



## 17 Abstract

18 Extracting data from studies is the norm in meta-analyses, enabling researchers to gen-  
19 erate effect sizes when raw data are otherwise not available. While there has been a  
20 general push for increased reproducibility throughout the many facets of meta-analysis,  
21 the transparency and reproducibility of the data extraction phase are still lagging be-  
22 hind. This particular meta-analytic facet is critical because it facilitates error-checking  
23 and enables users to update older meta-analyses. Unfortunately, there is little guidance  
24 of how to make the process of data extraction more transparent and shareable, in part  
25 this is as a result of relatively few data extraction tools currently offering such function-  
26 ality. Here, we suggest a simple framework that aims to help increase the reproducibility  
27 of data extraction for meta-analysis. We also provide suggestions of software that can  
28 further help users adopt open data policies. More specifically, we overview two GUI style  
29 software in the R environment, *shinyDigitise* and *juicr*, that both facilitate reproducible  
30 workflows while reducing the need for coding skills in R. Adopting the guiding principles  
31 listed here and using appropriate software will provide a more streamlined, transparent,  
32 and shareable form of data extraction for meta-analyses.

## 33 Introduction

34 In recent years, there has been a push to increase the reproducibility of meta-analyses,  
35 with the expectation that exact search strings, screening steps (e.g. the PRISMA flowchart  
36 [Moher et al., 2011](#); [O’Dea et al., 2021](#)), and metadata of accepted papers are now included  
37 alongside manuscripts. However, even as transparent reporting is improving over time,  
38 studies across a number of research fields have concluded that more still needs to be done  
39 to increase reproducibility and produce robust results ([Lakens et al., 2016, 2017](#); [Maassen  
40 et al., 2020](#); [Polanin et al., 2020](#); [Chow et al., 2021](#)). One of the key methodological steps  
41 of meta-analyses is data extraction, but, unlike the study selection process, there is still

42 no unified reporting framework. In addition, although effect sizes are typically associ-  
43 ated with a particular study when data are available, the source of the data within that  
44 study (e.g. be it from text, table, or figure) used to compute these effect sizes is seldom  
45 reported, meaning this process lacks both transparency and, importantly, reproducibil-  
46 ity. As a result, several papers have highlighted the prevalence of errors in meta-analysis,  
47 particularly surrounding the data extraction process (Götzsche *et al.*, 2007; Mathes *et al.*,  
48 2017; Wong & Bouchard, 2022).

49 Typically, a large quantity of the data needed for meta-analysis are locked away  
50 within figures. Indeed, Simmonds *et al.* (2022) recently showed that in 3%-64% of studies  
51 (depending on the field of research), had measures of uncertainty needed for meta-analysis  
52 (e.g. standard error or variance) only contained within figures. Consequently, some  
53 form of figure-based data extraction software is required (reviewed in Pick *et al.*, 2019).  
54 However, often the data extraction process is mentioned simply as a sentence within a  
55 study's methodology. Which tool was used to extract data, which effect sizes originated  
56 from figures, which particular figures the effect sizes came from, and any record of the  
57 extraction process are rarely reported. Ultimately, this severely limits the potential for  
58 error-checking and replicability.

59 Here, to assess the extent of the data extraction problem, we review the current state  
60 of the literature. Firstly, by searching recent meta-analyses in Ecology and Evolution for  
61 evidence of adequate reporting of data extraction software. Secondly, by investigating  
62 papers that cited the R package *metaDigitise* as a case study. We then introduce a simple  
63 framework to help improve the reproducibility of data extraction, and in particular data  
64 extraction from figures (these should still readily apply to data extraction from other  
65 sources, such as tables and text). Finally, we introduce two GUI-based software packages  
66 in the R environment (one of the most commonly used statistical languages), *shinyDigitise*  
67 and *juicR*, which have both been designed to aid transparency and reproducibility. The  
68 use and sharing of resulting files from these programs will facilitate error-checking and

69 updating of meta-analyses into the future.

