release coordination task across studies. Following Dougherty and Guillette [50] and Davies et al. [22], study, species, and phylogeny will be added to these models as random factors.

One potential issue will be with handling data from studies where analysis has been conducted at different levels (individual and group). The above analysis will be repeated but with data from only studies analysing at the individual level, and with data from only the studies analysing at the group level respectively. Where the same individuals have been tested across studies at the individual level, the average of their effect size will be taken and included in the dataset as opposed to including two measures for a single individual.

Study heterogeneity (τ , Q) can be used to assess if studies share a common effect size. Usually, random-effects models are able to assess a true effect across studies, even when we assume the effects of different studies may vary. But, when we also expect there to be variation in the true effects of different groups, a subgroup analysis is necessary. In evolutionary meta-analyses, where different species are tested in one task, it might be expected that there will be a different true effect size for each species, as we likely expect some species to be more successful than others in a specific task or paradigm, and this is valuable information for making evolutionary comparisons. Given that different estimates may be expected for studies of different species, true effect size for an individual species will be difficult to estimate by only analysing all studies together. Therefore, species subgroups will be analysed (when the amount of studies per subgroup exceeds 2) for effect size wherever possible alongside larger analyses, and species will be tested as a moderator variable in all meta-analyses where evolutionary factors are not analysed as moderator variables. This will hopefully make effect size for different species easier to estimate, and will allow analyses to be conducted regarding the differing levels of success of species. However, it is important to note that the number of studies included in these analyses are likely to be small, and results from analyses should be assessed with this in mind. Despite this, we hope this analysis will serve as a useful demonstration

372 of more powerful analyses that could be conducted in the future.

373 Publication bias will be analysed in multiple different ways, and assessments
374 for publication bias will be conducted for all studies testing the success of sub-
375 jects in a delay-release coordination task, and for subgroups of studies testing
376 the success of a single species in a delay-release coordination task if sufficient
377 data is available. P-value distributions will be assessed for all studies testing
378 coordination success in a delay-release string-pull task. Differences between
379 unpublished and published results will be assessed using a random effects meta-
380 analysis model with whether the publication was published or not used as a
381 moderator variable, and study, species, and phylogeny included as random fac-
382 tors [29, 50, 22]. Funnel plots and trim and fill methods will be used to assess
383 whether there are any potential studies missing from the dataset (using the ‘fun-
384 nel’ and ‘trimfill’ metafor functions). The ‘regtest’ function from the metafor
385 package will be used to test for publication biases regarding cooperative under-
386 standing with a mixed-effects meta-regression model for plot asymmetry [49].
387 Orwin’s fail-safe N (calculated using the metafor ‘fsn’ function) will be used
388 to assess the amount of potential studies that would have to exist in the liter-
389 ature, and to have been missed by the present meta-analysis, in order for an
390 effect that does exist to be non-existent in the population [51]. A cumulative
391 meta-analysis (calculated using the metafor ‘cumul’ function) will be conducted
392 using a fixed-effects model ranked by sample size from largest samples to small.
393 This will assess whether studies with smaller samples - which may mean the
394 effect detected is less accurate - are skewing the results of larger sample studies
395 - which should be more precise [29].

396 For an evolutionary analysis of good predictors of coordination success, sev-
397 eral moderator variables will be tested in separate models using the ‘rma.mv’
398 function from the metafor package [49]. These variables will be endocranial
399 volume, population group size and dietary breadth. Each will be used as an
400 individual predictor of success in a delay-release coordination task. Species and
401 study will be included in models as a random factor, given that some species
402 and studies will appear repeatedly in the dataset. Again, this analysis will

403 be repeated but with results from studies analysed at only the individual and
404 group level respectively. This evolutionary analysis will only be conducted if it is
405 deemed feasible after data collection, where feasible means there are more than
406 10 different species that can be included in analysis. Where conducted, these
407 tests will primarily be used to assess the precision (the size of the confidence
408 interval for estimates) of such models based on available data.

409 Generalised least squares models for phylogenetic analysis will be constructed
410 using the ‘gls’ function from the nlme r package [52]. Brownian motion models
411 of evolution generated from the phylogenetic trees will be fitted to these mod-
412 els to account for phylogenetic relationships between the species tested. This
413 will be achieved using the ‘corPagel’ function from the ape r package [35]. The
414 models will use endocranial volume, various measures of population group size
415 and various measures of dietary breadth as individual predictors of success in
416 a delay-release coordination task. Again, if this analysis is not deemed feasible
417 (as defined above) post data collection, then it will not be conducted; and where
418 conducted, it will be primarily used to address the precision (as defined above)
419 of such models with limited data.

420 P-values will be adjusted using the Holm method [53] where tests are: re-
421 peated in meta-analytic models and phylogenetic models; conducted to assess
422 differences between multiple species; conducted on all data and then repeated
423 with subsets of results from studies analysing at the individual and the group
424 levels; and where models are repeated to test publication bias.

425 The ‘forest’ function from the metafor package will be used to produce forest
426 plots, and the ‘funnel’ function will be used to produce funnel plots [49].

427 **3 Results**

428 The pre-registered protocol was carried out with only one change to the search
429 strategy as described in the following paragraph. The only changes made to
430 analysis code was to stop the code for the analyses that could not be conducted
431 - based on details outlined in the pre-registered methods - from running. The

pre-registration can be found at: <https://osf.io/v6qb7>. The final data and data analysis script can be found at: <https://osf.io/hr6ma/>.

