Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.

Elliot Gould, School of Agriculture Food and Ecosystem Sciences, University of Melbourne, Australia

Hannah S. Fraser, School of Historical and Philosophical Studies, University of Melbourne, Australia

Timothy H. Parker, Department of Biology, Whitman College, USA. Author for Correspondence: parkerth@whitman.edu

Shinichi Nakagawa, School of Biological, Earth & Environmental Sciences, University of New South Wales, Australia

Simon C. Griffith, School of Natural Sciences, Macquarie University, Australia

Peter A. Vesk, School of Agriculture Food and Ecosystem Sciences, University of Melbourne, Australia

Fiona Fidler, School of Historical and Philosophical Studies, University of Melbourne, Australia

Daniel G. Hamilton, School of Public Health and Preventive Medicine, Monash University, Australia

Robin N Abbey-Lee, Länsstyrelsen Östergötland, Sweden

Jessica K. Abbott, Biology Department, Lund University, Sweden

Luis A. Aguirre, Department of Biology, University of Massachusetts, USA

Carles Alcaraz, Marine and Continental Waters, IRTA, Spain

Irith Aloni, Deptartment of Life Sciences, Ben Gurion University of the Negev, Israel

Drew Altschul, Department of Psychology, The University of Edinburgh, UK

Kunal Arekar, Centre for Ecological Sciences, Indian Institute of Science, India

Jeff W. Atkins, Southern Research Station, USDA Forest Service, USA

Joe Atkinson, Center for Ecological Dynamics in a Novel Biosphere (ECONOVO), Department of Biology, Aarhus University, Denmark

Christopher M. Baker, School of Mathematics and Statistics, University of Melbourne, Australia

Meghan Barrett, Biology, Indiana University Purdue University Indianapolis, USA

Kristian Bell, School of Life and Environmental Sciences, Deakin University, Australia

Suleiman Kehinde Bello, Department of Arid Land Agriculture, King Abdulaziz University, Kingdom of Saudi Arabia

Iván Beltrán, Department of Biological Sciences, Macquarie University, Australia

Bernd J. Berauer, Department of Plant Ecology, University of Hohenheim, Institute of Landscape and Plant Ecology, Germany

Michael Grant Bertram, Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

Peter D. Billman, Department of Ecology and Evolutionary Biology, University of Connecticut, USA

Charlie K. Blake, STEM Center, Southern Illinois University Edwardsville, USA

Shannon Blake, University of Guelph, Canada

Louis Bliard, Department of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland

Andrea Bonisoli-Alquati, Department of Biological Sciences, California State Polytechnic University, Pomona, USA

Timothée Bonnet, Centre d'Études Biologiques de Chizé, UMR 7372 Université de la Rochelle - Centre National de la Recherche Scientifique, France

Camille Nina Marion Bordes, Faculty of Life Sciences, Bar Ilan University, Israel

Aneesh P. H. Bose, Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

Thomas Botterill-James, School of Natural Sciences, University of Tasmania, Australia

Melissa Anna Boyd, Whitebark Institute, USA

Sarah A. Boyle, Department of Biology, Rhodes College, USA

Tom Bradfer-Lawrence, Centre for Conservation Science, RSPB, UK

Jennifer Bradham, Environmental Studies, Wofford College, USA

Jack A. Brand, Department of Wildlife, Fish and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

Martin I. Brengdahl, IFM Biology, Linköping University, Sweden

Martin Bulla, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Czech Republic

Luc Bussière, Biological and Environmental Sciences & Gothenburg Global Biodiversity Centre, University of Gothenburg, Sweden

Ettore Camerlenghi, School of Biological Sciences, Monash University, Australia

Sara E. Campbell, Ecology and Evolutionary Biology, University of Tennessee Knoxville, USA

Leonardo L. F. Campos, Departamento de Ecologia e Zoologia, Universidade Federal de Santa Catarina, Brazil

Anthony Caravaggi, School of Biological and Forensic Sciences, University of South Wales, UK

Pedro Cardoso, Centre for Ecology, Evolution and Environmental Changes (cE3c) & CHANGE - Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, Portugal

Charles J.W. Carroll, Forest and Rangeland Stewardship, Colorado State University, USA

Therese A. Catanach, Department of Ornithology, Academy of Natural Sciences of Drexel University, USA

Xuan Chen, Biology, Salisbury University, USA

Heung Ying Janet Chik, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Netherlands

Emily Sarah Choy, Department of Biology, McMaster University, Canada

Alec Philip Christie, Department of Zoology, University of Cambridge, UK

Angela Chuang, Entomology and Nematology, University of Florida, USA

Amanda J. Chunco, Environmental Studies, Elon University, USA

Bethany L. Clark, BirdLife International, UK

Andrea Contina, School of Integrative Biological and Chemical Sciences, The University of Texas Rio Grande Valley, USA

Garth A. Covernton, Department of Ecology and Evolutionary Biology, University of Toronto, Canada

Murray P. Cox, Department of Statistics, University of Auckland, New Zealand

Kimberly A. Cressman, Catbird Stats, LLC, USA

Marco Crotti, School of Biodiversity, One Health & Veterinary Medicine, University of Glasgow, UK

Connor Davidson Crouch, School of Forestry, Northern Arizona University, USA

Pietro B. D'Amelio, Department of Behavioural Neurobiology, Max Planck Institute for Biological Intelligence, Germany

Alexandra Allison de Sousa, School of Sciences: Center for Health and Cognition, Bath Spa University, UK

Timm Fabian Döbert, Department of Biological Sciences, University of Alberta, Canada

Ralph Dobler, Applied Zoology, TU Dresden, Germany

Adam J. Dobson, School of Molecular Biosciences, College of Medical Veterinary & Life Sciences, University of Glasgow, UK

Tim S. Doherty, School of Life and Environmental Sciences, The University of Sydney, Australia

Szymon Marian Drobniak, Institute of Environmental Sciences, Jagiellonian University, Poland

Alexandra Grace Duffy, Biology Department, Brigham Young University, USA

Alison B. Duncan, Institute of Evolutionary Sciences Montpellier, University of Montpellier, CNRS, IRD., France

Robert P. Dunn, Baruch Marine Field Laboratory, University of South Carolina, USA

Jamie Dunning, Department of Life Sciences, Imperial College London, UK

Trishna Dutta, European Forest Institute, Germany

Luke Eberhart-Hertel, Department of Ornithology, Max Planck Institute for Biological Intelligence, Germany

Jared Alan Elmore, Forestry and Environmental Conservation, National Bobwhite and Grassland Initiative, Clemson University, USA

Mahmoud Medhat Elsherif, Department of Psychology and Vision Science, University of Birmingham, Baily Thomas Grant, UK

Holly M. English, School of Biology and Environmental Science, University College Dublin, Ireland

David C. Ensminger, Department of Biological Sciences, San José State University, USA

Ulrich Rainer Ernst, Apicultural State Institute, University of Hohenheim, Germany

Stephen M. Ferguson, Department of Biology, St. Norbert College, USA

Esteban Fernandez-Juricic, Department of Biological Sciences, Purdue University, USA

Thalita Ferreira-Arruda, Biodiversity, Macroecology & Biogeography, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Germany

John Fieberg, Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, USA

Elizabeth A. Finch, CABI, UK

Evan A. Fiorenza, Department of Ecology and Evolutionary Biology, School of Biological Sciences, University of California, Irvine, USA

David N. Fisher, School of Biological Sciences, University of Aberdeen, UK

Amélie Fontaine, Department of Natural Resource Sciences, McGill University, Canada

Wolfgang Forstmeier, Department of Ornithology, Max Planck Institute for Biological Intelligence, Germany

Yoan Fourcade, Institute of Ecology and Environmental Sciences (iEES), Univ. Paris-Est Creteil, France

Graham S. Frank, Department of Forest Ecosystems and Society, Oregon State University, USA

Cathryn A. Freund, Wake Forest University, USA

Eduardo Fuentes-Lillo, Laboratorio de Invasiones Biológicas (LIB), Instituto de Ecología y Biodiversidad, Chile

Sara L. Gandy, Institute for Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, UK

Dustin G. Gannon, Department of Forest Ecosystems and Society, College of Forestry, Oregon State University, USA

Ana I. García-Cervigón, Biodiversity and Conservation Area, Rey Juan Carlos University, Spain

Alexis C. Garretson, Graduate School of Biomedical Sciences, Tufts University, USA

Xuezhen Ge, Department of Integrative Biology, University of Guelph, Canada

William L. Geary, School of Life and Environmental Sciences (Burwood Campus), Deakin University, Australia

Charly Géron, CNRS, University of Rennes, France

Marc Gilles, Department of Behavioural Ecology, Bielefeld University, Germany

Antje Girndt, Fakultät für Biologie, Arbeitsgruppe Evolutionsbiologie, Universität Bielefeld, Germany

Daniel Gliksman, Chair of Meteorology, Institute for Hydrology and Meteorology, Faculty of Environmental Sciences, Technische Universität Dresden, Germany

Harrison B. Goldspiel, Department of Wildlife, Fisheries, and Conservation Biology, University of Maine, USA

Dylan G. E. Gomes, Department of Biological Sciences, Boise State University, USA

Megan Kate Good, School of Agriculture, Food and Ecosystem Sciences, The University of Melbourne, Australia

Sarah C. Goslee, Pastures Systems and Watershed Management Research Unit, USDA Agricultural Research Service, USA

J. Stephen Gosnell, Department of Natural Sciences, Baruch College, City University of New York, USA

Eliza M. Grames, Department of Biological Sciences, Binghamton University, USA

Paolo Gratton, Dipartimento di Biologia, Università di Roma "Tor Vergata", Italy

Nicholas M. Grebe, Department of Anthropology, University of Michigan, USA

Skye M. Greenler, College of Forestry, Oregon State University, USA

Maaike Griffioen, University of Antwerp, Belgium

Daniel M. Griffith, Earth & Environmental Sciences, Wesleyan University, USA

Frances J. Griffith, Yale School of Medicine, Department of Psychiatry, Yale University, USA