## 70 State of the field

71 To start, we wanted to quantify what percentage of meta-analyses adequately report  
72 software packages used to extract data from figures. To do this, YY, ML, and JR re-  
73 examined the 102 meta-analyses reviewed in the 2021 PRISMA-EcoEvo guidelines (O’Dea  
74 *et al.*, 2021). From these 102 studies, only 39 cited the data extraction software that was  
75 used to extract data from figures (representing 38% of the total number).

76 Next, to assess transparency of the data extraction process itself, EIC reviewed  
77 all studies listed as citing the *metaDigitise* R package (Pick *et al.*, 2019) on Google  
78 Scholar in August 2022 (all data and code located at the Github repository, Elvimey-  
79 Cook/DataExtraction). Papers citing *metaDigitise* provide a good insight into the trans-  
80 parency of the data-extraction process. The recently published *metaDigitise* package (R  
81 package on CRAN in 2018, accompanying paper published in 2019) was in part designed  
82 to help improve transparency and reproducibility of data extraction (Pick *et al.*, 2019). It  
83 provides a simple way of storing figures and associated extraction data which can easily  
84 be uploaded as part of the data archiving process. Of these 70 published papers, 55 were  
85 meta-analyses (the other 15 papers cited were either general meta-analysis guidelines or  
86 systematic reviews). From these meta-analyses EIC recorded:

- 87 a) If the data was open and accessible (interoperable; e.g. as a .csv or .txt file).
- 88 b) Whether the paper reported the origin of the effect size (text/table/figure).
- 89 c) Whether figures that had been used in the data extraction process were stored  
90 alongside the extracted data.
- 91 d) Whether a figure-specific data extraction record was stored alongside the appropri-  
92 ate figure.

93 The results of this survey are shown in Figure 1. 78% of the 55 meta-analyses (n =

94 43) had available data in an interoperable format, despite the open access policy of many  
95 journals, and increased awareness of the importance of open-data in those conducting  
96 meta-analysis. From these, only 24 (44% of the total) readily provided information about  
97 the origin of the effect sizes. These studies showed the large dependence of data from  
98 figures, with the number of effect sizes generated from figures ranging from 1-922 (median  
99 = 83) representing between 2-96% (median = 28%) of total effect sizes generated from  
100 figures. Finally, only four (7%) provided the figures from which data was extracted  
101 and only two provided the calibration data needed to recreate the extraction (5%) in  
102 addition to the figure and metadata required to reproduce the analysis (Figure 1). The  
103 low reporting rates are even lower if one considers only 40% meta-analyses reported what  
104 software package was used to extract effect sizes from figures.

## 105 **Framework for data extraction**

106 Based on this survey it is clear that we need to improve the transparency and repro-  
107 ducibility of data extraction in meta-analysis. To achieve this, we introduce a simple  
108 five-point framework.

109 1. **Provide data.** As discussed at length elsewhere ([Miyakawa, 2020](#)), providing data  
110 is a minimum requirement for reproducibility. We found that 78% of meta-analyses  
111 provide data, similar to the 77% in a recent review of ecology and evolution meta-  
112 analyses (2010 - 2019; [O’Dea \*et al.\*, 2021](#)). Although this shows an improvement  
113 over the last decade (from 31% shared data in ecology meta-analyses between the  
114 years 1996–2013 [Koricheva & Gurevitch, 2014](#)), and is substantially greater than in  
115 other fields (e.g. 3% of studies provided interoperable data in clinical psychological  
116 meta-analyses from 2000–2020; [López-Nicolás \*et al.\*, 2022](#)), data in meta-analysis  
117 typically come from open sources (i.e. published literature) and so there does not  
118 seem many obvious reasons why data should not be made public with the paper.

119 Furthermore, those conducting meta-analyses are in a prime position to understand  
120 the importance of easily accessible data (given the shared experience and agony of  
121 extracting information from individual studies). Meta-analysts should therefore be  
122 expected to lead by example and provide their own data.