3.1 The dataset

463 studies cited Melis et al. [7], 196 studies cited Hirata and Fuwa [8], and 354 studies cited Crawford [27]. The search proposed could not be conducted in the journal Animal Behaviour due to the database not accepting the search format (wildcards). Instead, this search was conducted in the journals Learning and Behaviour, and Behavioural Ecology and Sociobiology. 949 studies were extracted in searches in Behavioural Ecology and Sociobiology. 616 studies were extracted in searches conducted in the Animal Cognition journal database. 193 studies were extracted from the database of the journal Learning and Behaviour.

As shown in figure 1, of the 42 full-text articles assessed for eligibility and removed, 17 were removed because they did not report any of the necessary measures, and 17 were removed because participants could still succeed in a task without coordination. 2 were removed because the results were repeated in other publications, 3 because they employed the wrong iteration of the task for the result reported (i.e. they employed a delay task but an estimate of the relationship between affiliation and success, or a simultaneous release iteration and an estimate of coordination success, or used neither a simultaneous or delay release iteration), and 3 more were removed because they were inaccessible.

50% of studies analysed at the dyadic level, and 50% at the group level. At the group level: 94% were repeated measures; 53% of studies employed both simultaneous string pull and delay string pull, and 29% of studies employed just simultaneous, with 18% employing only the delay iteration of the task; 58% of studies claimed their subjects were successful in a task, 23% claimed no success, and the rest were ambiguous or irrelevant; 94% were journal articles,

Figure 1: A PRISMA diagram of the search protocol

458 and 94% were published. Of studies that drew a conclusion on the success of
459 their study subjects, 72% claimed their subjects were successful. The dataset
460 included eight studies of birds (7 different species: kea, ravens, rooks, grey
461 parrots, orange-winged parrots, blue-throated macaws and peach-fronted para-
462 keets) and nine studies of mammals (5 different species: chimpanzees, dolphins,
463 dogs, marmosets and elephants).

464 16 authors were contacted via email in an attempt to ascertain the existence
465 of any unpublished datasets that may be relevant to our study. 14 authors were
466 initially contacted, and two of those authors recommended we contact another
467 researcher, which we did. If authors did not respond to an initial email, a second
468 email was sent a month after the first. Of the 16 authors contacted, 4 responded
469 stating that they did not have any unpublished material.

470 **3.2 Quality assessment and reliability**

471 Quality assessment finds that while most studies describe the source population
472 well (1a in table 1 - numbers and letters in brackets refer to this same table in the
473 present paragraph), studies are frequently not representative of the population
474 at large (1b). The selection of explanatory variables were usually mostly based
475 on recent theory (2a), and confounding factors tended to be well controlled
476 in studies (2b). Apparatus were generally well adapted to the species tested
477 (2c). Studies were generally split in terms of their tests of rater reliability, with
478 half of the studies reporting these tests and half of the the studies not reporting
479 them (3a). Studies generally reported all important outcomes, and all important
480 outcomes were generally assessed (3b). Studies rarely reported the power of their
481 study designs (4a), but they generally considered multiple explanatory variables
482 in analyses (4b) and analytical methods were usually appropriate (4c). Studies
483 mostly reported the precision of effect sizes, despite the quality of reports being
484 variable between studies (4d). In summary, quality assessment found that study
485 results were internally valid (5a), but often were not very generalisable to larger
486 populations (5b).

Table 1: A table showing how many articles were coded in each category for each quality assessment question.

Question	++	+	-	NR
Population				
1a) Is the source population well described?	6	11	0	0
1b) Do the selected participants or areas represent the eligible population or area?	2	5	10	0
Methods				
2a) Was the selection of explanatory variables based on a sound theoretical basis?	1	15	1	0
2b) How well were likely confounding factors identified and controlled?	1	10	6	0
2c) Was the apparatus used appropriate for the species tested?	15	2	0	0
Outcomes				
3a) Were the outcome measures and procedures reliable?	8	1	0	8
3b) Were all the important outcomes assessed?	12	4	1	0
Analyses				
4a) Was the study sufficiently powered to detect an intervention effect?	0	0	1	16
4b) Were multiple explanatory variables considered in analyses?	4	8	5	0
4c) Were the analytical methods appropriate?	2	12	3	0
4d) Was the precision of association given or calculable?				
Is association meaningful?	3	6	8	0
Summary				
5a) Are the study results internally valid?	1	13	3	0
5b) Are the findings generalisable to the source population?	0	3	14	0

487 In some cases, where studies used several training conditions, the number
488 of training trials was recorded as the number of trials in which an individual
489 could interact with the string-pull apparatus prior to the trial used to collect
490 data on the result of interest for the meta-analysis. In some cases, some rope
491 lengths and delay times were made incremental in studies, and were recorded
492 in the present dataset as such. Risk ratios were calculated from comparisons of
493 success rates to an imaginary population of the same size with a success rate
494 of 50% to provide effect sizes for success in the delay task that are appropriate
495 for meta analysis. An external coder (RW) coded 5 papers from the dataset
496 and reliability analyses were conducted. Reliability between coders was mostly
497 high, as illustrated in table 2. However, the sample was not large enough to
498 provide highly reliable estimates of inter-rater similarity.

Table 2: Table of results from reliability analysis. Results are proportion of the exact same answers between coders. Note that, for continuous variables, this is often different from the test for reliability, which estimates how closely correlated coders were in their coding of continuous measures as well as agreement between coders. Reliability is the results from the reliability analysis. κ values are weighted κ values.