Jake J. Grossman, Biology Department and Environmental Studies Department, St. Olaf College, USA

Ali Güncan, Department of Plant Protection, Faculty of Agriculture, Ordu University, Turkey

Stef Haesen, Department of Earth and Environmental Sciences, KU Leuven, Belgium

James G. Hagan, Department of Marine Sciences, University of Gothenburg, Sweden

Heather A. Hager, Department of Biology, Wilfrid Laurier University, Canada

Jonathan Philo Harris, Natural Resource Ecology and Management, Iowa State University, USA

Natasha Dean Harrison, School of Biological Sciences, University of Western Australia, Australia

Sarah Syedia Hasnain, Department of Biological Sciences, Middle East Technical University, Turkey

Justin Chase Havird, Dept. of Integrative Biology, University of Texas at Austin, USA

Andrew J. Heaton, Grand Bay National Estuarine Research Reserve, USA

María Laura Herrera-Chaustre, Universidad de los Andes, Colombia

Tanner J. Howard

Bin-Yan Hsu, Department of Biology, University of Turku, Finland

Fabiola Iannarilli, Dept of Fisheries, Wildlife and Conservation Biology, University of Minnesota, USA

Esperanza C. Iranzo, Instituto de Ciencia Animal. Facultad de Ciencias Veterinarias, Universidad Austral de Chile, Chile

Erik N. K. Iverson, Department of Integrative Biology, The University of Texas at Austin, USA

Saheed Olaide Jimoh, Department of Botany, University of Wyoming, USA

Douglas H. Johnson, Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, USA

Martin Johnsson, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden

Jesse Jorna, Department of Biology, Brigham Young University, Brigham Young University, USA

Tommaso Jucker, School of Biological Sciences, University of Bristol, UK

Martin Jung, International Institute for Applied Systems Analysis (IIASA), Austria

Ineta Kačergytė, Department of Ecology, Swedish University of Agricultural Sciences, Sweden

Oliver Kaltz, Université de Montpellier, France

Alison Ke, Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, USA

Clint D. Kelly, Département des Sciences biologiques, Université du Québec à Montréal, Canada

Katharine Keogan, Institute of Evolutionary Biology, University of Edinburgh, UK

Friedrich Wolfgang Keppeler, Center for Limnology, Center for Limnology, University of Wisconsin - Madison, USA

Alexander K. Killion, Center for Biodiversity and Global Change, Yale University, USA

Dongmin Kim, Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, USA

David P. Kochan, Institute of Environment and Department of Biological Sciences, Florida International University, USA

Peter Korsten, Department of Life Sciences, Aberystwyth University, UK

Shan Kothari, Institut de recherche en biologie végétale, Université de Montréal, Canada

Jonas Kuppler, Institute of Evolutionary Ecology and Conservation Genomics, Ulm University, Germany

Jillian M. Kusch, Department of Biology, Memorial University of Newfoundland, Canada

Malgorzata Lagisz, Evolution & Ecology Research Centre and School of Biological, Earth & Environmental Sciences, University of New South Wales, Australia

Kristen Marianne Lalla, Department of Natural Resource Sciences, McGill University, Canada

Daniel J. Larkin, Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota-Twin Cities, USA

Courtney L. Larson, The Nature Conservancy, USA

Katherine S. Lauck, Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, USA

M. Elise Lauterbur, Ecology and Evolutionary Biology, University of Arizona, USA

Alan Law, Biological and Environmental Sciences, University of Stirling, UK

Don-Jean Léandri-Breton, Department of Natural Resource Sciences, McGill University, Canada

Jonas J. Lembrechts, Department of Biology, University of Antwerp, Belgium

Kiara L'Herpiniere, Natural sciences, Macquarie University, Australia

Eva J. P. Lievens, Aquatic Ecology and Evolution Group, Limnological Institute, University of Konstanz, Germany

Daniela Oliveira de Lima, Campus Cerro Largo, Universidade Federal da Fronteira Sul, Brazil

Shane Lindsay, School of Psychology and Social Work, University of Hull, UK

Martin Luquet, UMR 1224 ECOBIOP, Université de Pau et des Pays de l’Adour, France

Ross MacLeod, School of Biological & Environmental Sciences, Liverpool John Moores University, UK

Kirsty H. Macphie, Institute of Ecology and Evolution, University of Edinburgh, UK

Kit Magellan, Cambodia

Magdalena M. Mair, Statistical Ecotoxicology, Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Germany

Lisa E. Malm, Ecology and Environmental Science, Umeå University, Sweden

Stefano Mammola, Molecular Ecology Group (MEG), Water Research Institute (IRSA), National Research Council of Italy (CNR), Italy

Caitlin P. Mandeville, Department of Natural History, Norwegian University of Science and Technology, Norway

Michael Manhart, Center for Advanced Biotechnology and Medicine, Rutgers University Robert Wood Johnson Medical School, USA

Laura Milena Manrique-Garzon, Departamento de Ciencias Biológicas, Universidad de los Andes, Colombia

Elina Mäntylä, Department of Biology, University of Turku, Finland

Philippe Marchand, Institut de recherche sur les forêts, Université du Québec en Abitibi-Témiscamingue, Canada

Benjamin Michael Marshall, Biological and Environmental Sciences, University of Stirling, UK

Charles A. Martin, Université du Québec à Trois-Rivières, Canada

Dominic Andreas Martin, Institute of Plant Sciences, University of Bern, Switzerland

Jake Mitchell Martin, Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

April Robin Martinig, School of Biological, Earth and Environmental Sciences, University of New South Wales, Australia

Erin S. McCallum, Department of Wildlife, Fish and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

Mark McCauley, Whitney Laboratory for Marine Bioscience, University of Florida, USA

Sabrina M. McNew, Ecology and Evolutionary Biology, University of Arizona, USA

Scott J. Meiners, Biological Sciences, Eastern Illinois University, USA

Thomas Merkling, Centre d'Investigations Clinique Plurithématique - Institut Lorrain du Coeur et des Vaisseaux, Université de Lorraine, Inserm1433 CIC-P CHRU de Nancy, France

Marcus Michelangeli, Department of Wildlife, Fish and Environmental Studies, Swedish University of Agricultural Sciences, Sweden

Maria Moiron, Evolutionary biology department, Bielefeld University, Germany

Bruno Moreira, Department of Ecology and global change, Centro de Investigaciones sobre Desertificación, Consejo Superior de Investigaciones Cientificas (CIDE-CSIC/UV/GV), Spain

Jennifer Mortensen, Department of Biological Sciences, University of Arkansas, USA

Benjamin Mos, School of the Environment, Faculty of Science, The University of Queensland, Australia

Taofeek Olatunbosun Muraina, Department of Animal Health and Production, Oyo State College of Agriculture and Technology, Nigeria

Penelope Wrenn Murphy, Department of Forest & Wildlife Ecology, University of Wisconsin-Madison, USA

Luca Nelli, School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, UK

Petri Niemelä, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Finland

Josh Nightingale, South Iceland Research Centre, University of Iceland, Iceland

Gustav Nilsonne, Department of Clinical Neuroscience, Karolinska Institutet, Sweden

Sergio Nolazco, School of Biological Sciences, Monash University, Australia

Sabine S. Nooten, Animal Ecology and Tropical Biology, University of Würzburg, Germany

Jessie Lanterman Novotny, Biology, Hiram College, USA

Agnes Birgitta Olin, Department of Aquatic Resources, Swedish University of Agricultural Sciences, Sweden

Chris L. Organ, Department of Earth Sciences, Montana State University, USA

Kate L. Ostevik, Department of Evolution, Ecology, and Organismal Biology, University of California, Riverside, USA

Facundo Xavier Palacio, Sección Ornitología, Universidad Nacional de La Plata, Argentina

Matthieu Paquet, Department of Ecology, Swedish University of Agricultural Sciences, Sweden

Darren James Parker, Bangor University, UK

David J. Pascall, MRC Biostatistics Unit, University of Cambridge, UK

Valerie J. Pasquarella, Harvard Forest, Harvard University, USA

John Harold Paterson, Biological and Environmental Sciences, University of Stirling, Scotland

Ana Payo-Payo, Departamento de Biodiversidad, Ecología y Evolución., Universidad Complutense de Madrid, Spain

Karen Marie Pedersen, Biology Department, Technische Universität Darmstadt, Germany

Grégoire Perez, UMR 1309 ASTRE, CIRAD, France

Kayla I. Perry, Department of Entomology, The Ohio State University, USA

Patrice Pottier, Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, The University of New South Wales, Australia

Michael J. Proulx, Department of Psychology, University of Bath, UK

Raphaël Proulx, Chaire de recherche en intégrité écologique, Université du Québec à Trois-Rivières, Canada

Jessica L Pruett, Mississippi Based RESTORE Act Center of Excellence, University of Southern Mississippi, USA

Veronarindra Ramananjato, Department of Integrative Biology, University of California, Berkeley, USA

Finaritra Tolotra Randimbiarison, Mention Zoologie et Biodiversité Animale, Université d'Antananarivo, Madagascar

Onja H. Razafindratsima, Department of Integrative Biology, University of California, Berkeley, USA

Diana J. Rennison, Department of Ecology, Behavior and Evolution, University of California, San Diego, USA

Federico Riva, Institute for Environmental Sciences, VU Amsterdam, The Netherlands

Sepand Riyahi, Department of Evolutionary Anthropology, University of Vienna, Austria

Michael James Roast, Konrad Lorenz Institute for Ethology, University of Veterinary Medicine, Austria

Felipe Pereira Rocha, School of Biological Sciences, The University of Hong Kong, China