123 **2. Clearly state where each effect size was extracted from.** In addition to  
124 providing the title of the study and relevant additional information, a column in  
125 your meta-data file should clearly state whether your effect size was extracted from -  
126 text, table, figure or appendix/supplementary material, including a reference to the  
127 exact location, e.g. "Figure 2a", "table 3", "main text p275". [Curtis \*et al.\* \(2013\)](#)  
128 suggested a shorthand for reporting this information in tabular form (e.g., F2a, T3),  
129 and we extend this formatting to T=table, M=main text, F=figure, A=appendix,  
130 S=supplementary material, R=raw data, followed by the figure or page number  
131 where the data was extracted. This simple step greatly improves transparency,  
132 aids reproducibility and helps identify potential errors. Despite this, only 44% of  
133 meta-analyses in our sample gave this information, in line with the 44% in ecology  
134 and evolution meta-analyses ([O'Dea \*et al.\*, 2021](#)).

135 **3. Provide information on any transformation that may have been per-**  
136 **formed on the data.** Just providing effect sizes alone does not give information  
137 on how they have been generated. For example, transformations have to be used to  
138 generate means and standard deviations from the quantiles in a boxplot (e.g. [Wan  
139 \*et al.\*, 2014](#)). Other typical conversions include converting standard errors (SE) to  
140 standard deviations (SD), or calibrations of extracted data by back-transforming  
141 logarithms. These details are more challenging to report succinctly in tabular for-  
142 mat, as they may require equations, but a textual description is better than nothing.  
143 Providing raw data and code also greatly aid transparency in how effect sizes have  
144 been generated, and allows them to be reproduced.

145 **4. Provide the figures in an open data repository with a record of the data**

146 **extraction process.** Every figure that has undergone data extraction should be  
147 provided in a digital data repository alongside the generated effect size. However,  
148 figures alone don't provide sufficient for reproducibility and error-checking of the  
149 data-extraction process. Data extraction files including calibration data are also  
150 needed for any researcher to be able to recreate the extraction process.

151 **5. For software developers, enable the saving and reloading of the data**  
152 **extraction process.** Whilst there exists a multitude of data extraction tools, few  
153 allow users to easily save and reload the data extraction process. Therefore, to  
154 increase reproducibility, individuals looking to provide tools or software for data  
155 extraction should ensure this functionality. To further extend transparency and  
156 openness of these data, the file format of extractions should be tool agnostic with  
157 a format accessible to all (interoperable; e.g., a .csv file).

## 158 **Tools for increasing reproducibility in figure-based ex-** 159 **traction**

160 Here we highlight two R-based packages that allow for reproducible figure-based data ex-  
161 traction, *metaDigitise* (Pick *et al.*, 2019) and *juicr* (Lajeunesse, 2021). We also introduce  
162 a new GUI-based package, *shinyDigitise*. We focus on these packages because R is one of  
163 the most widely used statistical environments for analysing meta-analytic data. Making  
164 use of packages already within the R environment will unify data extraction and analysis  
165 workflows which will promote transparency and reproducibility.

166 *shinyDigitise* is a streamlined and intuitive GUI interface which makes use of functions  
167 within the *metaDigitise* package. This includes the ability to extract data from a wide  
168 variety of plot-types (Pick *et al.*, 2019), and automatically save calibration data so users  
169 have a historical record of the figure-based data extraction process. *shinyDigitise* should  
170 reduce the barrier of entry by needing very little experience of writing code and knowledge

171 of the R coding software. Users simply install the *shinyDigitise* package, which contains  
172 the functionality of *metaDigitise*, and direct *shinyDigitise* to the folder where images  
173 that need extracting are stored with a single line of R code. Using a simple shiny web  
174 application, which opens immediately, users progress through the various required steps  
175 of data extraction without having to engage with the R console directly. When the user  
176 has finished data extraction, a comma-delimited data file (.csv) containing extracted data,  
177 along with files containing a record of the data extraction process will automatically be  
178 created and ready to share. The raw extracted data can also easily be imported into  
179 R for further analysis if so required. The *shinyDigitise* enhancement to the popular  
180 *metaDigitise* package adds additional options such as image zooming, and will be more  
181 accessible to individuals with varying levels of coding experience in a clearer and cleaner  
182 data extraction environment.