Variable	Result	Reliability
Experience estimate	4/5	ICC=1, p<0.001
P values of delay estimates	3/5	ICC=0, p=0.49
Success in delay task estimates	2/5	ICC=1, p<0.001
No. of training trials	3/5	ICC=0.69, p=0.047
No. of test trials	4/5	ICC=1, p<0.001
Delay task rope length	5/5	κ =NA
Sample size	4/5	ICC=0.71, p=0.038
Delay time	4/5	ICC=0.92, p=0.0017
Success rate	3/5	ICC=1, p<0.001
Trial length	4/5	ICC=0.8, p=0.017
Simultaneous and/or delay task	5/5	κ =0.64
Previous participation	5/5	κ =0
Conclusion code	5/5	κ =1
Publication type	5/5	κ =1
Publication status	5/5	κ =0.56
Analysis level	4/5	κ =0
Repeated measures	5/5	κ =0

3.3 Phylogenetic trees

Comparisons between phylogenetic trees found that the tree with the lowest AICc was the tree constructed using neighbour joining tree estimation. The difference between the AICc of this tree and the next best (constructed using the unweighted paired group method) was 25.6.

3.4 Publication bias

There was substantial heterogeneity between levels of success in studies using the delay iteration of a coordination task ($\tau=0.2$, $Q=13.4$, $df=6$, adjusted $p=0.12$). The model found an estimate of success of 0.38 ($se=0.11$, $z=3.47$, adjusted $p=0.009$), as shown in figure 3. Figure 4 illustrates the change in effect size

as the effect sizes of studies using smaller and smaller samples are added to the same meta analysis. A trim-and-fill method applied to this model finds the same estimate of 0.38 (se=0.11, $z=3.47$, adjusted $p=0.009$). This model also estimated a similar level of heterogeneity when accounting for unpublished results ($\tau=0.2$, $Q=13.4$, $df=6$, adjusted $p=0.12$). However, a regression test for funnel plot asymmetry finds a z value of -3.55 (adjusted $p=0.009$), and finds a limit estimate of 1.47 (ci: 0.88-2.06). Similarly, a rank correlation test finds a τ value of -0.98 (adjusted $p=0.009$), and 0 missing studies. A fail-safe N analysis estimates that no extra studies, with an average effect size of 0.28 and a target effect size of 0.14, would need to be added to the analysis before the effect size shrunk to a trivial value.

Figure 2: Estimates of success for different species in delay iterations of coordination tasks.

Figure 3: Cumulative estimates of the success of different species in a delay iteration of coordination tasks, added by largest to smallest sample sizes.

520 No unpublished data were found when authors of studies were contacted
521 via email. Only one unpublished study was identified during searches, meaning
522 there is only one unpublished study in the analysed dataset. A model assessing
523 the difference between published and unpublished studies finds a significant dif-
524 ference between the two (estimate=-8.86, se=2.68, z=-3.31, adjusted p=0.009).
525 There is significant heterogeneity between both the moderator (Q=10.96, df=1,
526 adjusted p=0.009) and studies in general (Q=13.4, df=6, adjusted p=0.12).

527 3.5 The plausability of meta analyses

528 One analysis that could be conducted assessed the relationship between experi-
529 ment with a task and coordination success. As shown in figure 2, the relationship
530 over all extracted results between experience with the task and success in the
531 task was 0.4 (se=0.2, $z=2$, adjusted $p=0.28$). There was significant heterogene-
532 ity ($Q=40.59$, $df=8$, adjusted $p=0.01$), but a lot of this heterogeneity could
533 be accounted for by including species in the model as a moderator ($Q=39.95$,
534 $df=5$, adjusted $p=0.01$). When species was included, heterogeneity in effects
535 was reduced ($Q=0.63$, $df=3$, adjusted $p=1$).

536 When data from different dyads (as opposed to different groups) partici-
537 pating was synthesised, the relationship between experience and success was
538 high (est.=0.96, se=0.09, $z=10.47$, adjusted $p=0.01$), and heterogeneity in ef-
539 fects was low ($Q=0.5$, $df=2$, adjusted $p=1$). When analysing results of studies
540 that analysed data at the level of the whole group of subjects, the relationship
541 between experience with the task and cooperative success was 0.29 (s.e.=0.17,
542 $z=1.78$, adjusted $p=0.48$), heterogeneity was estimated at $Q=9.24$ ($df=5$, ad-
543 justed $p=0.5$). Species did not have a significant effect as a moderator in this
544 analysis ($Q=9.24$, adjusted $p=0.5$, $df=5$). But, when species was included as a
545 moderator variable, heterogeneity between study effect sizes was very low ($Q=0$,
546 adjusted $p=1$).

547 The other models proposed in the pre-registration could not be conducted
548 due to the samples of studies for those analyses being too low. These analyses
549 were assessments of the relationships between cooperative success and tolerance,
550 and cooperative success and affiliation, and meta analyses of studies analysing
551 at the dyadic level. Furthermore, phylogenetic analyses could not be conducted
552 since the number of different species in analyses never exceeded the required 10
553 different species. Analyses of individual species could not be conducted due to
554 low frequencies of different studies on any single species.

Figure 4: A forest plot of estimates for the relationship between experience with a coordination task and success in that task.

4 Discussion

The present study finds that, whilst studies of animal coordination are largely internally valid, reports of statistical power and inter-rater reliability are few or variable, and many samples used are not representative of animal populations. Unsurprisingly, where statistical analyses could be conducted, it was found that species differences accounted for a lot of variation in results. Only two of our research questions could be assessed: The relationship between number of trials and success in a coordination task; and the presence of publication bias in the field. The success of most species included in the analysis was correlated with their experience with the coordination task. Statistical tests for publication

565 bias find little evidence of such, despite the one unpublished study that could
566 be included in analyses providing different findings to the published studies.
567 But, these tests may be misleading due to the confirmatory nature of typical
568 assessments of animal cognition.