Dominique G. Roche, Institut de biologie, Université de Neuchâtel, Switzerland

Cristian Román-Palacios, School of Information, University of Arizona, USA

Michael S. Rosenberg, Center for Biological Data Science, Virginia Commonwealth University, USA

Jessica Ross, University of Wisconsin, USA

Freya E. Rowland, School of the Environment, Yale University, USA

Deusdedith Rugemalila, Institute of the Environment, Florida International University, USA

Avery L. Russell, Department of Biology, Missouri State University, USA

Suvi Ruuskanen, Department of Biological and Environmental Science, University of Jyväskylä, Finland

Patrick Saccone, Institute for Interdisciplinary Mountain Research, OeAW (Austrian Academy of Sciences), Austria

Asaf Sadeh, Department of Natural Resources, Newe Ya'ar Research Center, Agricultural Research Organization (Volcani Institute), Israel

Stephen M. Salazar, Department of Animal Behaviour, Bielefeld University, Germany

Kris Sales, Office for National Statistics, UK

Pablo Salmón, Institute of Avian Research "Vogelwarte Helgoland", Germany

Alfredo Sánchez-Tójar, Department of Evolutionary Biology, Bielefeld University, Germany

Leticia Pereira Santos, Ecology Department, Universidade Federal de Goiás, Brazil

Francesca Santostefano, University of Exeter, University of Exeter, UK

Hayden T. Schilling, New South Wales Department of Primary Industries Fisheries, Australia

Marcus Schmidt, Research Data Management, Leibniz Centre for Agricultural Landscape Research (ZALF), Germany

Tim Schmoll, Evolutionary Biology, Bielefeld University, Germany

Adam C. Schneider, Biology Department, University of Wisconsin-La Crosse, USA

Allie E. Schrock, Department of Evolutionary Anthropology, Duke University, USA

Julia Schroeder, Department of Life Sciences, Imperial College London, UK

Nicolas Schtickzelle, Earth and Life Institute, Ecology and Biodiversity, UCLouvain, Belgium

Nick L. Schultz, Future Regions Research Centre, Federation University Australia, Australia

Drew A. Scott, United States Department of Agriculture- Agricultural Research Service-, USA

Michael Peter Scroggie, Arthur Rylah Insitute for Environmental Research, Australia

Julie Teresa Shapiro, Epidemiology and Surveillance Support Unit, University of Lyon - French Agency for Food, Environmental and Occupational Health and Safety (ANSES), France

Nitika Sharma, UCLA Anderson Center for Impact, University of California, Los Angeles, USA

Caroline L. Shearer, Department of Evolutionary Anthropology, Duke University, USA

Diego Simón, Facultad de Ciencias, Universidad de la República, Uruguay

Michael I. Sitvarin, Independent researcher, USA

Fabrício Luiz Skupien, Programa de Pós-Graduação em Ecologia, Instituto de Biologia, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Brazil

Heather Lea Slinn, Vive Crop Protection, Canada

Grania Polly Smith, University of Cambridge, UK

Jeremy A. Smith, British Trust for Ornithology, UK

Rahel Sollmann, Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, USA

Kaitlin Stack Whitney, Science, Technology & Society Department, Rochester Institute of Technology, USA

Shannon Michael Still, Nomad Ecology, USA

Erica F. Stuber, Wildland Resources Department, Utah State University, USA

Guy F. Sutton, Center for Biological Control, Department of Zoology and Entomology, Rhodes University, South Africa

Ben Swallow, School of Mathematics and Statistics and Centre for Research in Ecological and Environmental Modelling, University of St Andrews, UK

Conor Claverie Taff, Department of Ecology and Evolutionary Biology, Cornell University, USA

Elina Takola, Department of Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Germany

Andrew J. Tanentzap, Ecosystems and Global Change Group, School of the Environment, Trent University, Canada

Rocío Tarjuelo, Instituto Universitario de Investigación en Gestión Forestal Sostenible (iuFOR), Universidad de Valladolid, Spain

Richard J. Telford, Department of Biological Sciences, University of Bergen, Norway

Christopher J. Thawley, Department of Biological Science, University of Rhode Island, USA

Hugo Thierry, Department of Geography, McGill University, Canada

Jacqueline Thomson, Integrative Biology, University of Guelph, Canada

Svenja Tidau, School of Biological and Marine Sciences, University of Plymouth, UK

Emily M. Tompkins, Biology Deptartment, Wake Forest University, USA

Claire Marie Tortorelli, Plant Sciences, University of California, Davis, USA

Andrew Trlica, College of Natural Resources, North Carolina State University, USA

Biz R. Turnell, Institute of Zoology, Technische Universität Dresden, Germany

Lara Urban, Helmholtz AI, Helmholtz Zentrum Muenchen, Germany

Stijn Van de Vondel, Department of Biology, University of Antwerp, Belgium

Jessica Eva Megan van der Wal, FitzPatrick Institute of African Ornithology, University of Cape Town, South Africa

Jens Van Eeckhoven, Department of Cell & Developmental Biology, Division of Biosciences, University College London, UK

Francis van Oordt, Natural Resource Sciences, McGill University, Canada

K. Michelle Vanderwel, Biology, University of Saskatchewan, Canada

Mark C. Vanderwel, Department of Biology, University of Regina, Canada

Karen J. Vanderwolf, Biology, University of Waterloo, Canada

Juliana Vélez, Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota, USA

Diana Carolina Vergara-Florez, Department of Ecology & Evolutionary Biology, University of Michigan, USA

Brian C. Verrelli, Center for Biological Data Science, Virginia Commonwealth University, USA

Marcus Vinícius Vieira, Dept. Ecologia, Instituto de Biologia, Universidade Federal do Rio de Janeiro, Brazil

Nora Villamil, Lothian Analytical Services, Public Health Scotland, UK

Valerio Vitali, Institute for Evolution and Biodiversity, University of Muenster, Germany

Julien Vollering, Department of Environmental Sciences, Western Norway University of Applied Sciences, Norway

Jeffrey Walker, Department of Biological Sciences, University of Southern Maine, USA

Xanthe J. Walker, Center for Ecosystem Science and Society, Northern Arizona University, USA

Jonathan A. Walter, Center for Watershed Sciences, University of California, Davis, USA

Pawel Waryszak, School of Agriculture and Environmental Science, University of Southern Queensland, Australia

Ryan J. Weaver, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, USA

Ronja E. M. Wedegärtner, Fram Project AS, Norway

Daniel L. Weller, Department of Food Science & Technology, Virginia Polytechnic Institute and State University, USA

Shannon Whelan, Department of Natural Resource Sciences, McGill University, Canada

Rachel Louise White, School of Applied Sciences, University of Brighton, UK

David William Wolfson, Department of Fisheries, Wildlife and Conservation Biology, University of Minnesota, USA

Andrew Wood, Department of Biology, University of Oxford, UK

Scott W. Yanco, Department of Integrative Biology, University of Colorado, Denver, USA

Jian D. L. Yen, Arthur Rylah Institute for Environmental Research, Australia

Casey Youngflesh, Ecology, Evolution, and Behavior Program, Michigan State University, USA

Giacomo Zilio, ISEM, University of Montpellier, CNRS, France

Cédric Zimmer, Laboratoire d'Ethologie Expérimentale et Comparée, LEEC, UR4443, Université Sorbonne Paris Nord, USA

Gregory Mark Zimmerman, Department of Science and Environment, Lake Superior State University, USA

Rachel A. Zitomer, Department of Forest Ecosystems and Society, Oregon State University, USA

Abstract

Although variation in effect sizes and predicted values among studies of similar phenomena is inevitable, such variation far exceeds what might be produced by sampling error alone. One possible explanation for variation among results is differences among researchers in the decisions they make regarding statistical analyses. A growing array of studies has explored this analytical variability in different fields and has found substantial variability among results despite analysts having the same data and research question. Many of these studies have been in the social sciences, but one small 'many analyst' study found similar variability in ecology. We expanded the scope of this prior work by implementing a large-scale empirical exploration of the variation in effect sizes and model predictions generated by the analytical decisions of different researchers in ecology and evolutionary biology. We used two unpublished datasets, one from evolutionary ecology (blue tit, *Cyanistes caeruleus*, to compare sibling number and nestling growth) and one from conservation ecology (*Eucalyptus*, to compare grass cover and tree seedling recruitment). The project leaders recruited 174 analyst teams, comprising 246 analysts, to investigate the answers to prespecified research questions. Analyses conducted by these teams yielded 141 usable effects (compatible with our meta-analyses and with all necessary information provided) for the blue tit dataset, and 85 usable effects for the *Eucalyptus* dataset. We found substantial heterogeneity among results for both datasets, although the patterns of variation differed between them. For the blue tit analyses, the average effect was convincingly negative, with less growth for nestlings living with more siblings, but there was near continuous variation in effect size from large negative effects to effects near zero, and even effects crossing the traditional threshold of statistical significance in the opposite direction. In contrast, the average relationship between grass cover and *Eucalyptus* seedling number was only slightly negative and not convincingly different from zero, and most effects ranged from weakly negative to weakly positive, with about a third of effects crossing the traditional threshold of significance in one direction or the other. However, there were also several striking outliers in the *Eucalyptus* dataset, with effects far from zero. For both datasets, we found substantial variation in the variable selection and random effects structures among analyses, as well as in the ratings of the analytical methods by peer reviewers, but we found no strong relationship between any of these and deviation from the meta-analytic mean. In other words, analyses with results that were far from the mean were no more or less likely to have dissimilar variable sets, use random effects in their models, or receive poor peer reviews than those analyses that found results that were close to the mean. The existence of substantial variability among analysis outcomes raises important questions about how ecologists and evolutionary biologists should interpret published results, and how they should conduct analyses in the future.

