A peer-reviewed version of this preprint was published in PeerJ on 4 August 2015.

View the peer-reviewed version (peerj.com/articles/1137), which is the preferred citable publication unless you specifically need to cite this preprint.

Andrew RL, Albert AYK, Renaut S, Rennison DJ, Bock DG, Vines T. 2015. Assessing the reproducibility of discriminant function analyses. PeerJ 3:e1137 https://doi.org/10.7717/peerj.1137

Assessing the reproducibility of discriminant function analyses

Rose L Andrew, Arianne YK Albert, Sebastien Renaut, Diana J Rennison, Dan G Bock, Tim Vines

Data are the foundation of empirical research, yet all too often the datasets underlying published papers are unavailable, incorrect, or poorly curated. This is a serious issue, because future researchers are then unable to validate published results or reuse data to explore new ideas and hypotheses. While data files may be securely stored and accessible, they must also be accompanied by accurate labels and identifiers. To assess how often problems with metadata or data curation affect the reproducibility of published results, we attempted to reproduce Discriminant Function Analyses (DFAs) from the field of organismal biology. DFA is a commonly used statistical analysis that has changed little since its inception almost eight decades ago, and therefore provides an excellent case study to test reproducibility. Out of 100 papers we initially surveyed, fourteen were excluded because they did not present the common types of quantitative result from their DFA, used complex and unique data transformations, or gave insufficient details of their DFA. Of the remaining 86 datasets, there were 16 cases for which we were unable to confidently relate the dataset we received to the one used in the published analysis. The reasons ranged from incomprehensible or absent variable labels, the DFA being performed on an unspecified subset of the data, or incomplete data sets. We focused on reproducing three common summary statistics from DFAs: the percent variance explained, the percentage correctly assigned and the largest discriminant function coefficient. The reproducibility of the first two was high (20 of 25 and 43 of 59 datasets, respectively), whereas our success rate with the discriminant function coefficients was lower (15 of 36 datasets). When considering all three summary statistics, we were able to completely reproduce 46 (66%) of 70 datasets. While our results are encouraging, they highlight the fact that science still has some way to go before we have the carefully curated and reproducible research that the public expects.

Title: Assessing the reproducibility of discriminant function analyses

Authors: Rose L. Andrew [1,2,*], Arianne Y.K. Albert [3], Sebastien Renaut [1,4], Diana J. Rennison [1], Dan G. Bock [1], Timothy H. Vines [1,5,]

Affiliations: [1]Biodiversity Research Centre, University of British Columbia, 6270 University Blvd Vancouver BC, Canada, V6T 1Z4.

[2]School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia.

[3]Women's Health Research Institute, 4500 Oak Street, Vancouver, BC, Canada V6H 3N1.

[4]Institut de recherche en biologie végétale, Département de sciences biologiques, Université de Montréal 4101 Sherbrooke est, Montréal, QC, Canada

[5]Molecular Ecology Editorial Office, 6270 University Blvd Vancouver BC, Canada, V6T 1Z4.

[*]Author for correspondence: Tim Vines, vines@zoology.ubc.ca

Abstract

Data are the foundation of empirical research, yet all too often the datasets underlying published papers are unavailable, incorrect, or poorly curated. This is a serious issue, because future researchers are then unable to validate published results or reuse data to explore new ideas and hypotheses. While data files may be securely stored and accessible, they must also be accompanied by accurate labels and identifiers. To assess how often problems with metadata or data curation affect the reproducibility of published results, we attempted to reproduce Discriminant Function Analyses (DFAs) from the field of organismal biology. DFA is a commonly used statistical analysis that has changed little since its inception almost eight decades ago, and therefore provides an excellent case study to test reproducibility. Out of 100 papers we initially surveyed, fourteen were excluded because they did not present the common types of quantitative result from their DFA, used complex and unique data transformations, or gave insufficient details of their DFA. Of the remaining 86 datasets, there were 16 cases for which we were unable to confidently relate the dataset we received to the one used in the published analysis. The reasons ranged from incomprehensible or absent variable labels, the DFA being performed on an unspecified subset of the data, or incomplete data sets. We focused on reproducing three common summary statistics from DFAs: the percent variance explained, the percentage correctly assigned and the largest discriminant function coefficient. The reproducibility of the first two was high (20 of 25 and 43 of 59 datasets, respectively), whereas our success rate with the discriminant function coefficients was lower (15 of 36 datasets). When considering all three summary statistics, we were able to completely reproduce 46 (66%) of 70 datasets. While our results are encouraging, they highlight the fact that science still has some way to go before we have the carefully curated and reproducible research that the public expects.

Introduction

Published literature is the foundation for future research, so it is important that the results reported in scientific papers be supported by the accompanying data. After all, we cannot easily predict which aspects of a paper will prove useful in the future (Wolkovich et al. 2012), and if a portion of the results are wrong or misleading then subsequent research effort may well be wasted (e.g. Begley & Ellis 2012). One relatively simple way to judge the validity of published research is to obtain the original data analyzed in the paper and attempt to repeat some or all of the analyses: this allows researchers to retrace the path the authors took between the raw data and their results. The idea of reproducibility in research is becoming a topic of great interest and this movement is gaining traction with journals (Announcement: Reducing our irreproducibility 2013; McNutt 2014). Correspondingly, there is clearly a need to quantify the validity of published research, yet there have been only a modest number of published studies that have tried to reproduce the results of published papers (e.g. Errington et al. 2014; Gilbert et al. 2012; Ioannidis et al. 2009), most likely because it is often difficult to access the underlying data (Drew et al. 2013; Savage & Vickers 2009; Vines et al. 2013; Wicherts et al. 2006).

Even when the data file is available, one common problem that hampers reanalysis is poor data curation: it is sometimes difficult to relate the dataset provided by the authors upon request or archived at publication to the one described in the paper (Gilbert et al. 2012; Ioannidis et al. 2009; Michener et al. 1997). For example, variable names may differ between the received dataset and the one described in the study, or there may be differences in the number of variables or data points. It is typically not possible to reproduce the authors' analyses in these cases, and moreover the data may not be considered sufficiently reliable for testing new hypotheses.

The current study had two goals: to assess how often we could reproduce the authors' results when the available dataset matched the one described in the paper, and to assess how often poor data curation prevents re-analysis of published data. We made use of 100 datasets acquired from authors as part of an earlier study assessing the impact of time since publication on data availability (Vines et al. 2014). The articles we chose had to i) contain morphometric data from plants or animals, ii) have analysed the morphometric data with a DFA, and iii) have not previously made the data available online. To make the data set

manageable in size, we selected only those studies published in odd years (between 1991 and 2011) as detailed in Vines et al. (2014).

We focused on morphometric data because it has been collected in a similar fashion for decades (e.g. with Vernier callipers or a binocular microscope), so datasets from a range of time periods are expected to be similar in size and format. Similarly, since its inception (Fisher 1936), DFA has frequently been applied to morphometric datasets. While computer processing power has greatly improved over the years, the way the analysis has been performed has changed little. We can therefore reasonably compare DFAs from papers with a wide range of publication dates, allowing us to investigate how changing analysis software or changing curation standards affect reproducibility. In combination with Vines et al. (2014), our results quantify the extent of the challenges facing science publication, both in terms of getting hold of the original data analysed in the paper, and in terms of the proportion of analyses that are poorly curated or cannot be reproduced.

Materials and methods

As part of the Vines et al. (2014) study, we received 100 datasets from authors (see Table 2). For papers reporting a classical DFA of morphological data, linear or quadratic DFA were considered, as were stepwise analyses where a) the variables in the final model were defined and b) at least one of three common metrics metrics (see below) was presented. This allowed us to attempt to reproduce the final model in the same way as a simple linear DFA. Studies employing stepwise analysis of relative warps or Fourier-transformed data were also excluded at this point, as these studies unfortunately did not indicate which variables were included in the final model. A study entirely written in a foreign language (Spanish) was also excluded.

For each remaining study, we followed the protocol below.

1) We first assessed the description of the methodology, checking whether the paper adequately described the groupings and morphometric variables used in the analysis.

2) We examined the data files (in some cases multiple files were supplied), which sometimes required specialised file formats to be converted. This was carried out using the R packages ‘foreign’ (R Core Team 2013) and ‘RODBC’ (Ripley & Lapsley

2013). If the data file was clearly wrong (e.g. a summary table, instead of raw data) we assigned the paper as 'Incorrect data file'.

3) We assessed whether the metadata contained in the data file, in other files supplied by the author or in the accompanying email were complete and could be related to their description in the paper. We classified papers missing sample names and those with unclear population groupings as having 'Insufficient metadata'. This category also included papers for which variable labels were in a foreign language and could be not be matched to the variables reported in the paper. However, we accepted files with unlabeled data columns, but where the number of columns matched number of variables described in the paper.