569 **4.1 Comparability between animal coordination studies**

570 Only 17 studies could be included in quality assessment due to many studies
571 not meeting the inclusion criteria. Even then, only some of these studies could
572 be included in certain analyses. The maximum number of studies that could be
573 included in any one analysis was 6. This illustrates the disparities in methods,
574 goals and reported outcomes of different studies of animal coordination. Al-
575 though it may have been possible to widen inclusion criteria, and thus include
576 more studies in some analyses, this would have limited the specificity of the
577 questions that could have been asked. Ideally, future studies should use more
578 standardised study designs when aiming to assess comparative, evolutionary
579 questions, making specific meta analyses more plausible [54, 46]. Tinbergen's
580 four questions can be used to distinguish when studies should be designed similar
581 to others, or can be more flexible in their design [55]. When studies are asking
582 evolutionary questions (as opposed to mechanistic, functional, or developmental
583 questions), these can only ultimately be answered by quantitative comparisons.
584 Although qualitative judgements can be useful, quantification of species perfor-
585 mance allows a much clearer measure for comparison. Thus comparability in
586 study design and reported results is of the utmost importance when aiming to
587 answer such questions.

588 **4.2 Quality assessment and reliability**

589 Quality assessment finds that one author of the present study did not have
590 reasons to doubt the internal validity of most studies of animal coordination
591 assessed. We also find that it is difficult to ensure that studies are externally
592 valid. However, we are limited in the generalisability of our conclusions due to

the subjective nature of the assessment. External validity is a known problem in comparative cognition generally [30], and is likely due to the populations of interest being difficult to access. Species are frequently represented in the research literature by small groups of the same, fairly idiosyncratic [50] individuals, and the same is true of the papers assessed in the present study. Furthermore, studies rarely reported statistical power. This may be due to low sample sizes frequently being a practical necessity in the field. However, studies with low sample sizes can still be high powered if appropriate study designs and statistical procedures are used [56]. If samples remain unrepresentative, this may make assessing the larger evolutionary questions exceedingly difficult. Given that the individual rather than the group may be the most reliable unit of interest in assessments of cognition [50, 30], comparisons at the species level may be unreliable due to potential individual differences in the population that are not present in the sample. Conducting analyses at both levels will likely produce the most informative and useful results [30] for assessing all levels of evolutionary questions [55].

Inter rater reliability was high for most coded variables. This is reassuring, given that recent trials of meta research in comparative cognition have been less successful on this front when it comes to coding papers in the field [1]. That the number of training trials had a slightly lower ICC value than most other variables is telling. This is a difficult variable to code for. Training trials varied a lot between both individual subjects within studies and also between studies. Furthermore, since statistical tests of success in a delay task are frequently conducted using various different methods both within and between papers, this may account for the low reliability of p values from those statistical tests. However, given the low sample size in reliability analyses, reliability of these analyses themselves is likely low.

620 4.3 Publication bias

621 Our statistical tests for publication bias (namely Orwin’s fail-safe N, a rank
622 correlation test, and the trim-and-fill method applied to the meta-analysis of
623 success measures in the delay task) relied on a small sample of studies which
624 themselves rely on small, heterogeneous samples. Thus we can infer very little
625 about the existence of publication bias in studies assessing animal coordination
626 from our statistical tests. This does not mean that there is no publication bias,
627 but that our sample is simply too small to make any such claims, and that
628 methods of assessing publication bias in a field of such heterogenous studies
629 and samples must be more carefully considered in future. Similarly, the lack of
630 unpublished literature in our dataset does not necessarily mean that such re-
631 sults are non-existent. In fact, it could be a worrying indicator of low retention
632 of datasets and results from unpublished studies, although this claim should be
633 assessed with further research. Overall effect size of success in a delay task was
634 small, and occasionally negative, or standard errors ranged from negative to pos-
635 itive. A cumulative meta analysis does suggest some pull towards no effect being
636 a result of studies with small samples, and the available unpublished material
637 did report much less successful performances from animals than the published
638 literature (but only 1 unpublished manuscript was obtained). Unfortunately,
639 a cumulative meta analysis is limited in this case by the fact that statistical
640 power in the studies assessed is less affected by sample size and more reliant
641 upon repeated measures. As such, the results of the meta analysis are conser-
642 vative. Until unpublished literature can be empirically assessed, the presence
643 of publication bias remains uncertain. If anything, the results from statistical
644 tests of publication bias suggest that these tests may be inappropriate for the
645 results tested in the present study. The field is largely based on performing con-
646 firmatory tests, with little room for disconfirmatory analyses [57]. For example,
647 statistical analyses of the behaviour of animals in cognitive tasks frequently
648 frame the question in terms of whether there was a significant difference from
649 chance in their behaviour. However, there is no measure of the extent of failure

in the actual task, and an animal would be expected to perform at a chance level if they did not understand the task. This means that an animal can only perform at chance or better than chance, but rarely worse. Thus, it would be unsurprising to find that animals quite frequently demonstrate small positive rates of success in the task over many repeated tests despite not really ‘understanding’ the task. This would be some evidence towards a potential systematic, normally distributed bias towards small, false positive effects. Because many of the effect sizes extracted in the present study were small, statistical tests of publication bias may still find that results are evenly distributed, and thus not skewed (biased) towards a specific result, despite the effects found being small false positives. This is potentially evidence of some systematic bias primarily in study design, rather than at publication.

The above statistical results do clash somewhat with the conclusions drawn by authors. Published authors mostly provided clear conclusions regarding the success or failure of species tested in coordination tasks. This contrasts with recent findings in the field of animal physical cognition, where the conclusions drawn by authors were frequently ambiguous [1]. We also found that the proportion of studies attributing success to species tested in their coordination abilities was similar to that same survey of the conclusions drawn in studies of animal physical cognition [1]. That is, authors in animal physical cognition (when they do draw clear conclusions) are doing so at a rate similar to researchers studying animal coordination. Unfortunately, this uniformity tells us little at present, but may be a fruitful area of research in the future. It could be a signal of systematic bias in study designs, analyses and results, or it could be that the sample of species tested across both domains are simply proportionally good at both coordination and physical tasks, amongst other potential explanations. The latter is an interesting finding in and of itself. However, if the effect is a signal of bias, this should be checked.