183     Alongside *metaDigitise* and *shinyDigitise*, *juicr* offers savable and shareable records  
184 of retrieved data from images. Although being developed independently, this package  
185 has also converged on a similar philosophy of archiving the extraction process, but has  
186 approached the problem through the perspective of upholding the reproducibility of au-  
187 tomated approaches. As in *metaDigitise* and *shinyDigitise*, *juicR* offers a point-and-click  
188 solution to extracting data from images; however for some tasks, decision-making of what  
189 to extract can be delegated to automated (full algorithmic) or semi-automated (algorith-  
190 mic with user assistance) tools. Although seemingly practical in terms of speed and accu-  
191 racy, automated tools are neutral in what is extracted from images, and users will often  
192 need to parameterise models to target data (e.g., diamonds over circles in scatter plots),  
193 supplement missed data points, or cleanup false detections. Further, extraction success  
194 is also dependent on image identity—which is highly variable due to the diversity of ways  
195 published figures are available (e.g., high/low resolution images from publisher websites,  
196 inconsistent image sizes from user-extracted clippings or screenshots from PDFs). All of  
197 which are decisions and sources of variability that need to be transparent for reproducibil-  
198 ity and validation, and *juicr* aims to embed this more detailed information in reports that



199 are both human- and machine-readable. Moving forward, and given that technology is  
200 rapidly growing in this space, transparency and reporting of automated systems should  
201 be a core feature, like in *juicr*, in the design of model-assisted extraction tools.

202 Importantly, all these software packages provide a user with both an effect size and a  
203 record for each figure they choose to extract data from. The extraction process and data  
204 are archived and, after depositing on an appropriate data repository, can be subsequently  
205 viewed and error checked by the user or by anyone with access to both files. Whilst  
206 this is an important step for reproducibility, and directly adheres to point four on the  
207 framework above, very few people have adopted the use this archiving functionality.  
208 Figure 1 exemplifies this by highlighting the low percentage of studies that share source  
209 figures, their extracted data, and information as to when the extraction software tool  
210 was used, in addition to providing historical records of data extraction. Clearly there is  
211 an urgent need to increase transparency of data extraction, and the framework outlined  
212 above would go some way to addressing this. However, not only is this a data sharing  
213 issue, but it is also a functionality issue. As such, one of the chief priorities for future  
214 extraction software should be reproducibility, with the ability to save and reload the data  
215 generating process (i.e., the accuracy of extraction) and not just focus on the data itself.

216 Finally, it is our hope that the simple data extraction framework outlined in this  
217 article can serve as an important first step to help close the reproducibility gap that  
218 exists for this facet of meta-analysis. This general framework can be readily adopted  
219 by the many research areas that conduct meta-analysis and by doing so, will allow for  
220 greater reproducibility across disciplines.

## 221 **Acknowledgements**

222 We are grateful to Malgorzata Lagisz, Yefeng Yang and Joanna Rutkowska who surveyed  
223 the 102 meta-analyses as a part of a larger project. Note, we refer to authors in text using

224 the MeRIT system (Method Reporting with Initials for Transparency) as per (Nakagawa  
225 *et al.*, 2023, Accepted).