4) We then identified discrepancies in sample sizes or number of variables, after deleting rows containing missing data or samples not included in the analysis, where appropriate. We assigned papers for which variables were missing or for which sample sizes did not match those reported in the paper as 'Data discrepancy'.

5) In addition to simple transformations (logarithm or square root), we conducted size adjustments based on multigroup principal components analysis (e.g. Burnaby's (1966) back-projection) using the *R* packages 'multigroup' (Eslami et al. 2014) and 'cpcbp' (Bolker & Phillips 2012).

6) When more than one DFA meeting our criteria was conducted in a paper, we selected only the first one. We recorded whether raw or standardised coefficients were presented, whether cross-validation was used in the classification of individuals, and the statistical software used. The year of publication was recorded for each paper.

Based on a preliminary survey of the papers, we identified three DFA metrics to reproduce: the percentage of variance explained (PVE), the percentage of samples assigned correctly (PAC), and the largest model coefficient. These three summary statistics are commonly reported for DFAs, and are useful for interpreting DFA in a meaningful manner (Reyment et al. 1984), although the detail in which DFAs are described varies greatly depending on the focus of the paper. PVE and PAC are complementary indicators of the discriminatory power of a discriminant function, whereas the function coefficients provide a formula for assigning new samples to one group or another.

Our reanalysis procedure was designed to produce a single value per paper for each metric. Where possible, we compared the PVE for the first axis, which explains the greatest amount of variance in the model. When PVE was reported as the sum of the first two or three axes, we compared the summed PVE. We calculated PAC overall, or to a particular group if the overall percentage was not reported in the paper. For the coefficient, we selected the variable that had the coefficient with the largest absolute value, and determined from the paper whether the raw or standardised coefficient was used.

Although the original analyses used diverse statistical packages, we performed all discriminant function reanalyses in the statistical software R v3.1.0 (R Core Development Team 2011), using the functions 'lda' (in the MASS package; Venables & Ripley 2002) with default parameters. For a small subset of the data sets (three studies), we also conducted the analyses in SPSS using the same options as in the R analysis to check for systematic differences. We also estimated each summary statistic using proportional or flat priors and used the value that was closest to the published value. For PAC, authors reported a variety of methods for assignment, ranging from standard classification functions based on all data, to omitting one quarter of the data as a validation set. In our reanalysis, classification was carried out using leave-one-out (jackknife) cross-validation or direct prediction in 'lda', based on the description of the analysis in the paper. When neither was stated, we performed both and selected the value that was closest to the published result. While this approach biases the results towards the published value, it is a conservative means to avoid unfair treatment of studies that used default parameters for their chosen software.

The R code used is provided in the Supplementary Materials. We considered the analysis to have been reproduced if the PVE, coefficient, or PAC 'matched' within 1% of the published value, or was 'close' if within 5% of the published value.

We used generalised linear models (the core 'glm' function in R) in order to assess whether publication year affected the likelihood of problems in the data sets that would prevent attempts to reproduce the DFA results. Given a binomial model, we tested the effect of publication year on the probability that metadata would be insufficient, or that there would be discrepancies in sample sizes or variable numbers. A Fisher's exact test was used to test the effect of statistical software on data problems and on the success of the reanalysis, combining software used in only a single study (S-Plus, STATGRAPHICS and LINDA) into one category ("other").

Although we contacted authors again to ask for their preferences regarding acknowledgment or anonymity (Table 2), we did not seek further information (e.g. metadata or analysis parameters) to inform our reanalysis.

Results

The current study used 100 data sets originally gathered by Vines et al. (2014). Fourteen of those data sets were excluded from our reanalysis attempt (Tables 1 & 2): one paper was entirely in a language other than English (Spanish); two did not perform classical DFA; two used non-morphological data in their DFA; six did not present any of the metrics that we were attempting to reproduce; and three were based on stepwise analysis for which the final set of Fourier-transformed variables or relative warps were not specified.

Of the 86 remaining studies, the data files provided for two (2.3%) were classified as 'Incorrect data file': summary tables instead of morphometric data, or the data set used for a different analysis from the same paper (Table 1). Seven others (8.1%) were assigned as 'Insufficient metadata', such that columns in the data files could not be matched to the variables described in the paper. This was due to a combination of abbreviations and the use of languages other than English. A further five data sets (5.8%) did not match the expected sample sizes, and two (2.3%) were missing variables. All seven were classified as 'Data discrepancy'.

We found no effect of publication year on the probability of having inadequate metadata (odds ratio 0.95, $P = 0.44$) and no effect on the probability of mismatched sample size or missing variables ('Data discrepancy': odds ratio 1.05, $P = 0.55$). Combining these main types of data problems preventing us from attempting reanalysis (incorrect data, insufficient metadata, missing variables or mismatched sample sizes), there was no effect of year (odds ratio 0.99, $P = 0.87$). Where stated, the type of software (SAS ® (SAS Institute, Cary, NC, USA), SYSTAT (SYSTAT Software Inc., Richmond, CA, USA), SPSS (SPSS Inc., Chicago, IL, USA), MATLAB (Mathworks, Natick, MA, USA), STATISTICA (Statsoft, Tulsa, OK, USA), JMP (SAS Institute, Cary, NC, USA), R (R Core Development Team 2011), S-Plus ® (TIBCO Software Inc., Palo Alto, CA, USA), STATGRAPHICS (StatPoint Inc., Rockville, MD, USA) and LINDA (Cavalcanti 1999)) used for the initial study had a significant effect on the probability of data problems (Fisher's exact test, $P = 0.012$). This was largely due to a

high likelihood of data problems (0.454) among data sets originally analysed with SAS, compared with 0.186 overall.

We attempted a reanalysis of the DFA for the remaining 70 studies, and the results are summarised in Table 2. Our results regarding the PVE were generally close to the published values (Pearson correlation coefficient, $r = 0.94$, $P < 0.0001$; Figure 2). Of the 25 reanalysed data sets reporting this statistic, our reproduced value was within 1% of the published value in 20 (80%) of cases, and within 5% of the published value in 23 (92%) of cases. The PAC statistic was also often reproduced (Pearson's $r = 0.95$, $P < 0.0001$; Figure 4). Of 59 analyses attempted, reanalysed values differed from the published value by 1% or less in 43 (73%) cases, while 55 (93%) were within 5%. Discriminant function coefficients were reproduced less frequently in the reanalysis. Using the absolute value of the coefficient to exclude sign differences, reproduced values were within 5% of the published value for 15 (65%) of the 26 data sets reanalysed for this statistic, and each of these values was also within 1%. There was still a strong correlation between the published value and our estimate (using absolute values, Pearson's $r = 0.96$, $P < 0.0001$; Figure 3).

Of all 110 reanalysed PVE, PAC and coefficient values, 78 (71%) were within 1% of the published value, and 93 (85%) were within 5% (Table 2). Considering the reported summary statistics together for each paper, our reanalysis failed to replicate any value in the paper at the most stringent level (within 1%) in 12 studies (17% of the total 70 data sets; Table 1); however, we were able to partially reproduce 12 (17%) studies and completely reproduce the results in 46 studies (66%). The reanalysed values were within 5% of the published value for all three statistics for 55 (79%) of studies.

There was no effect of publication year on discrepancies between the published and reproduced values for PVE, coefficients or PAC (test, $P > 0.2$ in each case). Sample sizes were sufficient for a reliable test of the software effect for PAC only and this effect was not significant (Fisher's exact test, $P = 0.67$). There was also no effect of software on the overall reanalysis success (Fisher's exact test, $P = 0.85$) and the results of analysis with SPSS matched those of analysis with R entirely.

Discussion

Confidence in scientific research is boosted when published results can be independently reproduced by other scientists (Price 2011). Assuming that the raw data can be obtained

(which is typically difficult, e.g. Vines et al. 2014; Vines et al. 2013; Wicherts et al. 2011; Wicherts et al. 2006), several obstacles still remain. First, poor data curation (e.g. unintelligible column headings or missing samples) or inadequate methods description can mean that the dataset obtained cannot be matched to the one described in the paper, preventing reanalysis at the outset. Second, even when the datasets do match, some aspects of the results may be inherently harder to reproduce than others, perhaps because there are multiple calculation methods for the same summary statistic, or because the calculation involves 'random walk' estimation(e.g. Gilbert et al. 2012).

In this paper we attempted to reproduce the results of DFAs for 100 datasets of papers published between 1991 and 2011. In contrast to the striking decline in data availability over time (Vines et al. 2014), we found no evidence that the reproducibility of DFAs decreased with time since publication. Encouragingly, there was also no relationship between publication year and the proportion of datasets with data problems that prevented reanalysis, or with the proportion of reproducible results.