But, despite some authors’ claims, the effect sizes extracted from studies of animal coordination did not always correspond to success. Frequently, error values traversed zero, which means evidence may not be extraordinary enough

681 to attribute extraordinary cognitive abilities (such as being able to understand
682 a complex cooperative problem) to the species tested [30]. This may be further
683 evidence of a systematic bias in study designs due to a lack of methods for
684 disconfirming theory and hypotheses. Note that the rate of successful claims
685 from authors is quite high (72%), despite a high proportion of low positive effect
686 sizes. Although, assessments of other subfields of comparative cognition should
687 assess the presence of such phenomena based on whether the subfield does make
688 use of potentially biased study designs. A change in statistical approach could
689 help to eliminate this bias by assessing the proportion of data in support of
690 success vs. the proportion of data against success, as demonstrated in a recent
691 study with marmosets [58]. There is likely substantial between species variation
692 in measures of failure at a task. Until such measures of failure are developed and
693 widely applied, highly reliable quantitative comparisons will be difficult. The
694 present study illustrates this difficulty, with a handful of species performing at
695 a similar low level of success.

696 Phylogenetic (meta) analyses may provide a further potential tonic to such
697 issues for certain questions in comparative cognition and psychology [59, 46].
698 Conducting more comparable studies across species and using data from these
699 studies to perform phylogenetic analyses would allow researchers to assess which
700 species perform better than others in a certain task, as opposed to basing con-
701 clusions simply on the success or lack of success of a single species. The complex
702 nature of cognition and factors affecting success in such tasks means that asking
703 whether or not a certain species ‘understands’ a task is a much more compli-
704 cated question than a simple binary outcome can answer. However, information
705 regarding which species perform better in a task than others will perhaps pro-
706 vide a more fruitful avenue for understanding factors involved in task success,
707 both evolutionary and cognitive.

708 4.4 The plausability of meta analyses

709 One result found in the present study was a relationship between experience with
710 a coordination task and success in that task. This result was fairly consistent
711 across most species included in the analysis, yet there was also heterogeneity
712 across species, as expected. This suggests that all species included in the analysis
713 had mechanisms available to them to decipher successful behavioural strategies
714 for the task, but that some species were more efficient learners than others.
715 However, exactly what was being learned/adapted to cannot be concluded. The
716 high level of success and low level of heterogeneity in the analysis when including
717 only analyses at the dyadic level is likely due to these results coming from a
718 single species (*Psittacus erithacus*). This species showed a high relationship
719 between experience and success. However, in the studies testing this species,
720 subjects were given high amounts of training, and were trained until they had
721 completed the task successfully in 90% of trials. The range of total training trials
722 per subject across studies included in the present study was 1-120, and some
723 could not be specified due to variable numbers of training trials or unspecified
724 total numbers of training trials. This illustrates an obvious bias across species
725 arising from differences in study design. It represents a significant barrier to
726 true quantitative comparisons in the field. Such conclusions can also be affected
727 by researchers analysing their subjects at different levels (analysing per dyad
728 vs. analysing over the whole group of subjects). Synthesising results over all
729 studies (both those analysed at group and dyadic levels) returns a higher overall
730 effect size than analyses conducted at only the group level. Since the meta
731 analysis treats such analyses at the dyadic level as individual, low powered
732 studies, if the effects found at the dyadic level are high then they can pull
733 the overall effect size in that direction. This again illustrates the potential
734 affect that individual differences within species can have on potential future
735 phylogenetic meta analyses [50, 30]. This is another potential barrier to making
736 reliable quantitative comparisons between species. When different species are
737 analysed at different levels, it is difficult to assess whether the differences in

effect sizes are due to individual or species differences. Thus quantitative inter-species comparison is difficult to interpret in the present analysis. Although analysing at the level of the individual or dyad might be useful for assessing individual differences both within and between species, if authors are attempting to contribute data towards asking evolutionary questions, then analysing at both levels is the most useful. Phylogenetic meta analyses will require measures of heterogeneity both within and between species to draw any reliable conclusions. At present, large scale quantitative comparisons of animal coordination remain implausible largely due to variability in study design and level of analysis.

4.5 Difficulties of data synthesis

The primary source of difficulty for us when conducting the present study was caused by the reporting of methodological details and results. Coding for the methodological details of studies was often hindered by researchers detailing these procedures in different ways and across different study designs. For example, it was difficult to determine the number of training and testing trials an animal had participated in, sometimes due to a lack of transparency in methods, but frequently due to some individuals participating in multiple experimental conditions of different kinds. Often, this was despite the fact that such papers were testing the same hypotheses across different species. This was a difficult barrier when trying to determine comparability.

Similarly, large variations in the reporting of statistical results meant that, quite often, effect sizes were not reported or calculable from reported results. No possible way of determining an effect size in these studies meant that many studies that could have been comparable in design could not be compared to those that did report effect sizes and their precision. A lot of this lack of effect sizes can be explained by a large reliance on significance testing statistical practices, which emphasise the reporting of a significance test statistic and p-value, with less emphasis on the reporting of effects. Even when we could extract effect sizes, we quite frequently had to calculate many of the effects of importance to

us directly from raw data or convert between effect sizes. Standardising, at least to some extent, the effect size of interest in comparative studies can not only make them more comparable and make meta analyses more plausible, but can also clarify exactly what researchers are trying to compare across this currently diverse literature. Thus, reporting effect sizes, and reporting them in a clear and reliable fashion, is of the utmost importance for studies in comparative cognition.