Introduction

One value of science derives from its production of replicable, and thus reliable, results. When we repeat a study using the original methods, we should be able to expect a similar result. However, perfect replicability is not a reasonable goal. Effect sizes will vary, and even reverse in sign, by chance alone (Gelman and Weakliem 2009). Observed patterns can differ for other reasons as well. It could be that we do not sufficiently understand the conditions that led to the original result so when we seek to replicate it, the conditions differ due to some 'hidden moderator'. This hidden moderator hypothesis is described by meta-analysts in ecology and evolutionary biology as 'true biological heterogeneity' (Senior et al. 2016). This idea of true heterogeneity is popular in ecology and evolutionary biology, and there are good reasons to expect it in the complex systems in which we work (Shavit and Ellison 2017). However, despite similar expectations in psychology, recent evidence in that discipline contradicts the hypothesis that moderators are common obstacles to replicability, as variability in results in a large 'many labs' collaboration was mostly unrelated to commonly hypothesized moderators such as the conditions under which the studies were administered (Klein et al. 2018). Another possible explanation for variation in effect sizes is that researchers often present biased samples of results, thus reducing the likelihood that later studies will produce similar effect sizes (Open Science Collaboration 2015; Parker et al. 2016; Forstmeier, Wagenmakers, and Parker 2017; Fraser et al. 2018; Parker and Yang 2023). It also may be that although researchers did successfully replicate the conditions, the experiment, and measured variables, analytical decisions differed sufficiently among studies to create divergent results (Simonsohn, Simmons, and Nelson 2015; Silberzahn et al. 2018).

Analytical decisions vary among studies because researchers have many options. Researchers need to decide how to exclude possibly anomalous or unreliable data, how to construct variables, which variables to include in their models, and which statistical methods to use. Depending on the dataset, this short list of choices could encompass thousands or millions of possible alternative specifications (Simonsohn, Simmons, and Nelson 2015). However, researchers making these decisions presumably do so with the goal of doing the best possible analysis, or at least the best analysis within their current skill set. Thus, it seems likely that some specification options are more probable than others, possibly because they have previously been shown (or claimed) to be better, or because they are more well known. Of course, some of these different analyses (maybe many of them) may be equally valid alternatives. Regardless, on probably any topic in ecology and evolutionary biology, we can encounter differences in choices of data analysis. The extent of these differences in analyses and the degree to which these differences influence the outcomes of analyses and therefore studies' conclusions are important empirical questions. These questions are especially important given that many papers draw conclusions after applying a single method, or even a single statistical model, to analyze a dataset.

The possibility that different analytical choices could lead to different outcomes has long been recognized (Gelman and Loken 2013), and various efforts to address this possibility have been pursued in the literature. For instance, one common method in ecology and evolutionary biology involves creating a set of candidate models, each consisting of a different (though often similar) set of predictor variables, and then, for the predictor variable of interest, averaging the slope across all models (i.e. model averaging) (Burnham and Anderson 2002; Grueber et al. 2011). This method reduces the chance that a conclusion is contingent upon a single model specification, though use and interpretation of this method is not without challenges (Grueber et al. 2011). Further, the models compared to each other typically differ only in the inclusion or exclusion of certain predictor variables and not in other important ways, such as methods of parameter estimation. More explicit examination of outcomes of differences in model structure, model type, data exclusion, or other analytical choices can be implemented through sensitivity analyses (e.g., Noble et al. 2017). Sensitivity analyses, however, are typically rather narrow in scope, and are designed to assess the sensitivity of analytical outcomes to a particular analytical choice rather than to a large universe of choices. Recently, however, analysts in the social sciences have proposed extremely thorough sensitivity analysis, including 'multiverse analysis' (Steegen et al. 2016) and the 'specification curve' (Simonsohn, Simmons, and Nelson 2015), as a means of increasing the reliability of results. With these methods, researchers identify relevant decision points encountered during analysis and conduct the analysis many times to incorporate many plausible decisions made at each of these points. The study's conclusions are then based on a broad set of the possible analyses and so allow the analyst to distinguish between robust conclusions and those that are highly contingent on particular model specifications. These are useful outcomes, but specifying a universe of possible modelling decisions is not a trivial undertaking. Further, the analyst's knowledge and biases will influence decisions about the boundaries of that universe, and so there will always be room for disagreement among analysts about what to include. Including more specifications is not necessarily better. Some analytical decisions are better justified than others, and including biologically implausible specifications may undermine this process. Regardless, these powerful methods have yet to be adopted, and even the more limited forms of sensitivity analyses are not particularly widespread. Most studies publish a small set of analyses and so the existing literature does not provide much insight into the degree to which published results are contingent on analytical decisions.

Despite the potential major impacts of analytical decisions on variance in results, the outcomes of different individuals' data analysis choices have only recently begun to receive much empirical attention. The only formal exploration of this that we were aware of when we submitted our Stage 1 manuscript were (1) an analysis in social science that asked whether male professional football (soccer) players with darker skin tone were more likely to be issued red cards (ejection from the game for rule violation) than players with lighter skin tone (Silberzahn et al. 2018) and (2) an analysis in neuroimaging which evaluated nine separate hypotheses involving the neurological responses detected with fMRI in 108 participants divided between two treatments in a decision making task (Botvinik-Nezer et al. 2020). Several others have been published since (e.g., Huntington-Klein et al. 2021; Schweinsberg et al. 2021; Breznau et al. 2022; Coretta et al. 2023), and we recently learned of an earlier small study in ecology (Stanton-Geddes, Freitas, and Sales Dambros 2014). In the red card study, 29 teams designed and implemented analyses of a dataset provided by the study coordinators (Silberzahn et al. 2018). Analyses were peer reviewed (results blind) by at least two other participating analysts; a level of scrutiny consistent with standard pre-publication peer review. Among the final 29 analyses, odds-ratios varied from 0.89 to 2.93, meaning point estimates varied from having players with lighter skin tones receive more red cards (odds ratio < 1) to a strong effect of players with darker skin tones receiving more red cards (odds ratio > 1). Twenty of the 29 teams found a statistically-significant effect in the predicted direction of players with darker skin tones being issued more red cards. This degree of variation in peer-reviewed analyses from identical data is striking, but the generality of this finding has only just begun to be formally investigated (e.g., Huntington-Klein et al. 2021; Schweinsberg et al. 2021; Breznau et al. 2022; Coretta et al. 2023).

In the neuroimaging study, 70 teams evaluated each of the nine different hypotheses with the available fMRI data (Botvinik-Nezer et al. 2020). These 70 teams followed a divergent set of workflows that produced a wide range of results. The rate of reporting of statistically significant support for the nine hypotheses ranged from 21% to 84%, and for each hypothesis on average, 20% of research teams observed effects that differed substantially from the majority of other teams. Some of the variability in results among studies could be explained by analytical decisions such as choice of software package, smoothing function, and parametric versus non-parametric corrections for multiple comparisons. However, substantial variability among analyses remained unexplained, and presumably emerged from the many different decisions each analyst made in their long workflows. Such variability in results among analyses from this dataset and from the very different red-card dataset suggests that sensitivity of analytical outcome to analytical choices may characterize many distinct fields, as several more recent many-analyst studies also suggest (Huntington-Klein et al. 2021; Schweinsberg et al. 2021; Breznau et al. 2022).

To further develop the empirical understanding of the effects of analytical decisions on study outcomes, we chose to estimate the extent to which researchers' data analysis choices drive differences in effect sizes, model predictions, and qualitative conclusions in ecology and evolutionary biology. This is an important extension of the meta-research agenda of evaluating factors influencing replicability in ecology, evolutionary biology, and beyond (Fidler et al. 2017). To examine the effects of analytical decisions, we used two different datasets and recruited researchers to analyze one or the other of these datasets to answer a question we defined. The first question was "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?" To answer this question, we provided a dataset that includes brood size manipulations from 332 broods conducted over three years at Wytham Wood, UK. The second question was "How does grass cover influence *Eucalyptus* spp. seedling recruitment?" For this question, analysts used a dataset that includes, among other variables, number of seedlings in different size classes, percentage cover of different life forms, tree canopy cover, and distance from canopy edge from 351 quadrats spread among 18 sites in Victoria, Australia.

We explored the impacts of data analysts' choices with descriptive statistics and with a series of tests to attempt to explain the variation among effect sizes and predicted values of the dependent variable produced by the different analysis teams for both datasets separately. To describe the variability, we present forest plots of the standardized effect sizes and predicted values produced by each of the analysis teams, estimate heterogeneity (both absolute, τ^2, and proportional, I^2) in effect size and predicted values among the results produced by these different teams, and calculate a similarity index that quantifies variability among the predictor variables selected for the different statistical models constructed by the different analysis teams. These descriptive statistics provide the first estimates of the extent to which explanatory statistical models and their outcomes in ecology and evolutionary biology vary based on the decisions of different data analysts. We then quantified the degree to which the variability in effect size and predicted values could be explained by (1) variation in the quality of analyses as rated by peer reviewers and (2) the similarity of the choices of predictor variables between individual analyses.

Methods

This project involved a series of steps (1-6) that began with identifying datasets for analyses and continued through recruiting independent groups of scientists to analyze the data, allowing the scientists to analyze the data as they saw fit, generating peer review ratings of the analyses (based on methods, not results), evaluating the variation in effects among the different analyses, and producing the final manuscript.

Step 1: Select datasets

We used two previously unpublished datasets, one from evolutionary ecology and the other from ecology and conservation.

Evolutionary ecology

Our evolutionary ecology dataset is relevant to a sub-discipline of life-history research which focuses on identifying costs and trade-offs associated with different phenotypic conditions. These data were derived from a brood-size manipulation experiment imposed on wild birds nesting in boxes provided by researchers in an intensively studied population. Understanding how the growth of nestlings is influenced by the numbers of siblings in the nest can give researchers insights into factors such as the evolution of clutch size, determination of provisioning rates by parents, and optimal levels of sibling competition (Vander Werf 1992; DeKogel 1997; Royle et al. 1999; Verhulst, Holveck, and Riebel 2006; Nicolaus et al. 2009). Data analysts were provided this dataset and instructed to answer the following question: "To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?"