We attempted re-analyses for 81% (70 of 86 papers) of data sets after rejecting those with obvious problems in the data file. These problems included the wrong data file being provided, missing data (individuals or variables), differences in the labels of variables between data files and published work, or unspecified subsetting of the data files prior to the analytical steps. While some of these problems could be solved through further communication with the authors, our study reflected the long-term reusability of the data, as contact with authors is likely to become increasingly difficult as time passes (Vines et al. 2014). Digital information is rapidly moving towards a more centralised online system ("the cloud", Armbrust et al. (2010)). Similarly, the responsibility for data preservation is being lifted from scientists to online repositories (e.g.: Dryad (www.datadryad.org), figshare (www.figshare.com), NCBI (www.ncbi.nlm.nih.gov)). Given this paradigm shift, we recommend more attention given to the quality of the metadata and curation of the specific files that are stored (Michener et al. 1997). For instance, if data are size-adjusted or manipulated in other ways, both pre- and post-transformation data should be archived. Perhaps the most critical piece of information is the link between column labels in the data file and the variables described in the paper. We were unable to determine the correct columns or rows in 8% of data sets. While we were able to convert all data files to text format, the loss of metadata may stem from this conversion (in one case, this had to be typed by hand, because data file provided was from a scanned hardcopy of the data in a MSc thesis

appendix). In line with previous authors on this topic (Borer et al. 2009; Whitlock 2011), we recommend storing data in text-based data formats, as these are most accessible across the range of statistical software packages. Also in line with previous recommendations (Wolkovich et al. 2012), we recommend publishing the code used in analysis (as part of the supplementary material or online repository such as GitHub, see Ram 2013), as it is often difficult to provide a full description of the parameters used for a given analysis in the methods section of a journal article.

Among the 70 data sets that were suitable to be reanalysed, we were able to reproduce, to within 1% of the published value, at least one of the three statistics that we focused on (PVE, PAC and the largest (absolute) coefficient) for 58 studies (83%). There were strong positive correlations between published and reanalysed values for statistics reported in DFA, which suggests that replication, in the broad sense, is possible when the proper metadata are provided and with adequate curation of the data file; however, the reanalysed metrics matched the published values precisely for only 46 of 70 studies (66%). Slight discrepancies could be due to differences in rounding, as well as data handling by statistical programs. The default parameters differ between *R* and SPSS, for example, although for three papers that we compare using SPSS and R, results were entirely consistent when the parameters were identical. Although obvious data file problems appear to be associated with different analysis software, there was no effect of software on the reproduction of the published results in our reanalyses.

Evaluating whether the DFA metrics analysed here fall within 5% of the published values is, in our view, a reasonable test of of reproducibility. However, it is uncertain how much the original conclusions from these studies would change based on the values we have obtained. The reproducibility of inference is an aspect of reproducibility that we admittedly did not explicitly address in this study. Additionally, while DFA was not always a central or essential component of the original study, its reproducibility is an important indicator of the underlying data's quality and/or completeness. Such checks are an essential consideration when archived data are re-used for new purposes.

The reproducibility of the analysis varied dramatically among statistics, ranging from 65% for the coefficient to 80% for PVE of reanalysed data sets, with a similar reproducibility percentage (73%) for the more complex PAC analyses. With a wider criterion for success (i.e. within 5% of the published value), 65% to 93% of reanalyses gave broadly similar

results. The discriminant function coefficients were far less likely to be reproduced, even when PVE and/or PAC matched. However, the procedures used to standardise model coefficients and calculate PAC differ among statistical packages and studies, potentially yielding overly optimistic results (see Dechaume-Moncharmont et al. 2011). This influences our ability to reproduce the results. For instance, if jackknifing had been used for all PAC reanalyses, only 56% of published values would have been reproduced (results not shown). These results suggest that while general patterns in multivariate data are likely to be robust, predictive models built on these data may be more sensitive to rounding and other minor errors in the archived data. While this clearly does not invalidate the original results, it does highlight another obstacle to successfully reproducing the authors' results: some summary statistics may be inherently harder to reproduce, particularly when there are numerous calculation methods, as is the case here, or when the estimation procedure makes use of stochastic numerical optimisation methods (e.g. Gilbert et al. 2012).

In comparison with our previous study of reproducibility of analysis using STRUCTURE (Pritchard et al. 2000), both the proportion of inadequate data or metadata and the reproducibility of basic results were similar for DFA reanalysis. However, the correlation between published and reanalyzed results was consistently greater for DFA (r = 0.94-0.96) than for STRUCTURE (r = 0.59). DFA is a much simpler statistical procedure, although other differences also exist; for instance, the STRUCTURE data sets were all in the same format. In attempts to reanalyse microarray data sets, which are much more complex than morphological data sets, approximately half of the results could be reproduced from available data (Ioannidis et al. 2009). It is not surprising that analyses with more steps and parameter choices are harder to reproduce, and this is echoed within our study, where we had to explore a wide range of analysis options to obtain close matches for the most complex DFA statistic, PAC.

Shared data is an important substrate for science and is one of the levers that may be used to improve the reliability of research (Ioannidis 2014). The system of having data re-users directly contact data generators to obtain access to their data has been in place for decades, and is absolutely necessary for data re-use within embargo periods (Roche et al. 2014), but it is not a long-term solution for the preservation of research data (Vines et al. 2014). We argue that in order for archived data to retain their full value, all of the necessary data and metadata must be stored at the time of archiving, which typically happens at or soon before/after publication. We have determined some of the common problems that can occur in self-

archived data even when authors can be contacted and are able to share their data. The same factors are relevant to communal data archives. While sequence repositories such as NCBI Genbank have made the provision of metadata a key part of the submission, the decision of what additional information to archive lies with the author for more generalised databases such as Dryad and Nature's Scientific Data. The results presented here and those of previous studies (Drew et al. 2013; Gilbert et al. 2012; Savage & Vickers 2009; Vines et al. 2014; Vines et al. 2013) illustrate the need for our research community to make data availability and curation a central part of the research and publication process.

Acknowledgments

We are extremely grateful to the authors who kindly provided their data, without which this research would not have been possible. We also thank our collaborators on the first part of this project, Florence Débarre, Michelle Franklin, Kim Gilbert and Jean-Sébastien Moore. We thank Michael Whitlock and Heather Piwowar for useful discussions during the planning of the project and Mary O'Connor for thoughtful comments on our manuscript.

References

Amado S, Armada-da-Silva PA, João F, Maurício AC, Luís AL, Simões MJ, and Veloso AP. 2011. The sensitivity of two-dimensional hindlimb joint kinematics analysis in assessing functional recovery in rats after sciatic nerve crush. *Behavioural Brain Research* 225:562-573.

Amini F, Zamini A, and Ahmadi M. 2007. Intergeneric hybridization between Kutum, *Rutilus frisii kutum*, and Bream, *Abramis brama orientalis*, of the Caspian Sea. *Journal of the World Aquaculture Society* 38:497-505.

Announcement: Reducing our irreproducibility. 2013. Nature. p 398.

Aparicio E, García-Berthou E, Araguas R, Martínez P, and García-Marín J. 2005. Body pigmentation pattern to assess introgression by hatchery stocks in native *Salmo trutta* from Mediterranean streams. *Journal of Fish Biology* 67:931-949.

Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, and Zaharia M. 2010. A view of cloud computing. *Communications of the ACM* 53:50-58.

Asanidze Z, Akhalkatsi M, and Gvritishvili M. 2011. Comparative morphometric study and relationships between the Caucasian species of wild pear (*Pyrus* spp.) and local

cultivars in Georgia. *Flora-Morphology, Distribution, Functional Ecology of Plants* 206:974-986.

Audisio P, Belfiore C, De Biase A, and Antonini G. 2001. Identification of *Meligethes matronalis* and *M. subaeneus* based on morphometric and ecological characters (Coleoptera: Nitidulidae). *European Journal of Entomology* 98:87-98.

Begley CG, and Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* 483:531-533.

Berzins LL, Gilchrist HG, and Burness G. 2009. No assortative mating based on size in black guillemots breeding in the Canadian Arctic. *Waterbirds* 32:459-463.

Bolker B, and Phillips PC. 2012. cpcbp: common principal components/back-projection analysis. R package version 0.3.3.

Borer ET, Seabloom EW, Jones MB, and Schildhauer M. 2009. Some simple guidelines for effective data management. *Bulletin of the Ecological Society of America* 90:205-214.

Bourgeois K, Curé C, Legrand J, Gómez-Díaz E, Vidal E, Aubin T, and Mathevon N. 2007. Morphological versus acoustic analysis: what is the most efficient method for sexing yelkouan shearwaters *Puffinus yelkouan*? *Journal of Ornithology* 148:261-269.