4.6 Conclusions

At present, results in studies of animal coordination are just not yet comparable enough to conduct reliable phylogenetic analyses. In the small field of animal coordination, studies need to be replicated to ensure validity of conclusions for populations of the samples studied, and more species need to be studied more systematically and with more standardised methods in order to make asking evolutionary questions, quantitative comparisons, and phylogenetic analyses possible. This would be the ideal situation, but we recognise that there are extensive practical barriers to such research. Large and varied samples are often difficult to access, especially where species are endangered or difficult to keep in captivity, and studying species in the wild, despite being more ecologically valid, means a loss of experimental control. Comparative cognition as a field at large has faced similar issues in the past [60, 61]. Thankfully, some subfields of comparative cognition have been able to make some reliable quantitative comparisons [59, 46]. Furthermore, initiatives like the Many Primates project are making such analyses more possible for certain animal taxa [54]. However, to make reliable estimates of the effects of selective pressures vs. evolutionary history on certain cognitive traits, data from species from across the evolutionary bush are required [62]. Phylogenetic methods are integral to these efforts. They can be used to synthesise results from both highly controlled experiments and studies with animal groups in the wild. They can also give some added control to studies with wild groups, whilst also accounting for evolutionary relationships

796 when testing species experimentally.

797 It is important to continually assess the specific challenges and possibili-
 798 ties that each subfield and comparative paradigm presents. We hope we have
 799 demonstrated the importance of this for animal coordination paradigms in the
 800 present study. We also hope that we have demonstrated how methods of sys-
 801 tematic review (quality assessment in particular), and meta analysis can be used
 802 to assess the extent to which comparative questions can reliably be asked at a
 803 given point in time. This will be an ongoing process until it is deemed that there
 804 are enough results that are both reliable and similar enough to be synthesised
 805 into a phylogenetic analysis. Even then, such quantitative analyses should be
 806 continually assessed in this manner to detect (and ultimately mitigate) bias and
 807 increase reliability. This is a vital result as it can be used to direct research
 808 in comparative cognition in a reliably productive and specific direction. Thus,
 809 at present, comparative and evolutionary questions cannot be asked of animal
 810 coordination studies using the data available in the current, sampled literature.
 811 Although many of these existing studies can be used to assess other questions of
 812 Tinbergen's [55], the evolutionary questions cannot be assessed quantitatively
 813 based on the currently available data. Further, large scale collaborative research
 814 can remedy this situation [54, 59].

815 Data Accessibility

816 Data and analysis script for this study is available on the Open Science Frame-
 817 work: <https://osf.io/hr6ma/>.

818 Declaration of interests

819 The authors report no conflicts of interest regarding the authorship or publica-
 820 tion of this article.

821 Authors' contributions

822 The conceptual and methodological details of the present study were developed
823 by all authors. LK carried out the search strategy, data extraction and sta-
824 tistical analyses. The results were written by LK. All authors contributed to
825 the discussion. An individual external to the study authors double coded a
826 proportion of papers collected for the study.

827 Acknowledgements

828 The authors would like to thank Ben Farrar and one other, anonymous reviewer
829 for their very helpful comments on study protocol and the paper itself. We
830 would also like to thank Robin Watson (RW) for double coding part of our
831 dataset.

832 Funding

833 This work was supported by the NINE Doctoral Training Partnership.

834 References

- 835 [1] B. G. Farrar, D. M. Altschul, J. Fischer, J. van der Mescht, S. Placì, C. A.
836 Troisi, A. Vernouillet, N. S. Clayton, L. Ostojić, Trialling meta-research in
837 comparative cognition: Claims and statistical inference in animal physical
838 cognition, *Animal behavior and cognition* 7 (3) (2020) 419.
- 839 [2] E. Fehr, U. Fischbacher, The nature of human altruism, *Nature* 425 (6960)
840 (2003) 785.
- 841 [3] M. Tomasello, A. P. Melis, C. Tennie, E. Wyman, E. Herrmann, Two Key
842 Steps in the Evolution of Human Cooperation, *Current Anthropology* 53 (6)
843 (2012) 673–692. doi:10.1086/668207.
844 URL <http://www.journals.uchicago.edu/doi/10.1086/668207>

- 845 [4] R. Boyd, P. J. Richerson, Culture and the evolution of human coopera-
846 tion, *Philosophical Transactions of the Royal Society B: Biological Sciences*
847 364 (1533) (2009) 3281–3288. doi:10.1098/rstb.2009.0134.
- 848 [5] A. P. Melis, D. Semmann, How is human cooperation different?, *Philosoph-*
849 *ical Transactions of the Royal Society B: Biological Sciences* 365 (1553)
850 (2010) 2663–2674. doi:10.1098/rstb.2010.0157.
- 851 [6] T. Clutton-Brock, Cooperation between non-kin in animal societies, *Nature*
852 462 (7269) (2009) 51.
- 853 [7] A. P. Melis, B. Hare, M. Tomasello, Engineering cooperation in chim-
854 panzees: tolerance constraints on cooperation, *Animal Behaviour* 72 (2)
855 (2006) 275–286. doi:10.1016/j.anbehav.2005.09.018.
- 856 [8] S. Hirata, K. Fuwa, Chimpanzees (pan troglodytes) learn to act with other
857 individuals in a cooperative task, *Primates* 48 (1) (2007) 13–21.
- 858 [9] K. Jaakkola, E. Guarino, K. Donegan, S. L. King, Bottlenose dolphins
859 can understand their partner’s role in a cooperative task, *Proceedings*
860 *of the Royal Society B: Biological Sciences* 285 (1887) (2018) 20180948.
861 doi:10.1098/RSPB.2018.0948.
862 URL [http://rspb.royalsocietypublishing.org/lookup/doi/10.](http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2018.0948)
863 [1098/rspb.2018.0948](http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2018.0948)
- 864 [10] J. M. Plotnik, R. Lair, W. Suphachoksahakun, F. B. M. de Waal, Elephants
865 know when they need a helping trunk in a cooperative task, *Proceedings*
866 *of the National Academy of Sciences* 108 (12) (2011) 5116–5121.
- 867 [11] M. Heaney, R. D. Gray, A. H. Taylor, Keas perform similarly to chim-
868 panzees and elephants when solving collaborative tasks, *PLoS ONE* 12 (2)
869 (2017) 1–13. doi:10.1371/journal.pone.0169799.
- 870 [12] A. M. Seed, N. S. Clayton, N. J. Emery, Cooperative problem solving in
871 rooks (*Corvus frugilegus*), *Proceedings of the Royal Society B: Biological*
872 *Sciences* 275 (1641) (2008) 1421–1429. doi:10.1098/rspb.2008.0111.