Researchers conducted brood size manipulations and population monitoring of blue tits at Wytham Wood, a 380 ha woodland in Oxfordshire, U.K (1º 20'W, 51º 47'N). Researchers regularly checked approximately 1100 artificial nest boxes at the site and monitored the 330 to 450 blue tit pairs occupying those boxes in 2001-2003 during the experiment. Nearly all birds made only one breeding attempt during the April to June study period in a given year. At each blue tit nest, researchers recorded the date the first egg appeared, clutch size, and hatching date. For all chicks alive at age 14 days, researchers measured mass and tarsus length and fitted a uniquely numbered, British Trust for Ornithology (BTO) aluminium leg ring. Researchers attempted to capture all adults at their nests between day 6 and day 14 of the chick-rearing period. For these captured adults, researchers measured mass, tarsus length, and wing length and fitted a uniquely numbered BTO leg ring. During the 2001-2003 breeding seasons, researchers manipulated brood sizes using cross fostering. They matched broods for hatching date and brood size and moved chicks between these paired nests one or two days after hatching. They sought to either enlarge or reduce all manipulated broods by approximately one fourth. To control for effects of being moved, each reduced brood had a portion of its brood replaced by chicks from the paired increased brood, and vice versa. Net manipulations varied from plus or minus four chicks in broods of 12 to 16 to plus or minus one chick in broods of 4 or 5. Researchers left approximately one third of all broods unmanipulated. These unmanipulated broods were not selected systematically to match manipulated broods in clutch size or laying date. We have mass and tarsus length data from 3720 individual chicks divided among 167 experimentally enlarged broods, 165 experimentally reduced broods, and 120 unmanipulated broods. The full list of variables included in the dataset is publicly available (https://osf.io/hdv8m), along with the data (https://osf.io/qjzby).

Ecology and conservation

Additional Explanation:

Shortly after beginning to recruit analysts, several analysts noted a small set of related errors in the blue tit dataset. We corrected the errors, replaced the dataset on our OSF site, and emailed the analysts on 19 April 2020 to instruct them to use the revised data. The email to analysts is available here (https://osf.io/4h53z). The errors are explained in that email.

Our ecology and conservation dataset is relevant to a sub-discipline of conservation research which focuses on investigating how best to revegetate private land in agricultural landscapes. These data were collected on private land under the Bush Returns program, an incentive system where participants entered into a contract with the Goulburn Broken Catchment Management Authority and received annual payments if they executed predetermined restoration activities. This particular dataset is based on a passive regeneration initiative, where livestock grazing was removed from the property in the hopes that the *Eucalyptus* spp. overstorey would regenerate without active (and expensive) planting. Analyses of some related data have been published (Miles 2008; Vesk et al. 2016) but those analyses do not address the question analysts answered in our study. Data analysts were provided this dataset and instructed to answer the following question: "How does grass cover influence *Eucalyptus* spp. seedling recruitment?".

Researchers conducted three rounds of surveys at 18 sites across the Goulburn Broken catchment in northern Victoria, Australia in winter and spring 2006 and autumn 2007. In each survey period, a different set of 15 x 15 m quadrats were randomly allocated across each site within 60 m of existing tree canopies. The number of quadrats at each site depended on the size of the site, ranging from four at smaller sites to 11 at larger sites. The total number of quadrats surveyed across all sites and seasons was 351. The number of *Eucalyptus* spp. seedlings was recorded in each quadrat along with information on the GPS location, aspect, tree canopy cover, distance to tree canopy, and position in the landscape. Ground layer plant species composition was recorded in three 0.5 x 0.5 m sub-quadrats within each quadrat. Subjective cover estimates of each species as well as bare ground, litter, rock and moss/lichen/soil crusts were recorded. Subsequently, this was augmented with information about the precipitation and solar radiation at each GPS location. The full list of variables included in the dataset is publicly available (https://osf.io/r5gbn), along with the data (https://osf.io/qz5cu).

Step 2: Recruitment and initial survey of analysts

The lead team (TP, HF, SN, EG, SG, PV, DH, FF) created a publicly available document providing a general description of the project (https://osf.io/mn5aj/). The project was advertised at conferences, via Twitter, using mailing lists for ecological societies (including *Ecolog*, *Evoldir*, and lists for *the Environmental Decisions Group*, and *Transparency in Ecology and Evolution*), and via word of mouth. The target population was active ecology, conservation, or evolutionary biology researchers with a graduate degree (or currently studying for a graduate degree) in a relevant discipline. Researchers could choose to work independently or in a small team. For the sake of simplicity, we refer to these as 'analysis teams' though some comprised one individual. We aimed for a minimum of 12 analysis teams independently evaluating each dataset (see sample size justification below). We simultaneously recruited volunteers to peer review the analyses conducted by the other volunteers through the same channels. Our goal was to recruit a similar number of peer reviewers and analysts, and to ask each peer reviewer to review a minimum of four analyses. If we were unable to recruit at least half the number of reviewers as analysis teams, we planned to ask analysts to serve also as reviewers (after they had completed their analyses), but this was unnecessary. Therefore, no data analysts peer reviewed analyses of the dataset they had analyzed. All analysts and reviewers were offered the opportunity to share co-authorship on this manuscript and we planned to invite them to participate in the collaborative process of producing the final manuscript. All analysts signed [digitally] a consent (ethics) document (https://osf.io/xyp68/) approved by the Whitman College Institutional Review Board prior to being allowed to participate.

> **Preregistration Deviation:**
>
> Due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We identified our minimum number of analysts per dataset by considering the number of effects needed in a meta-analysis to generate an estimate of heterogeneity (τ^2) with a 95% confidence interval that does not encompass zero. This minimum sample size is invariant regardless of τ^2. This is because the same t-statistic value will be obtained by the same sample size regardless of variance

(τ^2). We see this by first examining the formula for the standard error, SE for variance, (τ^2) or ($\mathrm{SE}\tau^2$) assuming normality in an underlying distribution of effect sizes (Knight 2000):

$$\mathrm{SE}(\tau^2) = \sqrt{\frac{2\tau^4}{n-1}}$$

and then rearranging the above formula to show how the t-statistic is independent of τ^2, as seen below.

$$t = \frac{\tau^2}{\mathrm{SE}(\tau^2)} = \sqrt{\frac{n-1}{2}}$$

We then find a minimum $n = 12$ according to this formula.

Step 3: Primary data analyses

Analysis teams registered and answered a demographic and expertise survey (https://osf.io/seqzy/). We then provided them with the dataset of their choice and requested that they answer a specific research question. For the evolutionary ecology dataset that question was “To what extent is the growth of nestling blue tits (*Cyanistes caeruleus*) influenced by competition with siblings?” and for the conservation ecology dataset it was “How does grass cover influence *Eucalyptus* spp. seedling recruitment?” Once their analysis was complete, they answered a structured survey (https://osf.io/neyc7/), providing analysis technique, explanations of their analytical choices, quantitative results, and a statement describing their conclusions. They also were asked to upload their analysis files (including the dataset as they formatted it for analysis and their analysis code [if applicable]) and a detailed journal-ready statistical methods section.

Additional Information:

As is common in many studies in ecology and evolutionary biology, the datasets we provided contained many variables, and the research questions we provided could be addressed by our datasets in many different ways. For instance, volunteer analysts had to choose the dependent (response) variable and the independent variable, and make numerous other decisions about which variables and data to use and how to structure their model.

Preregistration Deviation:

We originally planned to have analysts complete a single survey (https://osf.io/neyc7/), but after we evaluated the results of that survey, we realized we would need a second survey (https://osf.io/8w3v5/) to adequately collect the information we needed to evaluate heterogeneity of results (step 5). We provided a set of detailed instructions with the follow-up survey, and these instructions are publicly available and can be found within the following files (blue tit: https://osf.io/kr2g9, *Eucalyptus*: https://osf.io/dfvym).

Step 4: Peer reviews of analyses

At minimum, each analysis was evaluated by four different reviewers, and each volunteer peer reviewer was randomly assigned methods sections from at least four analyst teams (the exact number varied). Each peer reviewer registered and answered a demographic and expertise survey identical to that asked of the analysts, except we did not ask about 'team name' since reviewers did not work in teams. Reviewers evaluated the methods of each of their assigned analyses one at a time in a sequence determined by the project leaders. We systematically assigned the sequence so that, if possible, each analysis was allocated to each position in the sequence for at least one reviewer. For instance, if each reviewer were assigned four analyses to review, then each analysis would be the first analysis assigned to at least one reviewer, the second analysis assigned to another reviewer, the third analysis assigned to yet another reviewer, and the fourth analysis assigned to a fourth reviewer. Balancing the order in which reviewers saw the analyses controls for order effects, e.g. a reviewer might be less critical of the first methods section they read than the last.

The process for a single reviewer was as follows. First, the reviewer received a description of the methods of a single analysis. This included the narrative methods section, the analysis team's answers to our survey questions regarding their methods, including analysis code, and the dataset. The reviewer was then asked, in an online survey (https://osf.io/4t36u/), to rate that analysis on a scale of 0-100 based on this prompt: "Rate the overall appropriateness of this analysis to answer the research question (*one of the two research questions inserted here*) with the available data. To help you calibrate your rating, please consider the following guidelines:

- 100. A perfect analysis with no conceivable improvements from the reviewer
- 75. An imperfect analysis but the needed changes are unlikely to dramatically alter outcomes
- 50. A flawed analysis likely to produce either an unreliable estimate of the relationship or an over-precise estimate of uncertainty
- 25. A flawed analysis likely to produce an unreliable estimate of the relationship and an over-precise estimate of uncertainty
- 0. A dangerously misleading analysis, certain to produce both an estimate that is wrong and a substantially over-precise estimate of uncertainty that places undue confidence in the incorrect estimate.