Brysting A, Elven R, and Nordal I. 1997. The hypothesis of hybrid origin of *Poa jemtlandica* supported by morphometric and isoenzyme data. *Nordic Journal of Botany* 17:199-214.

Buczkó K, Wojtal AZ, and Jahn R. 2009. *Kobayasialla* species of the Carpathian region: morphology, taxonomy and description of *K. tintinnus* spec. nov. *Diatom Research* 24:1-21.

Bulgarella M, Wilson RE, Kopuchian C, Valqui TH, and McCracken KG. 2007. Elevational variation in body size of crested ducks (*Lophonetta specularioides*) from the central high Andes, Mendoza, and Patagonia. *Ornitologia Neotropical* 18:587-602.

Burnaby TP. 1966. Growth-invariant discriminant functions and generalized distances. *Biometrics* 22:96-110.

Cadrin SX. 1995. Discrimination of American lobster (*Homarus americanus*) stocks off southern New England on the basis of secondary sex character allometry. *Canadian Journal of Fisheries and Aquatic Sciences* 52:2712-2723.

Capoccioni F, Costa C, Aguzzi J, Menesatti P, Lombarte A, and Ciccotti E. 2011. Ontogenetic and environmental effects on otolith shape variability in three Mediterranean European eel (*Anguilla anguilla*, L.) local stocks. *Journal of Experimental Marine Biology and Ecology* 397:1-7.

390 Cavalcanti MJ. 1999. LINDA—linear discriminant analysis and comparison of multivariate
391 samples with randomisation tests. http://life.biosunysb.edu/morph/
392 Conde-Padín P, Grahame J, and Rolán-Alvarez E. 2007. Detecting shape differences in
393 species of the *Littorina saxatilis* complex by morphometric analysis. *Journal of*
394 *Molluscan Studies* 73:147-154.
395 Contrafatto G. 2005. Species with fuzzy borders: the taxonomic status and species limits of
396 Saunders' vlei rat, *Otomys saundersiae* Roberts, 1929 (Rodentia, Muridae, Otomyini).
397 *Mammalia* 69:297-322.
398 Darbyshire S, and Cayouette J. 1995. Identification of the species in the *Panicum capillare*
399 complex (Poaceae) from eastern Canada and adjacent New York State. *Canadian*
400 *Journal of Botany* 73:333-348.
401 de la Hera I, Pérez-Tris J, and Telleria JL. 2007. Testing the validity of discriminant function
402 analyses based on bird morphology: the case of migratory and sedentary blackcaps
403 *Sylvia atricapilla* wintering in southern Iberia. *Ardeola* 54:81-91.
404 Dechaume-Moncharmont F-X, Monceau K, and Cezilly F. 2011. Sexing birds using
405 discriminant function analysis: a critical appraisal. *The Auk* 128:78-86.
406 Drew BT, Gazis R, Cabezas P, Swithers KS, Deng J, Rodriguez R, Katz LA, Crandall KA,
407 Hibbett DS, and Soltis DE. 2013. Lost Branches on the Tree of Life. *PLoS Biol*
408 11:e1001636.
409 Ekrt L, Travnicek P, Jarolimova V, Vit P, and Urfus T. 2009. Genome size and morphology
410 of the *Dryopteris affinis* group in Central Europe. *Preslia* 81:261-280.
411 Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, and Nosek BA. 2014. *An open*
412 *investigation of the reproducibility of cancer biology research*.
413 Eslami A, Qannari E, Bougeard S, and Sanchez G. 2014. multigroup: methods for multigroup
414 data analysis. R package version 0.4.2. .
415 Fernández IÁ, and Feliner GN. 2001. A multivariate approach to assess the taxonomic utility
416 of morphometric characters in *Doronicum* (Asteraceae, Senecioneae). *Folia*
417 *Geobotanica* 36:423-444.
418 Fisher RA. 1936. The use of multiple measurements in taxonomic problems. *Annals of*
419 *Eugenics* 7:179-188.
420 Floate KD, and Whitham TG. 1995. Insects as traits in plant systematics: their use in
421 discriminating between hybrid cottonwoods. *Canadian Journal of Botany* 73:1-13.

Foggi B, Rossi G, and Signorini M. 1999. The *Festuca violacea* aggregate (Poaceae) in the Alps and Apennines (central southern Europe). *Canadian Journal of Botany* 77:989-1013.

Forster MA, Ladd B, and Bonser SP. 2010. Optimal allocation of resources in response to shading and neighbours in the heteroblastic species, *Acacia implexa*. *Annals of Botany* 107:219-228.

Gabrielson PW, Miller KA, and Martone PT. 2011. Morphometric and molecular analyses confirm two distinct species of *Calliarthron* (Corallinales, Rhodophyta), a genus endemic to the northeast Pacific. *Phycologia* 50:298-316.

Gilbert KJ, Andrew RL, Bock DG, Franklin MT, Kane NC, Moore J-S, Moyers BT, Renaut S, Rennison DJ, Veen T, and Vines TH. 2012. Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Molecular Ecology* 21:4925-4930.

Ginoris Y, Amaral A, Nicolau A, Coelho M, and Ferreira E. 2007. Development of an image analysis procedure for identifying protozoa and metazoa typical of activated sludge system. *Water Research* 41:2581-2589.

Gordo FP, and Bandera CC. 1997. Differentiation of Spanish strains of *Echinococcus granulosus* using larval rostellar hook morphometry. *International Journal for Parasitology* 27:41-49.

Gouws G, Stewart BA, and Reavell PE. 2001. A new species of freshwater crab (Decapoda, Potamonautidae) from the swamp forests of Kwazulu-Natal, South Africa: biochemical and morphological evidence. *Crustaceana* 74:137-160.

Gugerli F. 1997. Hybridization of *Saxifraga oppositifolia* and *S. biflora* (Saxifragaceae) in a mixed alpine population. *Plant Systematics and Evolution* 207:255-272.

Hata Y, Hashiba T, Nakamura T, Kitamura M, Ishida TA, Akimoto S-i, Sato H, and Kimura MT. 2011. Differences in leafminer (Phyllonorycter, Gracillariidae, Lepidoptera) and aphid (Tuberculatus, Aphididae, Hemiptera) composition among *Quercus dentata*, *Q. crispula*, *Q. serrata*, and their hybrids. *Journal of Forest Research* 16:309-318.

Hendriks IE, Van Duren LA, and Herman PM. 2005. Image analysis techniques: A tool for the identification of bivalve larvae? *Journal of Sea Research* 54:151-162.

Heraty JM, and Woolley JB. 1993. Separate species or polymorphism: a recurring problem in *Kapala* (Hymenoptera: Eucharitidae). *Annals of the Entomological Society of America* 86:517-531.

Hermida M, San Miguel E, Bouza C, Castro J, and Martínez P. 2009. Morphological variation in a secondary contact between divergent lineages of brown trout (*Salmo trutta*) from the Iberian Peninsula. *Genetics and Molecular Biology* 32:42-50.

Ibáñez AL, and O'Higgins P. 2011. Identifying fish scales: The influence of allometry on scale shape and classification. *Fisheries Research* 109:54-60.

Ioannidis JPA. 2014. How to make more published research true. *PLoS Med* 11:e1001747.

Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, and van Noort V. 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41:149-155.

Katoh M, and Tokimura M. 2001. Genetic and morphological identification of *Sebastiscus tertius* in the East China Sea (Scorpaeniformes: Scorpaenidae). *Ichthyological Research* 48:247-255.

López-González C, Presley SJ, Owen RD, and Willig MR. 2001. Taxonomic status of *Myotis* (Chiroptera: vespertilionidae) in Paraguay. *Journal of Mammalogy* 82:138-160.

Magud BD, Stanisavljević LŽ, and Petanović RU. 2007. Morphological variation in different populations of *Aceria anthocoptes* (Acari: Eriophyoidea) associated with the Canada thistle, Cirsium arvense, in Serbia. *Experimental and Applied Acarology* 42:173-183.

Malenke JR, Johnson KP, and Clayton DH. 2009. Host specialization differentiates cryptic species of feather-feeding lice. *Evolution* 63:1427-1438.

Marhold K, Jongepierová I, Krahulcová A, and Kucera J. 2005. Morphological and karyological differentiation of *Gymnadenia densiflora* and *G. conopsea* in the Czech Republic and Slovakia. *Preslia* 77:159-176.

McNutt M. 2014. Journals unite for reproducibility. *Science* 346:679.

Michener WK, Brunt JW, Helly JJ, Kirchner TB, and Stafford SG. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7:330-342.