- [13] J. J. Massen, C. Ritter, T. Bugnyar, Tolerance and reward equity predict cooperation in ravens (*Corvus corax*), *Scientific Reports* 5 (2015) 1–11. doi:10.1038/srep15021. URL <http://dx.doi.org/10.1038/srep15021>
- [14] F. Péron, L. Rat-Fischer, M. Lalot, L. Nagle, D. Bovet, Cooperative problem solving in African grey parrots (*Psittacus erithacus*), *Animal Cognition* 14 (4) (2011) 545–553. doi:10.1007/s10071-011-0389-2.
- [15] L. Ostojić, N. S. Clayton, Behavioural coordination of dogs in a cooperative problem-solving task with a conspecific and a human partner, *Animal Cognition* 17 (2) (2014) 445–459, **Academic journal search engine**. doi:10.1007/s10071-013-0676-1.
- [16] B. Hare, A. P. Melis, V. Woods, S. Hastings, R. Wrangham, Tolerance Allows Bonobos to Outperform Chimpanzees on a Cooperative Task, *Current Biology* 17 (7) (2007) 619–623. doi:10.1016/j.cub.2007.02.040.
- [17] C. Tassin de Montaigu, K. Durdevic, D. Brucks, A. Krasheninnikova, A. von Bayern, Blue-throated macaws (*ara glaucogularis*) succeed in a cooperative task without coordinating their actions, *Ethology* 126 (2) (2020) 267–277.
- [18] K. Asakawa-Haas, M. Schiestl, T. Bugnyar, J. J. Massen, Partner choice in raven (*corvus corax*) cooperation, *PLoS ONE* 11 (6) (2016) 1–15. doi:10.1371/journal.pone.0156962.
- [19] S. Marshall-Pescini, C. Basin, F. Range, A task-experienced partner does not help dogs be as successful as wolves in a cooperative string-pulling task, *Scientific reports* 8 (1) (2018) 1–12.
- [20] M. Suchak, T. M. Eppeley, M. W. Campbell, F. B. de Waal, Ape duos and trios: spontaneous cooperation with free partner choice in chimpanzees, *PeerJ* 2 (2014) e417. doi:10.7717/peerj.417. URL <https://peerj.com/articles/417>

- 901 [21] M. Suchak, J. Watzek, L. F. Quarles, F. B. de Waal, Novice chimpanzees
902 cooperate successfully in the presence of experts, but may have limited
903 understanding of the task, *Animal Cognition* 21 (1) (2018) 87–98. doi:
904 10.1007/s10071-017-1142-2.
- 905 [22] A. D. Davies, Z. Lewis, L. R. Dougherty, A meta-analysis of factors influ-
906 encing the strength of mate-choice copying in animals, *Behavioral Ecology*
907 31 (6) (2020) 1279–1290.
- 908 [23] A. M. Seed, K. Jensen, Large-scale cooperation, *Nature* 472 (7344) (2011)
909 424–425.
- 910 [24] L. Ostojic, The evolution of social cognition: the case of animal cooperative
911 problem solving (2020).
- 912 [25] J. C. Valentine, Judging the quality of primary research, *The handbook of*
913 *research synthesis and meta-analysis* 2 (2009) 129–146.
- 914 [26] W. Wood, A. H. Eagly, Advantages of certainty and uncertainty, *The Hand-*
915 *book of Research Synthesis and Meta-Analysis* (2009) 455.
- 916 [27] M. P. Crawford, The cooperative solving of problems by young chim-
917 panzees, *Comp Psychol Monogr* 14 (1937) 1–88.
- 918 [28] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, P. Group, et al., Pre-
919 ferred reporting items for systematic reviews and meta-analyses: the prisma
920 statement, *PLoS med* 6 (7) (2009) e1000097.
- 921 [29] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein, *Introduction*
922 *to meta-analysis*, John Wiley & Sons, 2011.
- 923 [30] B. Farrar, K. Voudouris, N. Clayton, Replications, comparisons, sampling
924 and the problem of representativeness in animal behavior and cognition
925 research.
- 926 [31] N. Borrego, Socially tolerant lions (*panthera leo*) solve a novel cooperative
927 problem, *Animal Cognition* 23 (2) (2020) 327–336.