*Please note that these values are meant to calibrate your ratings. We welcome ratings of any number between 0 and 100.

After providing this rating, the reviewer was presented with this prompt, in multiple-choice format: "Would the analytical methods presented produce an analysis that is (a) publishable as is, (b) publishable with minor revision, (c) publishable with major revision, (d) deeply flawed and unpublishable?" The reviewer was then provided with a series of text boxes and the following prompts: "Please explain your ratings of this analysis. Please evaluate the choice of statistical analysis type. Please evaluate the process of choosing variables for and structuring the statistical model. Please evaluate the suitability of the variables included in (or excluded from) the statistical model. Please evaluate the suitability of the structure of the statistical model. Please evaluate choices to exclude or not exclude subsets of the data. Please evaluate any choices to transform data (or, if there were no transformations, but you think there should have been, please discuss that choice)." After submitting this review, a methods section from a second analysis was then made available to the reviewer. This same sequence was followed until all analyses allocated to a given reviewer were provided and reviewed. After providing the final review, the reviewer was simultaneously provided with all four (or more) methods sections the reviewer had just completed reviewing, the option to revise their original ratings, and a text box to provide an explanation. The invitation to revise the original ratings was as follows: “If, now that you have seen all the analyses you are reviewing, you wish to revise your ratings of any of these analyses, you may do so now.” The text box was prefaced with this prompt: “Please explain your choice to revise (or not to revise) your ratings.”

Additional Information: unregistered analysis

To determine how consistent peer reviewers were in their ratings, we assessed inter-rater reliability among reviewers for both the categorical and quantitative ratings combining blue tit and *Eucalyptus* data using Krippendorff’s alpha for ordinal and continuous data respectively. This provides a value that is between -1 (total disagreement between reviewers) and 1 (total agreement between reviewers).

Step 5: Evaluate variation

Additional Information: analysis schematic

The lead team conducted a range of preregistered and exploratory analyses to understand variation between analyses and their results. Figure 1 is intended to clarify the analyses described below.

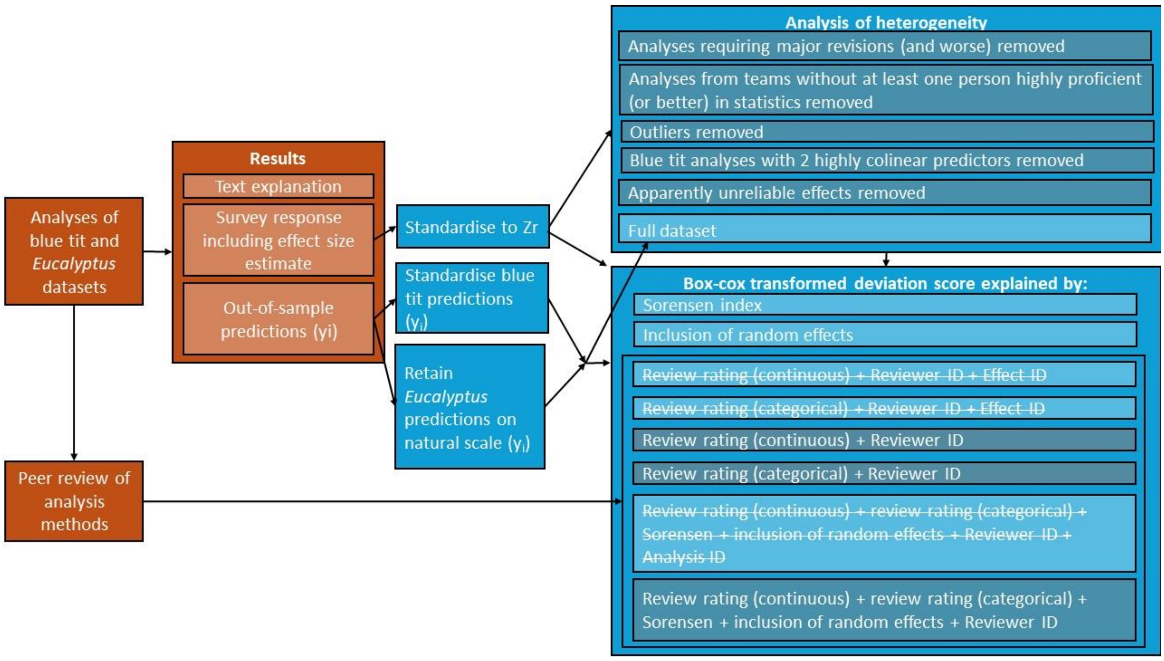

Figure 1: Schematic of research process showing recruited analyst and reviewer contributions in orange and core team contributions in blue. Items that are crossed out were preregistered but could not be conducted. Items with a greyed background were added as exploratory analyses after preregistration.

The lead team conducted the analyses outlined in this section. We described the variation in model specification in several ways. We calculated summary statistics describing variation among analyses, including mean, SD, and range of number of variables per model included as fixed effects, the number of interaction terms, the number of random effects, and the mean, SD, and range of sample sizes. We also present the number of analyses in which each variable was included. We summarized the variability in standardized effect sizes and predicted values of dependent variables among the individual analyses using standard random effects meta-analytic techniques. First, we derived standardized effect sizes from each individual analysis. We did this for all linear models or generalized linear models by converting the *t* value and the degree of freedom (*df*) associated with regression coefficients (e.g. the effect of the number of siblings [predictor] on growth [response] or the effect of grass cover [predictor] on seedling recruitment [response]) to the correlation coefficient, *r,* using the following:

$$r = \frac{t^2}{(t^2 + df)}$$

This formula can only be applied if *t* and *df* values originate from linear or generalized linear models [GLMs; Nakagawa and Cuthill (2007)]. If, instead, linear mixed-effects models (LMMs) or generalized linear mixed-effects models (GLMMs) were used by a given analysis, the exact *df* cannot be estimated. However, adjusted *df* can be estimated, for example, using the Satterthwaite approximation of *df*, df_S, [note that SAS uses this approximation to obtain *df* for LMMs and GLMMs; Luke (2017)]. For analyses using either LMMs or GLMMs that do not produce df_S we planned to obtain df_S by rerunning the same (G)LMMs using the lmer() or glmer() function in the *lmerTest* package in R (Kuznetsova, Brockhoff, and Christensen 2017; R Core Team 2024).

Preregistration Deviation:

Rather than re-run these analyses ourselves, we sent a follow-up survey (referenced above under "Primary data analyses") to analysts and asked them to follow our instructions for producing this information. The instructions are publicly available and can be found within the following files (blue tit: https://osf.io/kr2g9, *Eucalyptus*: https://osf.io/dfvym).

We then used the *t* values and df_S from the models to obtain *r* as per the formula above. All *r* and accompanying *df* (or df_S) were converted to Fisher's Z_r.

$$Z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

and its sampling variance; 1/(n–3) where n=*df*+1. Any analyses from which we could not derive a signed Z_r, for instance one with a quadratic function in which the slope changed sign, were considered unusable for analyses of Z_r. We expected such analyses would be rare. In fact, most submitted analyses excluded from our meta-analysis of Z_r were excluded because of a lack of sufficient information provided by the analyst team rather than due to the use of effects that could not be converted to Z_r. Regardless, as we describe below, we generated a second set of standardized effects (predicted values) that could (in principle) be derived from any explanatory model produced by these data.

Besides Z_r, which describes the strength of a relationship based on the amount of variation in a dependent variable explained by variation in an independent variable, we also examined differences in the shape of the relationship between the independent and dependent variables. To accomplish this, we derived a point estimate (out-of-sample predicted value) for the dependent variable of interest for each of three values of our primary independent variable. We originally described these three values as associated with the 25th percentile, median, and 75th percentile of the independent variable and any covariates.

Preregistration Deviation:

The original description of the out-of-sample specifications did not account for the facts that (a) some variables are not distributed in a way that allowed division in percentiles and that (b) variables could be either positively or negatively correlated with the dependent variable. We provide a more thorough description here:

We derived three point-estimates (out-of-sample predicted values) for the dependent variable of interest; one for each of three values of our primary independent variable that we specified. We also specified values for all other variables that could have been included as independent variables in analysts' models so that we could derive the predicted values from a fully specified version of any model produced by analysts. For all potential independent variables, we selected three values or categories. Of the three we selected, one was associated with small, one with intermediate, and one with large values of one typical dependent variable (day 14 chick weight for the blue tit data and total number of seedlings for the *Eucalyptus* data; analysts could select other variables as their dependent variable, but the others typically correlated with the two identified here). For continuous variables, this means we identified the 25th percentile, median, and 75th percentile and, if the slope of the linear relationship between this variable and the typical dependent variable was positive, we left the quartiles ordered as is. If, instead, the slope was negative, we reversed the order of the independent variable quartiles so that the 'lower' quartile value was the one associated with the lower value for the dependent variable. In the case of categorical variables, we identified categories associated with the 25th percentile, median, and 75th percentile values of the typical dependent variable after averaging the values for each category. However, for some continuous and categorical predictors, we also made selections based on the principle of internal consistency between certain related variables, and we fixed a few categorical variables as identical across all three levels where doing so would simplify the modelling process (specification tables available: blue tit: https://osf.io/86akx; *Eucalyptus*: https://osf.io/jh7g5).

We used the 25th and 75th percentiles rather than minimum and maximum values to reduce the chance of occupying unrealistic parameter space. We planned to derive these predicted values from the model information provided by the individual analysts. All values (predictions) were first transformed to the original scale along with their standard errors (SE); we used the delta method (Ver Hoef 2012) for the transformation of SE. We used the square of the SE associated with predicted values as the sampling variance in the meta-analyses described below, and we planned to analyze these predicted values in exactly the same ways as we analyzed Z_r in the following analyses.