Mills SC, and Côté IM. 2003. Sex-related differences in growth and morphology of blue mussels. *Journal of the Marine Biological Association of the UK* 83:1053-1057.

Nishida S, Naiki A, and Nishida T. 2005. Morphological variation in leaf domatia enables coexistence of antagonistic mites in *Cinnamomum camphora*. *Canadian Journal of Botany* 83:93-101.

Okuda N, Ito S, and Iwao H. 2003. Female mimicry in a freshwater goby Rhinogobius sp. OR. *Ichthyological Research* 50:198-200.

488 Palma L, Mira S, Cardia P, Beja P, Guillemaud T, Ferrand N, and Cancela ML. 2001. Sexing
489 Bonelli's Eagle nestlings: Morphometrics versus molecular techniques. *Journal of*
490 *Raptor Research* 35:187-193.
491 Parent GJ, Plourde S, and Turgeon J. 2011. Overlapping size ranges of *Calanus* spp. off the
492 Canadian Arctic and Atlantic Coasts: impact on species' abundances. *Journal of*
493 *Plankton Research* 33:1654-1665.
494 Pearce TA, Fields MC, and Kurita K. 2007. Discriminating shells of *Gastrocopta pentodon*
495 (Say, 1822) and *G. tappaniana* (CB Adams, 1842)(Gastropoda: Pulmonata) with an
496 example from the Delmarva Peninsula, eastern USA. *Nautilus* 121:66-75.
497 Pérez-Farrera MA, Vovides AP, Martinez-Camilo R, Melendez NM, and Iglesias C. 2009. A
498 reassessment of the *Ceratozamia miqueliana* species complex (Zamiaceae) of
499 southeastern Mexico, with comments on species relationships. *Systematics and*
500 *Biodiversity* 7:433-443.
501 Price M. 2011. To replicate or not to replicate? *Science Careers*.
502 Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using
503 multilocus genotype data. *Genetics* 155:945-959.
504 R Core Development Team. 2011. R: A Language and Environment for Statistical
505 Computing. Version 3.1.0. Vienna, Austria: R Foundation for Statistical Computing.
506 R Core Team. 2013. foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat,
507 dBase, ...
508 Radloff SE, Hepburn HR, Fuchs S, Otis GW, Hadisoesilo S, Hepburn C, and Ken T. 2005.
509 Multivariate morphometric analysis of the *Apis cerana* populations of oceanic Asia.
510 *Apidologie* 36:475.
511 Ram K. 2013. Git can facilitate greater reproducibility and increased transparency in science.
512 *Source Code for Biology and Medicine* 8:1-8.
513 Reyment RA, Blackith RE, and Campbell NR. 1984. *Multivariate Morphometrics*. London:
514 Academic Press.
515 Rigby MC, and Font WF. 2001. Statistical reanalysis of the distinction between
516 *Spirocamallanus istiblenni* and *S. monotaxis* (Nematoda: Camallanidae). *Journal of*
517 *Parasitology* 87:1210-1213.
518 Rioux-Paquette S, and Lapointe F-J. 2007. The use of shell morphometrics for the
519 management of the endangered malagasy radiated tortoise (*Geochelone radiata*).
520 *Biological Conservation* 134:31-39.
521 Ripley B, and Lapsley M. 2013. RODBC: ODBC Database Access.

522 Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions
523 MD, and Kruuk LEB. 2014. Troubleshooting public data archiving: suggestions to
524 increase participation. *PLoS Biol* 12:e1001779.
525 Ruedi M. 1995. Taxonomic revision of shrews of the genus *Crocidura* from the Sunda Shelf
526 and Sulawesi with description of two new species (Mammalia: Soricidae). *Zoological*
527 *Journal of the Linnean Society* 115:211-265.
528 Russell JC, Ringler D, Trombini A, and Le Corre M. 2011. The island syndrome and
529 population dynamics of introduced rats. *Oecologia* 167:667-676.
530 Salcedo N, Rodriguez D, Strauss R, and Baker R. 2011. The Fitzcarrald Arch: a vicariant
531 event for *Chaetostoma* (Siluriformes: Loricariidae) speciation? *Copeia* 2011:503-512.
532 Santiago-Alarcon D, and Parker PG. 2007. Sexual size dimorphism and morphological
533 evidence supporting the recognition of two subspecies in the Galápagos Dove. *The*
534 *Condor* 109:132-141.
535 Savage CJ, and Vickers AJ. 2009. Empirical Study of Data Sharing by Authors Publishing in
536 PLoS Journals. *PLoS ONE* 4:e7078.
537 Schagerl M, and Kerschbaumer M. 2009. Autecology and morphology of selected *Vaucheria*
538 species (Xanthophyceae). *Aquatic Ecology* 43:295-303.
539 Semple JC, Chmielewski JG, and Leeder C. 1991. A multivariate morphometric study and
540 revision of *Aster* subg. *Doellingeria* sect. *Triplopappus* (Compositae: Astereae): the
541 *Aster umbellatus* complex. *Canadian Journal of Botany* 69:256-276.
542 Svagelj WS, and Quintana F. 2007. Sexual size dimorphism and sex determination by
543 morphometric measurements in breeding imperial shags (*Phalacrocorax atriceps*).
544 *Waterbirds* 30:97-102.
545 Thorogood R, Brunton D, and Castro I. 2009. Simple techniques for sexing nestling hihi
546 (*Notiomystis cincta*) in the field. *New Zealand Journal of Zoology* 36:115-121.
547 Vanclay JK, Gillison AN, and Keenan RJ. 1997. Using plant functional attributes to quantify
548 site productivity and growth patterns in mixed forests. *Forest Ecology and*
549 *Management* 94:149-163.
550 Venables WN, and Ripley BD. 2002. *Modern Applied Statistics with S*. New York: Springer.
551 Vines Timothy H, Albert Arianne YK, Andrew Rose L, Débarre F, Bock Dan G, Franklin
552 Michelle T, Gilbert Kimberly J, Moore J-S, Renaut S, and Rennison Diana J. 2014.
553 The availability of research data declines rapidly with article age. *Current Biology*
554 24:94-97.

555 Vines TH, Andrew RL, Bock DG, Franklin MT, Gilbert KJ, Kane NC, Moore J-S, Moyers
556 BT, Renaut S, Rennison DJ, Veen T, and Yeaman S. 2013. Mandated data archiving
557 greatly improves access to research data. *The FASEB Journal* 27:1304-1308.
558 Wasowicz P, and Rostanski A. 2009. The use of quantitative characters in determination of
559 frequently misdiagnosed species within *Lepidium* L. sect. *Dileptium* [Brassicaceae].
560 *Acta Societatis Botanicorum Poloniae* 78:221-227.
561 Whitlock MC. 2011. Data archiving in ecology and evolution: best practices. *Trends in*
562 *Ecology & Evolution* 26:61-65.
563 Wicherts JM, Bakker M, and Molenaar D. 2011. Willingness to Share Research Data Is
564 Related to the Strength of the Evidence and the Quality of Reporting of Statistical
565 Results. *PLoS ONE* 6:e26828.
566 Wicherts JM, Borsboom D, Kats J, and Molenaar D. 2006. The poor availability of
567 psychological research data for reanalysis. *American Psychologist* 61:726-728.
568 Wicht B, Moretti M, Preatoni D, Tosi G, and Martinoli A. 2003. The presence of Soprano
569 pipistrelle *Pipistrellus pygmaeus* (Leach, 1825) in Switzerland: first molecular and
570 bioacustic evidences. *Revue Suisse de Zoologie* 110:411-426.
571 Williams CT, Dean Kildaw S, and Loren Buck C. 2007. Sex-specific differences in body
572 condition indices and seasonal mass loss in Tufted Puffins. *Journal of Field*
573 *Ornithology* 78:369-378.
574 Wolkovich EM, Regetz J, and O'Connor MI. 2012. Advances in global change research
575 require open science by individual researchers. *Global Change Biology* 18:2102-
576 2110.
577 Zaitoun IS, Tabbaa MJ, and Bdour S. 2005. Differentiation of native goat breeds of Jordan on
578 the basis of morphostructural characteristics. *Small Ruminant Research* 56:173-182.

579

580
581

Tables

Table 1. Summary of papers excluded from or included in the study, in total and listed by the statistical software originally used to analyse the data. Those included in the study are further broken down by the reasons that reanalysis was not attempted or by the results of the reanalysis. A "partial match" occurred when both matching and non-matching metrics resulted from the reanalysed, compared to the published results. The metrics considered were PVE, a discriminant function coefficient, and PAC.