## 226 **Highlights**

227 In meta-analysis, there is often a large quantity of data that is contained within figures  
228 that needs some form of data extraction in order to convert it into a usable effect size.  
229 However, often the process and record of the data extraction process is limited, both  
230 in the methods section and when data is later uploaded to an open data repository.  
231 Therefore, in order to increase the reproducibility of the data extraction facet of meta-  
232 analysis, we introduce a simple five-point framework which includes suggestions for future  
233 research. Furthermore, we highlight two packages in R that readily facilitate reproducible  
234 workflows and allow for shareable records of the data extraction process. Adopting the  
235 principles and suggestions we have provided here will help towards ensuring that the  
236 entire meta-analysis process is more transparent, open, and reproducible.

## 237 **References**

- 238 Chow, J.C., Sjogren, A.L. & Zhao, H. (2021) Reporting and Reproducibility of Meta-  
239 Analysis in Speech, Language, and Hearing Research. *Journal of Speech, Language,*  
240 *and Hearing Research*, **64**, 2786–2793.
- 241 Curtis, P.S., Mengersen, K., Lajeunesse, M.J., Rothstein, H.R. & Stewart, G.B. (2013)  
242 Extraction and critical appraisal of data. J. Koricheva, J. Gurevitch & K. Mengersen,  
243 eds., *Handbook of meta-analysis in ecology and evolution*, pp. 52–60. Princeton Univer-  
244 sity Press, Princeton, New Jersey, USA.
- 245 Götzsche, P.C., Hróbjartsson, A., Marić, K. & Tendal, B. (2007) Data Extraction

- 246 Errors in Meta-analyses That Use Standardized Mean Differences. *JAMA*, **298**.  
247 <https://dx.doi.org/10.1001/jama.298.4.430>.
- 248 Koricheva, J. & Gurevitch, J. (2014) Uses and misuses of meta-analysis in plant ecology.  
249 *Journal of Ecology*, **102**, 828–844. ISBN: 1365-2745, <https://dx.doi.org/10.1111/1365->  
250 [2745.12224](https://dx.doi.org/10.1111/1365-2745.12224).
- 251 Lajeunesse, M.J. (2021) Squeezing data from scientific images using the juicr package for  
252 R. R package version 0.1.
- 253 Lakens, D., Page-Gould, E., Assen, M.A.L.M.v., Spellman, B., Schönbrodt, F., Hassel-  
254 man, F., Corker, K.S., Grange, J.A., Sharples, A., Cavender, C., Augusteijn, H.E.M.,  
255 Augusteijn, H., Gerger, H., Locher, C., Miller, I.D., Anvari, F. & Scheel, A.M. (2017)  
256 Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report.
- 257 Lakens, D., Hilgard, J. & Staaks, J. (2016) On the reproducibility of  
258 meta-analyses: six practical recommendations. *BMC Psychology*, **4**, 24.  
259 <https://dx.doi.org/10.1186/s40359-016-0126-3>.
- 260 López-Nicolás, R., López-López, J.A., Rubio-Aparicio, M. & Sánchez-Meca, J. (2022)  
261 A meta-review of transparency and reproducibility-related reporting practices in pub-  
262 lished meta-analyses on clinical psychological interventions (2000–2020). *Behavior Re-*  
263 *search Methods*, **54**, 334–349. <https://dx.doi.org/10.3758/s13428-021-01644-z>.
- 264 Maassen, E., Assen, M.A.L.M.v., Nuijten, M.B., Olsson-Collentine, A. & Wicherts, J.M.  
265 (2020) Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS*  
266 *ONE*, **15**, e0233107. <https://dx.doi.org/10.1371/journal.pone.0233107>.
- 267 Mathes, T., Klaffen, P. & Pieper, D. (2017) Frequency of data extraction errors and  
268 methods to increase data extraction quality: a methodological review. *BMC Medical*  
269 *Research Methodology*, **17**, 152. <https://dx.doi.org/10.1186/s12874-017-0431-4>.

270 Miyakawa, T. (2020) No raw data, no science: another possible source of the reproducibil-  
271 ity crisis. *Molecular Brain*, **13**, 24. <https://dx.doi.org/10.1186/s13041-020-0552-2>.