We plotted individual effect size estimates (Z_r) and predicted values of the dependent variable (y_i) and their corresponding 95% confidence / credible intervals in forest plots to allow visualization of

Preregistration Deviation:

1. Standardizing blue tit out-of-sample predictions (y_i)

Because analysts of blue tit data chose different dependent variables on different scales, after transforming out-of-sample values to the original scales, we standardized all values as *z* scores ('standard scores') to put all dependent variables on the same scale and make them comparable. This involved taking each relevant value on the original scale (whether a predicted point estimate or a SE associated with that estimate) and subtracting the value in question from the mean value of that dependent variable derived from the full dataset and then dividing this difference by the standard deviation, SD, corresponding to the mean from the full dataset (Supplementary Material B, Equation B.1).

Note that we were unable to standardise some analyst-constructed variables, so these analyses were excluded from the final out-of-sample estimates meta-analysis, see Supplementary Material B, section B.1.2.1 for details and explanation.

2. Log-transforming *Eucalyptus* out-of-sample predictions *yi*

All analyses of the *Eucalyptus* data chose dependent variables that were on the same scale, that is, *Eucalyptus* seedling counts. Although analysts may have used different size-classes of *Eucalyptus* seedlings for their dependent variable, we considered these choices to be akin to subsetting, rather than as different response variables, since changing the size-class of the dependent variable ultimately results in observations being omitted or included. Consequently, we did not standardise *Eucalyptus* out-of-sample predictions.

We were unable to fit quasi-Poisson or Poisson meta-regressions, as desired (O'Hara and Kotze 2010), because available meta-analysis packages (e.g. metafor:: and metainc::) do not provide implementation for outcomes as estimates-only, methods are only provided for outcomes as ratios or rate-differences between two groups. Consequently, we log-transformed the out-of-sample predictions for the *Eucalyptus* data and use the mean estimate for each prediction scenario as the dependent variable in our meta-analysis with the associated SE as the sampling variance in the meta-analysis (Nakagawa et al. 2023, Table 2).

the range and precision of effect size and predicted values. Further, we included these estimates in random effects meta-analyses (Higgins et al. 2003; Borenstein et al. 2017) using the *metafor* package in R (Viechtbauer 2010; R Core Team 2024):

$$Z_r \sim 1 + (1|Effect\ ID)$$

$$y_i \sim 1 + (1|Effect\ ID)$$

where y_i is the predicted value for the dependent variable at the 25th percentile, median, or 75th percentile of the independent variables. The individual Z_r effect sizes were weighted with the inverse of sampling variance for Z_r. The individual predicted values for dependent variable (y_i) were weighted by the inverse of the associated SE^2 (original registration omitted "inverse of the" in error). These analyses provided an average Z_r score ($\bar{Z}_r$) or an average y_i ($\bar{y}_i$) with corresponding 95% confidence interval and allowed us to estimate two heterogeneity indices, τ^2 and I^2. The former, τ^2, is the absolute measure of heterogeneity or the between-study variance (in our case, between-effect variance) whereas I^2 is a relative measure of heterogeneity. We obtained the estimate of relative heterogeneity (I^2) by dividing the between-effect variance by the sum of between-effect and within-effect variance (sampling error variance). I^2 is thus, in a standard meta-analysis, the proportion of variance that is due to heterogeneity as opposed to sampling error. When calculating I^2, within-study variance is amalgamated across studies to create a "typical" within-study variance which serves as the sampling error variance (Higgins et al. 2003; Borenstein et al. 2017). Our goal here was to visualize and quantify the degree of variation among analyses in effect size estimates (Nakagawa and Cuthill 2007). We did not test for statistical significance.

Additional explanation:

Our use of I^2 to quantify heterogeneity violates an important assumption, but this violation does not invalidate our use of I^2 as a metric of how much heterogeneity can derive from analytical decisions. In standard meta-analysis, the statistic I^2 quantifies the proportion of variance that is greater than we would expect if differences among estimates were due to sampling error alone (Rosenberg 2013). However, it is clear that this interpretation does not apply to our value of I^2 because I^2 assumes that each estimate is based on an independent sample (although these analyses can account for non-independence via hierarchical modelling), whereas all our effects were derived from largely or entirely overlapping subsets of the same dataset. Despite this, we believe that I^2 remains a useful statistic for our purposes. This is because, in calculating I^2, we are still setting a benchmark of expected variation due to sampling error based on the variance associated with each separate effect size estimate, and we are assessing how much (if at all) the variability among our effect sizes exceeds what would be expected had our effect sizes been based on independent data. In other words, our estimates can tell us how much proportional heterogeneity is possible from analytical decisions alone when sample sizes (and therefore meta-analytic within-estimate variance) are similar to the ones in our analyses. Among other implications, our violation of the independent sample assumption means that we (dramatically) over-estimate the variance expected due to sampling error, and because I^2 is a proportional estimate, we thus underestimate the actual proportion of variance due to differences among analyses other than sampling error. However, correcting this underestimation would create a trivial value since we designed the study so that much of the variance would derive from analytic decisions as opposed to differences in sampled data. Instead, retaining the I^2 value as typically calculated provides a useful comparison to I^2 values from typical meta-analyses.

Interpretation of τ^2 also differs somewhat from traditional meta-analysis, and we discuss this further in the Results.

Finally, we assessed the extent to which deviations from the meta-analytic mean by individual effect sizes (Z_r) or the predicted values of the dependent variable (y_i) were explained by the peer rating of each analysis team's method section, by a measurement of the distinctiveness of the set of predictor variables included in each analysis, and by the choice of whether or not to include random effects in the model. The deviation score, which served as the dependent variable in these analyses, is the absolute value of the difference between the meta-analytic mean $\bar{Z}_r$ (or $\bar{y}_i$) and the individual Z_r (or y_i) estimate for each analysis. We used the Box-Cox transformation on the absolute values of deviation scores to achieve an approximately normal distribution (c.f. Fanelli and Ioannidis 2013; Fanelli, Costas, and Ioannidis 2017). We described variation in this dependent variable with both a series of univariate analyses and a multivariate analysis. All these analyses were general linear (mixed) models. These analyses were secondary to our estimation of variation in effect sizes described above. We wished to quantify relationships among variables, but we had no *a priori* expectation of effect size and made no dichotomous decisions about statistical significance.

When examining the extent to which reviewer ratings (on a scale from 0 to 100) explained deviation from the average effect (or predicted value), each analysis had been rated by multiple peer reviewers, so for each reviewer score to be included, we include each deviation score in the analysis multiple times. To account for the non-independence of multiple ratings of the same analysis, we planned to include analysis identity as a random effect in our general linear mixed model in the *lme4* package in R (Bates et al. 2015; R Core Team 2024). To account for potential differences among reviewers in their scoring of analyses, we also planned to include reviewer identity as a

Additional explanation:

In our meta-analyses based on Box-Cox transformed deviation scores, we leave these deviation scores unweighted. This is consistent with our registration, which did not mention weighting these scores. However, the fact that we did not mention weighting the scores was actually an error: we had intended to weight them, as is standard in meta-analysis, using the inverse variance of the Box-Cox transformed deviation scores Supplementary Material C, equation C.1. Unfortunately, when we did conduct the weighted analyses, they produced results in which some weighted estimates differed radically from the unweighted estimate because the weights were invalid. Such invalid weights can sometimes occur when the variance (upon which the weights depend) is partly a function of the effect size, as in our Box-Cox transformed deviation scores (Nakagawa et al. 2022). In the case of the *Eucalyptus* analyses, the most extreme outlier was weighted much more heavily (by close to two orders of magnitude) than any other effect sizes because the effect size was, itself, so high. Therefore, we made the decision to avoid weighting by inverse variance in all analyses of the Box-Cox transformed deviation scores. This was further justified because (a) most analyses have at least some moderately unreliable weights, and (b) the sample sizes were mostly very similar to each other across submitted analyses, and so meta-analytic weights are not particularly important here (Buck et al. 2022). We systematically investigated the impact of different weighting schemes and random effects on model convergence and results, see Supplementary Material C, section C.8 for more details.

random effect:

$$\mathrm{DeviationScore}_j = BoxCox\left(\mathrm{DeviationFromMean}_j\right)$$

$$\mathrm{DeviationScore}_{ij} \sim \mathrm{Rating}_{ij} + \mathrm{ReviewerID}_i + EffectID_j$$

$$\mathrm{ReviewerID}_i \sim N(0, \sigma_i^2)$$

$$EffectID \sim N(0, \sigma_j^2)$$

Where $\mathrm{DeviationFromMean}_j$ is the deviation from the meta-analytic mean for the *j*th analysis, Reviewer ID_i is the random intercept assigned to each *i* reviewer, and Effect ID_j is the random intercept assigned to each *j* analysis, both of which are assumed to be normally distributed with a mean of 0 and a variance of σ^2. Absolute deviation scores were Box-Cox transformed using the step_box_cox() function from the *timetk* package in R (Dancho and Vaughan 2023; R Core Team 2024).

We conducted a similar analysis with the four categories of reviewer ratings ((1) deeply flawed and unpublishable, (2) publishable with major revision, (3) publishable with minor revision, (4) publishable as is) set as ordinal predictors numbered as shown here. As with the analyses above, we planned for these analyses to also include random effects of analysis identity and reviewer identity. Both of these analyses (1: 1-100 ratings as the fixed effect, 2: categorical ratings as the fixed effects) were planned to be conducted eight times for each dataset. Each of the four responses (Z_r, y_{25}, y_{50}, y_{75}) were to be compared once to the initial ratings provided by the peer reviewers, and again based on the revised ratings provided by the peer reviewers.