Software	Excluded	Included	Incorrect data	Incomplete metadata	Data discrepancy	Reanalysed		
						No match	Partial match	Complete match
TOTAL	14	86	2 (2.3%)	7 (8.1%)	7 (8.1%)	12 (14%)	46 (53.5%)	12 (14%)
JMP	2	2	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)
MATLAB	1	2	0 (0%)	0 (0%)	1 (50%)	0 (0%)	0 (0%)	1 (50%)
R	0	5	2 (40%)	0 (0%)	1 (20%)	1 (20%)	0 (0%)	1 (20%)
SAS	1	15	0 (0%)	3 (20%)	2 (13%)	3 (20%)	2 (13%)	5 (33%)
SPSS	6	30	0 (0%)	0 (0%)	2 (7%)	5 (17%)	6 (20%)	17 (57%)
STATIST ICA	0	9	0 (0%)	1 (11%)	1 (11%)	2 (22%)	0 (0%)	5 (56%)
SYSTAT	0	8	0 (0%)	0 (0%)	0 (0%)	1 (12%)	2 (25%)	5 (62%)
Other	1	2	0 (0%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	1 (50%)
Unknown	3	13	0 (0%)	1 (8%)	0 (0%)	0 (0%)	2 (15%)	10 (77%)

Table 2. Published results and reanalyzed values of DFAs based on data files received from authors. DFAs included in the current study were categorized according to the adequacy of data files and metadata, and the reproducibility of three metrics (percent variance explained, the largest coefficient and percent assigned correctly) among those that were able to be reanalyzed. Category indicates whether the data set was excluded from the study (E), was incorrect (I), had inadequate metadata (M), displayed data discrepancies (D) or was reanalysed (R). The reasons for excluding data sets from the study or preventing us from reanalyzing the data are summarized. The reanalysis outcome was classified as a complete match (C) when all reanalyzed summary statistics were within 1% of the published values, a partial match (P) when at least one (but not all) met this criterion, and no match (N) when none met this criterion. The same classification was applied to studies using the 'close' criterion (within 5%).

Study no.	Year	Software	PVE		COEF		PAC		Categ.	Reason	Reanalysis outcome		Citation*
			Published	Reanalyzed	Published	Reanalyzed	Published	Reanalyzed			Match (within 1%)	Close (within 5%)	
1	1991	SAS	47.3	45.8			93.2	93.2	R		P	C	(Semple et al. 1991)
2	1993	SAS	83.2	84.2	18.94	20.609			R		N	P	(Heraty & Woolley 1993)
3	1995	Other	79.1	79.1	2.87	-2.868	72	71.9	R		C	C	(Darbyshire & Cayouette 1995)
4	1995	SPSS			0.892	0.7	100	100	R		P	P	(Cadrin 1995)
5	1995	SPSS	57.3	57.3			91.4	91.4	R		C	C	
6	1995	SPSS			4.02	-3.805	100	100	R		P	P	(Ruedi 1995)
7	1995	SYSTAT			-1.09	1.091	92	86.9	R		P	P	
8	1995	SYSTAT			2.115	-2.115	100	100	R		C	C	(Floate & Whitham 1995)
9	1997	Not stated							E	Not all variables are morphological			(Vanclay et al. 1997)
10	1997	SPSS	67	66.9					R		C	C	(Brysting et al. 1997)
11	1997	SPSS	96.7	92.6	1.5	-2.488	100	98.6	R		N	P	(Gordo & Bandera 1997)
12	1997	SYSTAT	99.5	99	-0.57	0.611	89	88.7	R		P	P	(Gugerli 1997)
13	1999	Not stated							M	Row groupings don't match paper			

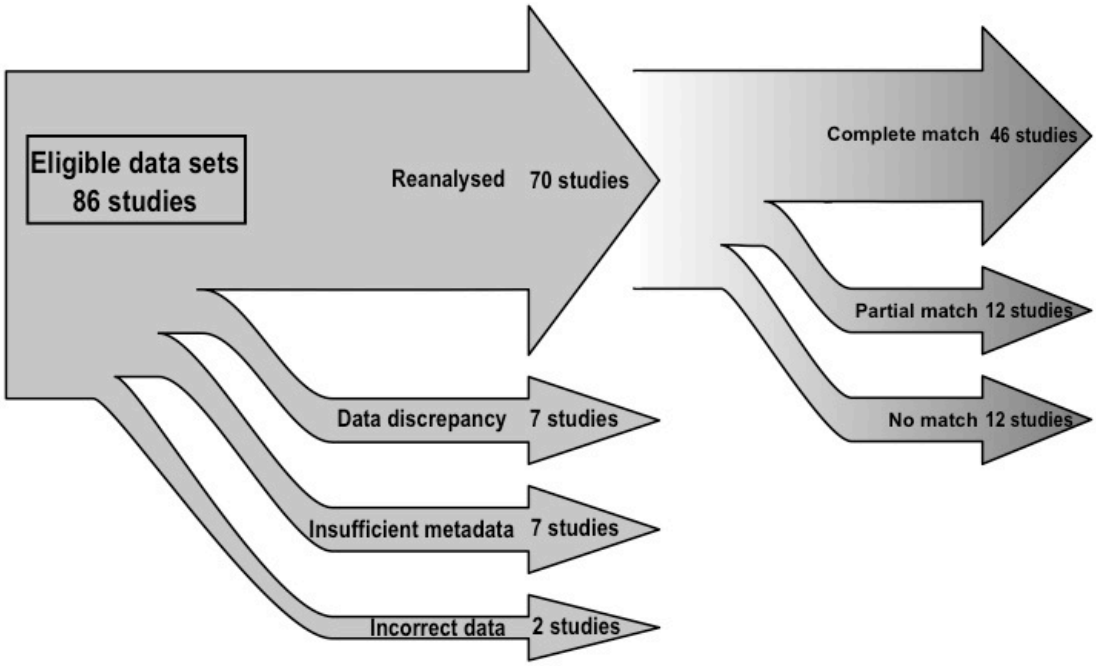
Study no.	Year	Software	PVE		COEF		PAC		Categ.	Reason	Reanalysis outcome		Citation*
			Published	Reanalyzed	Published	Reanalyzed	Published	Reanalyzed			Match (within 1%)	Close (within 5%)	
14	1999	Not stated							E	No PVE, coef or PAC			
15	1999	SAS							M	Column labels missing			
16	1999	SPSS							E	No PVE, coef or PAC			
17	1999	SPSS					73.4	73.4	R		C	C	
18	1999	SYSTAT					90	91.7	R		N	C	
19	2001	Not stated					100	100	R		C	C	(Rigby & Font 2001)
20	2001	SAS	96.7	96.4					R		C	C	
21	2001	SAS					71.3	93.8	R		N	N	
22	2001	SPSS			-1.072	-1.072	96	100	R		P	C	(Palma et al. 2001)
23	2001	SPSS					100	100	R		C	C	
24	2001	SPSS	96	96					R		C	C	(Fernández & Feliner 2001)
25	2001	SPSS			5.228	-5.228	86	82.6	R		P	C	(Katoh & Tokimura 2001)
26	2001	STATISTICA					94.4	94.4	R		C	C	
27	2003	Not stated					90.3	90.3	R		C	C	(Okuda et al. 2003)
28	2003	Not stated			-2.176	-2.176	90.6	90.6	R		C	C	
29	2003	SAS							M	Column labels in Spanish			
30	2003	SAS							D	Extra rows			
31	2003	SPSS			1.011	1.011	100	100	R		C	C	
32	2003	SPSS			3.5		81		D	Extra rows			(Mills & Côté 2003)
33	2003	SPSS							D	Missing rows and row assignments unclear			
34	2003	SPSS					88.9	87.5	R		N	C	
35	2003	SPSS			0.772	0.766	84.3	84.3	R		C	C	
36	2003	STATISTICA							M	Column labels unclear			
37	2003	SYSTAT			1.28	-1.275	81	80.6	R		C	C	(Wicht et al. 2003)
38	2005	JMP							E	No PVE, coef or PAC			(Nishida et al. 2005)
39	2005	Not stated					79.9	79.7	R		C	C	(Hendriks et al. 2005)
40	2005	Not stated	83	83.1			73	74.3	R		P	C	
41	2005	Not stated					100	100	R		C	C	(Radloff et al. 2005)
42	2005	Other							E	No PVE, coef or PAC			
43	2005	Other							M	Unclear groups			(Contrafatto 2005)