- 928 [32] N. I. for Health, C. E. G. Britain), Methods for the development of NICE
929 public health guidance, National Institute for Health and Care Excellence,
930 2012.
- 931 [33] J. Cohen, A coefficient of agreement for nominal scales, Educational and
932 psychological measurement 20 (1) (1960) 37–46.
- 933 [34] A. Del Re, M. A. Del Re, Package ‘compute. es’ (2012).
- 934 [35] E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics
935 and evolutionary analyses in r, Bioinformatics 35 (3) (2019) 526–528.
- 936 [36] K. P. Schliep, phangorn: phylogenetic analysis in r, Bioinformatics 27 (4)
937 (2011) 592–593.
- 938 [37] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman,
939 J. Ostell, E. W. Sayers, Genbank, Nucleic acids research 41 (D1) (2012)
940 D36–D42.
- 941 [38] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and
942 high throughput, Nucleic acids research 32 (5) (2004) 1792–1797.
- 943 [39] C. Heibl, Phyloch: R language tree plotting tools and interfaces to di-
944 verse phylogenetic software packages, Available online at: <http://www.christophheibl.de/Rpackages.html> (2008).
- 945
- 946 [40] O. Gascuel, Bionj: an improved version of the nj algorithm based on a
947 simple model of sequence data., Molecular biology and evolution 14 (7)
948 (1997) 685–695.
- 949 [41] K. P. Burnham, D. R. Anderson, K. P. Huyvaert, Aic model selection and
950 multimodel inference in behavioral ecology: some background, observa-
951 tions, and comparisons, Behavioral ecology and sociobiology 65 (1) (2011)
952 23–35.
- 953 [42] E.-J. Wagenmakers, S. Farrell, Aic model selection using akaike weights,
954 Psychonomic bulletin & review 11 (1) (2004) 192–196.

- 955 [43] R. Freckleton, P. Harvey, M. Pagel, Phylogenetic dependence and ecological
956 data: a test and review of evidence, *Am. Nat* 160 (2002) 716–726.
- 957 [44] M. Pagel, Inferring the historical patterns of biological evolution, *Nature*
958 401 (6756) (1999) 877–884.
- 959 [45] A. N. Iwaniuk, J. E. Nelson, Can endocranial volume be used as an estimate
960 of brain size in birds?, *Canadian Journal of Zoology* 80 (1) (2002) 16–23.
- 961 [46] E. L. MacLean, B. Hare, C. L. Nunn, E. Addessi, F. Amici, R. C. Anderson,
962 F. Aureli, J. M. Baker, A. E. Bania, A. M. Barnard, et al., The evolution
963 of self-control, *Proceedings of the National Academy of Sciences* 111 (20)
964 (2014) E2140–E2148.
- 965 [47] A. R. DeCasien, S. A. Williams, J. P. Higham, Primate brain size is pre-
966 dicted by diet but not sociality, *Nature ecology & evolution* 1 (5) (2017)
967 1–7.
- 968 [48] M. W. Colbert, R. Racicot, T. Rowe, Anatomy of the cranial endocast
969 of the bottlenose dolphin, *tursiops truncatus*, based on hrxt, *Journal of*
970 *Mammalian Evolution* 12 (1-2) (2005) 195–207.
- 971 [49] W. Viechtbauer, Conducting meta-analyses in r with the metafor package,
972 *Journal of statistical software* 36 (3) (2010) 1–48.
- 973 [50] L. R. Dougherty, L. M. Guillette, Linking personality and cognition: a
974 meta-analysis, *Philosophical Transactions of the Royal Society B: Biological*
975 *Sciences* 373 (1756) (2018) 20170282.
- 976 [51] R. G. Orwin, A fail-safe n for effect size in meta-analysis, *Journal of edu-*
977 *cational statistics* 8 (2) (1983) 157–159.
- 978 [52] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, S. Heisterkamp, B. Van Willi-
979 gen, R. Maintainer, Package ‘nlme’, Linear and nonlinear mixed effects
980 models, version 3 (1) (2017).

- 981 [53] S. Holm, A simple sequentially rejective multiple test procedure, Scandi-
982 navian journal of statistics (1979) 65–70.
- 983 [54] M. Primates, D. M. Altschul, M. J. Beran, M. Bohn, J. Call, S. DeTroy,
984 S. J. Duguid, C. L. Egelkamp, C. Fichtel, J. Fischer, et al., Establishing an
985 infrastructure for collaboration in primate cognition research, PLoS One
986 14 (10) (2019) e0223675.
- 987 [55] P. Bateson, K. N. Laland, Tinbergen’s four questions: an appreciation and
988 an update, Trends in ecology & evolution 28 (12) (2013) 712–718.
- 989 [56] P. L. Smith, D. R. Little, Small is beautiful: In defense of the small-n
990 design, Psychonomic bulletin & review 25 (6) (2018) 2083–2101.
- 991 [57] B. Farrar, L. Ostojic, The illusion of science in comparative cognition
992 (2019).
- 993 [58] J. S. Martin, S. Koski, T. Bugnyar, A. V. Jaeggi, J. J. Massen, Proso-
994 ciality, social tolerance and partner choice facilitate mutually beneficial
995 cooperation in common marmosets, callithrix jacchus, Animal Behaviour
996 173 (2021) 115–136.
- 997 [59] E. L. MacLean, L. J. Matthews, B. A. Hare, C. L. Nunn, R. C. Anderson,
998 F. Aureli, E. M. Brannon, J. Call, C. M. Drea, N. J. Emery, et al., How does
999 cognition evolve? phylogenetic comparative psychology, Animal cognition
1000 15 (2) (2012) 223–238.
- 1001 [60] S. J. Shettleworth, The evolution of comparative cognition: is the snark
1002 still a boojum?, Behavioural processes 80 (3) (2009) 210–217.
- 1003 [61] F. A. Beach, The snark was a boojum., American Psychologist 5 (4) (1950)
1004 115.
- 1005 [62] A. Krasheninnikova, P. K. Y. Chow, A. M. von Bayern, Comparative cog-
1006 nition: Practical shortcomings and some potential ways forward., Cana-
1007 dian Journal of Experimental Psychology/Revue canadienne de psychologie
1008 expérimentale 74 (3) (2020) 160.