272 Moher, D., Altman, D.G., Liberati, A. & Tetzlaff, J. (2011) PRISMA statement. *Epi-*  
273 *demiology*, **22**, 128.

274 Nakagawa, S., Ivimey-Cook, E.R., Grainger, M.J., O’Dea, R.E., Burke, S., Drobniak,  
275 S.M., Gould, E., Macartney, E.L., Martnig, A.R., Paquet, M., Morrison, K., Pick,  
276 J.L., Pottier, P., Ricolfi, L., Wilkinson, D.P., Willcox, A., Williams, C., Wilson, L.A.B.,  
277 Windecker, S.M., Yang, Y. & Lagisz., M. (2023, Accepted) Method reporting with ini-  
278 tials for transparency (merit) promotes more granularity and accountability for author  
279 contributions. *Nature Communications*.

280 O’Dea, R.E., Lagisz, M., Jennions, M.D., Koricheva, J., Noble, D.W., Parker, T.H.,  
281 Gurevitch, J., Page, M.J., Stewart, G., Moher, D. & Nakagawa, S. (2021) Pre-  
282 ferred reporting items for systematic reviews and meta-analyses in ecology and  
283 evolutionary biology: a PRISMA extension. *Biological Reviews*, **96**, 1695–1722.  
284 <https://dx.doi.org/10.1111/brv.12721>.

285 Pick, J.L., Nakagawa, S. & Noble, D.W.A. (2019) Reproducible, flexible and high-  
286 throughput data extraction from primary literature: The metaDigitise r package.  
287 *Methods in Ecology and Evolution*, **10**, 426–431. [https://dx.doi.org/10.1111/2041-](https://dx.doi.org/10.1111/2041-210X.13118)  
288 [210X.13118](https://dx.doi.org/10.1111/2041-210X.13118).

289 Polanin, J.R., Hennessy, E.A. & Tsuji, S. (2020) Transparency and Reproducibility of  
290 Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*,  
291 **15**, 1026–1041. <https://dx.doi.org/10.1177/1745691620906416>.

292 Simmonds, E.G., Adjei, K.P., Andersen, C.W., Aspheim, J.C.H., Battistin, C., Bulso,  
293 N., Christensen, H., Cretois, B., Cubero, R., Davidovich, I.A., Dickel, L., Dunn, B.,  
294 Dunn-Sigouin, E., Dyrstad, K., Einum, S., Giglio, D., Gjerlow, H., Godefroidt, A.,  
295 Gonzalez-Gil, R., Cogno, S.G., Grosse, F., Halloran, P., Jensen, M.F., Kennedy, J.J.,

- 296 Langsaether, P.E., Laverick, J.H., Lederberger, D., Li, C., Mandeville, E., Mandeville,  
297 C., Moe, E., Schroder, T.N., Nunan, D., Parada, J.S., Simpson, M.R., Skarstein, E.S.,  
298 Spensberger, C., Stevens, R., Subramanian, A., Svendsen, L., Theisen, O.M., Watret,  
299 C. & OHara, R.B. (2022) How is model-related uncertainty quantified and reported in  
300 different disciplines? *arXiv preprint*. <https://dx.doi.org/10.48550/arXiv.2206.12179>.
- 301 Wan, X., Wang, W., Liu, J. & Tong, T. (2014) Estimating the sample mean and standard  
302 deviation from the sample size, median, range and/or interquartile range. *BMC medical*  
303 *research methodology*, **14**, 135. ISBN: 1471-2288 (Electronic)\r1471-2288 (Linking),  
304 <https://dx.doi.org/10.1186/1471-2288-14-135>.
- 305 Wong, J.S. & Bouchard, J. (2022) Do Meta-Analyses of Intervention/Prevention  
306 Programs in the Field of Criminology Meet the Tests of Transparency and  
307 Reproducibility? *Trauma, Violence, & Abuse*, p. 15248380211073839.  
308 <https://dx.doi.org/10.1177/15248380211073839>.