Study no.	Year	Software	PVE		COEF		PAC		Categ.	Reason	Reanalysis outcome		Citation*
			Published	Reanalyzed	Published	Reanalyzed	Published	Reanalyzed			Match (within 1%)	Close (within 5%)	
44	2005	SAS							M	Column labels missing			(Zaitoun et al. 2005)
45	2005	SAS					94.3	94.9	R		C	C	(Marhold et al. 2005)
46	2005	SPSS					46	38.2	R		N	N	(Aparicio et al. 2005)
47	2005	SPSS	55.1	55.6	0.352	0.779	71.8	70.3	R		P	P	
48	2005	STATISTICA	67.5	67					R		C	C	
49	2005	STATISTICA					97	98.8	R		N	C	
50	2005	SYSTAT					100	100	R		C	C	
51	2007	MATLAB							D	Missing columns and insufficient metadata			
52	2007	Not stated			1.1	1.097	97	96.6	R		C	C	(Svagelj & Quintana 2007)
53	2007	Not stated					87.9	87.9	R		C	C	(de la Hera et al. 2007)
54	2007	SAS			8.623	3.495	97.3	98.6	R		N	P	
55	2007	SAS					76	76.6	R		C	C	(Williams et al. 2007)
56	2007	SAS							D	Missing columns			(Pearce et al. 2007)
57	2007	SPSS							E	No PVE, coef or PAC			
58	2007	SPSS					76.9	76.9	R		C	C	(Rioux-Paquette & Lapointe 2007)
59	2007	SPSS			0.689	0.647	100	85.4	R		N	N	(Santiago-Alarcon & Parker 2007)
60	2007	SPSS	61.8	61.6					R		C	C	
61	2007	SPSS							E	Final model not given			(Conde-Padín et al. 2007)
62	2007	SPSS					84	83.3	R		C	C	
63	2007	STATISTICA					96.1	96.2	R		C	C	
64	2007	STATISTICA	93.3	93.3	-0.951	-0.951	89.2	89.2	R		C	C	
65	2007	STATISTICA			1.68	1.678	83.7	83.7	R		C	C	(Bourgeois et al. 2007)
66	2007	SYSTAT	90.4	90.4			90	90	R		C	C	
67	2009	Not stated					91.2	91.2	R		C	C	
68	2009	Not stated							E	Not DFA			
69	2009	Not stated	40.8	41.1			79	78.3	R		C	C	(Hermida et al. 2009)
70	2009	Not stated			0.242	0.084	100	100	R		P	P	(Buczkó et al. 2009)
71	2009	SAS	69	69.2	1.05	-1.053			R		C	C	(Pérez-Farrera et al. 2009)
72	2009	SAS			0.95	0.604	80	80	R		P	P	
73	2009	SPSS					100	100	R		C	C	

Study no.	Year	Software	PVE		COEF		PAC		Categ.	Reason	Reanalysis outcome		Citation*
			Published	Reanalyzed	Published	Reanalyzed	Published	Reanalyzed			Match (within 1%)	Close (within 5%)	
74	2009	SPSS							E	Data not Morphological			
75	2009	SPSS					76.4	77	R		C	C	(Thorogood et al. 2009)
76	2009	STATISTICA							D	Missing rows			
77	2009	STATISTICA					100	98.1	R		N	C	
78	2009	SYSTAT			2.8	2.795	91	91.5	R		C	C	(Berzins et al. 2009)
79	2011	JMP							E	No PVE, coef or PAC			(Hata et al. 2011)
80	2011	JMP							M	Column headings unclear			
81	2011	JMP			-7.06	7.063	100	100	R		C	C	(Gabrielson et al. 2011)
82	2011	MATLAB	65.5	65					R		C	C	(Salcedo et al. 2011)
83	2011	MATLAB							E	not classical DFA			(Capoccioni et al. 2011)
84	2011	Not stated	90	90.5					R		C	C	(Russell et al. 2011)
85	2011	R							D	Missing rows			
86	2011	R							I	Wrong file			
87	2011	R							I	Wrong file			
88	2011	R	58	88.3			56	57.1	R		N	P	
89	2011	R					80.4	80.4	R		C	C	(Dechaume-Moncharmont et al. 2011)
90	2011	SAS							E	Spanish			
91	2011	SAS					100	100	R		C	C	(Parent et al. 2011)
92	2011	SPSS	81.8	81.7					R		C	C	(Forster et al. 2010)
93	2011	SPSS	97.7	97.7			87.5	87.5	R		C	C	(Amado et al. 2011)
94	2011	SPSS	58.3	58.3			62.9	62.9	R		C	C	(Ibáñez & O'Higgins 2011)
95	2011	SPSS	87.7	87.5					R		C	C	
96	2011	SPSS							E	Final model not given			
97	2011	SPSS							E	Final model not given			(Asanidze et al. 2011)
98	2011	SPSS					100	100	R		C	C	
99	2011	SPSS					95.7	93.9	R		N	C	
100	2011	SPSS	96	89.7	1.202	0.068	100	100	R		P	P	

*Authors were contacted individually once reanalyses were performed. Only authors wishing to be identified are cited above. In addition, several authors agreed to be cited, but not identified directly (Amini et al. 2007; Audisio et al. 2001; Bulgarella et al. 2007; Ekrt et al. 2009; Foggi et al. 1999; Ginoris et al. 2007; Gouws et al. 2001; López-González et al. 2001; Magud et al. 2007; Malenke et al. 2009; Schagerl & Kerschbaumer 2009; Wasowicz & Rostanski 2009)

606 Figure 1



607

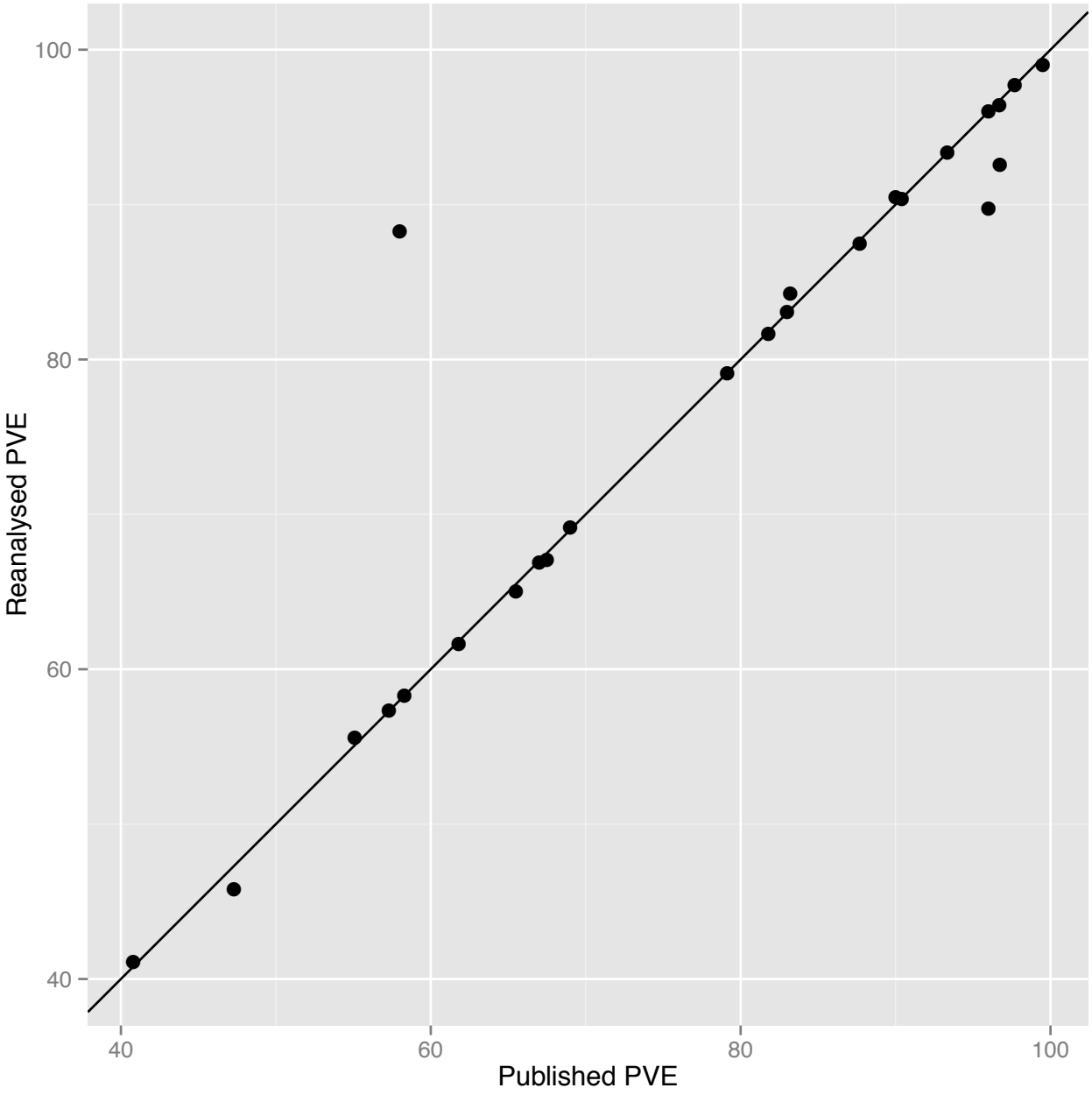
608

609 Figure 1. Summary of the reproducibility of the 70 reanalyzed data sets and of the problems
610 preventing reanalysis of 16 papers.

611

612

613 Figure 2



614

615

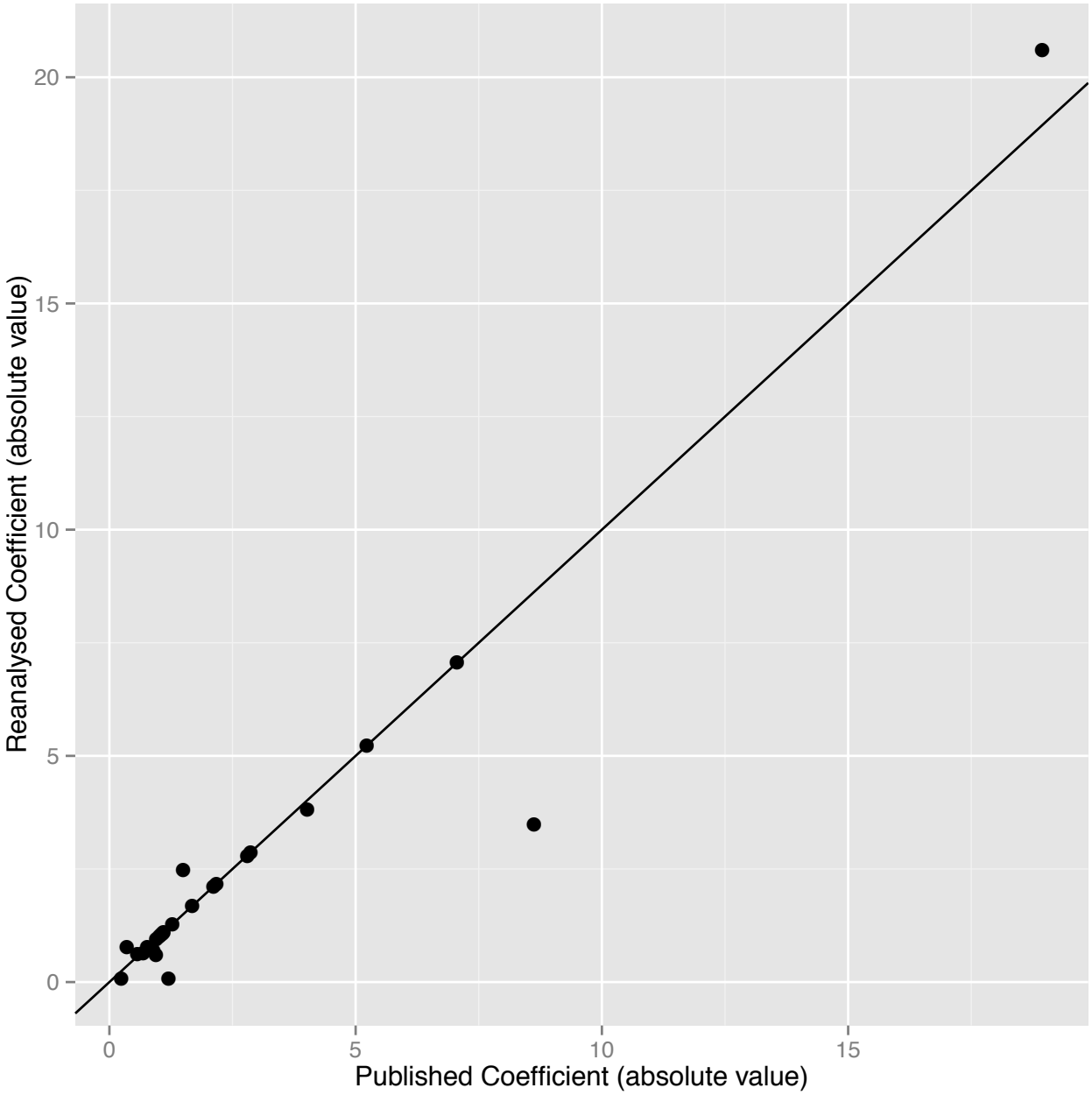
616 Figure 2. PVE values from reanalysis versus published DFA. Points on the 1:1 line represent
617 analyses differing by 1% or less.

618

619

620

621 Figure 3



622

623

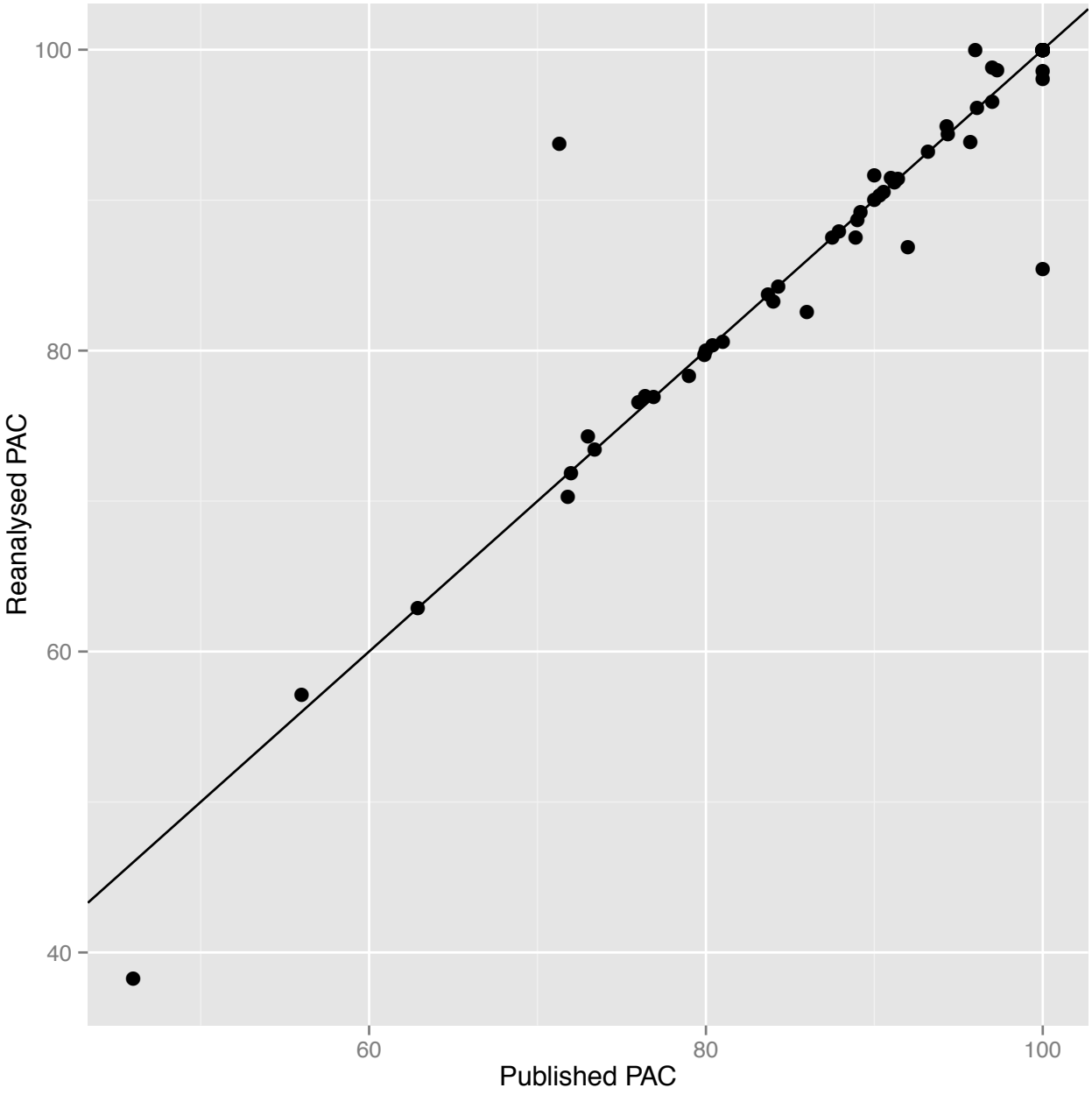
624 Figure 3. Discriminant function coefficients from the reanalysis versus the published results.
625 Absolute values are used because the signs of coefficients depends on the order of variables.
626 Points on the 1:1 line represent analyses differing by 1% or less.

627

628

629

630 Figure 4



631

632

633 Figure 4. PAC by reanalysis versus published DFA. Points on the 1:1 line represent analyses
634 differing by 1% or less.

635