Preregistration deviation:

1. We planned to include random effects of both analysis identity and reviewer identity in these models comparing reviewer ratings with deviation scores. However, after we received the analyses, we discovered that a subset of analyst teams had either conducted multiple analyses and/or identified multiple effects per analysis as answering the target question. We therefore faced an even more complex potential set of random effects. We decided that including Team ID and Effect ID along with Reviewer ID as random effects in the same model would almost certainly lead to model fit problems, and so we started with simpler models including just Effect ID and Reviewer ID. However, even with this simpler structure, our dataset was sparse, with reviewers rating a small number of analyses, resulting in models with singular fit (Supplementary Material C, section C.2). Removing one of the random effects was necessary for the models to converge. For both models of deviation from the meta-analytic mean explained by categorical or continuous reviewer ratings, we removed the random effect of Effect ID, leaving Reviewer ID as the only random effect.
2. We conducted analyses only with the final peer ratings after the opportunity for revision, not with the initial ratings. This was because when we recorded the final ratings, the initial ratings were over-written, therefore we did not have access to those initial values.

The next set of univariate analyses sought to explain deviations from the mean effects based on a measure of the distinctiveness of the set of variables included in each analysis. As a 'distinctiveness' score, we used Sorensen's Similarity Index (an index typically used to compare species composition across sites), treating variables as species and individual analyses as sites. To generate an individual Sorensen's value for each analysis required calculating the pairwise Sorensen's value for all pairs of analyses (of the same dataset), and then taking the average across these Sorensen's values for each analysis. We calculated the Sorensen's index values using the *betapart* package (Baselga et al. 2023) in R:

$$\beta Sorensen = \frac{b + c}{2a + b + c}$$

where a is the number of variables common to both analyses, b is the number of variables that occur in the first analysis but not in the second and c is the number of variables that occur in the second analysis. We then used the per-model average Sorensen's index value as an independent variable to predict the deviation score in a general linear model, and included no random effect since each analysis is included only once, in R (R Core Team 2024):

$$DeviationScore_j \sim \beta Sorensen_j$$

Additional explanation:

When we planned this analysis, we anticipated that analysts would identify a single primary effect from each model, so that each model would appear in the analysis only once. Our expectation was incorrect because some analysts identified >1 effect per analysis, but we still chose to specify our model as registered and not use a random effect. This is because most models produced only one effect and so we expected that specifying a random effect to account for the few cases where >1 effect was included for a given model would prevent model convergence.

Note that this analysis contrasts with the analyses in which we used reviewer ratings as predictors because in the analyses with reviewer ratings, each effect appeared in the analysis approximately four times due to multiple reviews of each analysis, and so it was much more important to account for that variance through a random effect.

Next, we assessed the relationship between the inclusion of random effects in the analysis and the deviation from the mean effect size. We anticipated that most analysts would use random effects in a mixed model framework, but if we were wrong, we wanted to evaluate the differences in outcomes when using random effects versus not using random effects. Thus, if there were at least 5 analyses that did and 5 analyses that did not include random effects, we would add a binary predictor variable "random effects included (yes/no)" to our set of univariate analyses and would add this predictor variable to our multivariate model described below. This standard was only met for the *Eucalyptus* analyses, and so we only examined inclusion of random effects as a predictor variable in meta-analysis of this set to analyses.

Finally, we conducted a multivariate analysis with the five predictors described above (peer ratings 0-100 and peer ratings of publishability 1-4; both original and revised and Sorensen's index, plus a sixth for *Eucalyptus*, presence / absence of random effects) with random effects of analysis identity and reviewer identity in the *lme4* package in R (Bates et al. 2015; R Core Team 2024). We had stated here in the text that we would use only the revised (final) peer ratings in this analysis, so the absence of the initial ratings is not a deviation from our plan:

$$\mathrm{DeviationScore}_j = BoxCox\left(\mathrm{DeviationFromMean}_j\right)$$

$$\mathrm{DeviationScore}_{ij} \sim \mathrm{RatingContinuous}_{ij} + \mathrm{RatingCategorical}_{ij} + \beta\mathrm{Sorensen}_j + \mathrm{ReviewerID}_i + \mathrm{Effect\ ID}_j$$

$$\mathrm{ReviewerID}_i \sim N(0, \sigma_i^2)$$

$$\mathrm{EffectID}_j \sim N(0, \sigma_j^2)$$

We conducted all the analyses described above eight times; for each of the four responses (Z_r, y_{25}, y_{50}, y_{75}) one time for each of the two datasets.

We have publicly archived all relevant data, code, and materials on the Open Science Framework (https://osf.io/mn5aj/). Archived data includes the original datasets distributed to all analysts, any edited versions of the data analyzed by individual groups, and the data we analyzed with our meta-analyses, which include the effect sizes derived from separate analyses, the statistics describing variation in model structure among analyst groups, and the anonymized answers to our surveys of analysts and peer reviewers. Similarly, we have archived both the analysis code used for each individual analysis (where available) and the code from our meta-analyses. We have also archived copies of our survey instruments from analysts and peer reviewers.

Our rules for excluding data from our study were as follows. We excluded from our synthesis any individual analysis submitted after we had completed peer review or those unaccompanied by analysis files that allow us to understand what the analysts did. We also excluded any individual analysis that did not produce an outcome that could be interpreted as an answer to our primary question (as posed above) for the respective dataset. For instance, this means that in the case of the data on blue tit chick growth, we excluded any analysis that did not include something that can be interpreted as growth or size as a dependent (response) variable, and in the case of the *Eucalyptus* establishment data, we excluded any analysis that did not include a measure of grass cover among the independent (predictor) variables. Also, as described above, any analysis that could not produce an effect that could be converted to a signed Z_r was excluded from analyses of Z_r.

Preregistration Deviation:

Some analysts had difficulty implementing our instructions to derive the out-of-sample predictions, and in some cases (especially for the *Eucalyptus* data), they submitted predictions with implausibly extreme values. We believed these values were incorrect and thus made the conservative decision to exclude out-of-sample predictions where the estimates were > 3 standard deviations from the mean value from the full dataset provided to teams for analysis.

Additional explanation: unregistered analyses

1. Evaluating model fit.

We evaluated all fitted models using the performance::performance() function from the *performance* package (Lüdecke, Ben-Shachar, et al. 2021) and the glance() function from the *broom.mixed* package (Bolker et al. 2024). For all models, we calculated the square root of the residual variance (Sigma) and the root mean squared error (RMSE). For GLMMs performance::performance() calculates the marginal and conditional R^2 values as well as the contribution of random effects (ICC), based on Nakagawa et al. (2017). The conditional R^2 accounts for both the fixed and random effects, while the marginal R^2 considers only the variance of the fixed effects. The contribution of random effects is obtained by subtracting the marginal R^2 from the conditional R^2.

2. Exploring outliers and analysis quality.

After seeing the forest plots of *Zr* values and noticing the existence of a small number of extreme outliers, especially from the *Eucalyptus* analyses, we wanted to understand the degree to which our heterogeneity estimates were influenced by these outliers. To explore this question, we removed the highest two and lowest two values of *Zr* in each dataset and re-calculated our heterogeneity estimates.

To help understand the possible role of the quality of analyses in driving the heterogeneity we observed among estimates of *Zr*, we created forest plots and recalculated our heterogeneity estimates after removing all effects from analysis teams that had received at least one rating of “deeply flawed and unpublishable” and then again after removing all effects from analysis teams with at least one rating of either “deeply flawed and unpublishable” or “publishable with major revisions”. We also used self-identified levels of statistical expertise to examine heterogeneity when we retained analyses only from analysis teams that contained at least one member who rated themselves as “highly proficient” or “expert” (rather than “novice” or “moderately proficient”) in conducting statistical analyses in their research area in our intake survey.

Additionally, to assess potential impacts of highly collinear predictor variables on estimates of *Zr* in blue tit analyses, we created forest plots (Supplementary Material B, Figure B.5) and recalculated our heterogeneity estimates after we removed analyses that contained the brood count after manipulation and the highly correlated (correlation of 0.89, Supplementary Material D, Figure D.2) brood count at day 14. This removal included the one effect based on a model that contained both these variables and a third highly correlated variable, the estimate of number of chicks fledged (the only model that included the estimate of number of chicks fledged). We did not conduct a similar analysis for the *Eucalyptus* dataset because there were no variables highly collinear with the primary predictors (grass cover variables) in that dataset (Supplementary Material D, Figure D.1).

3. Exploring possible impacts of lower quality estimates of degrees of freedom.

Our meta-analyses of variation in Z_r required variance estimates derived from estimates of the degrees of freedom in original analyses from which Z_r estimates were derived. While processing the estimates of degrees of freedom submitted by analysts, we identified a subset of these estimates in which we had lower confidence because two or more effects from the same analysis were submitted with identical degrees of freedom. We therefore conducted a second set of (more conservative) meta-analyses that excluded these Z_r estimates with identical estimates of degrees of freedom and we present these analyses in the supplement.

Additional explanation: Best practices in many-analysts research

After we initiated our project, a paper was published outlining best practices in many-analysts studies (Aczel et al. 2021). Although we did not have access to this document when we implemented our project, our study complies with these practices nearly completely. The one exception is that although we requested analysis code from analysts, we did not require submission of code.

Step 6: Facilitated Discussion and Collaborative Write-Up of Manuscript

We planned for analysts and initiating authors to discuss the limitations, results, and implications of the study and collaborate on writing the final manuscript for review as a stage-2 Registered Report.

Preregistration deviation:

As described above, due to the large number of recruited analysts and reviewers and the anticipated challenges of receiving and integrating feedback from so many authors, we limited analyst and reviewer participation in the production of the final manuscript to an invitation to call attention to serious problems with the manuscript draft.

We built an R package, ManyEcoEvo:: to conduct the analyses described in this study (Gould et al. 2023), which can be downloaded from https://github.com/egouldo/ManyEcoEvo/ to reproduce our analyses or replicate the analyses described here using alternate datasets. Data cleaning and preparation of analysis-data, as well as the analysis, is conducted in R (R Core Team 2024) reproducibly using the targets package (Landau 2021). This data and analysis pipeline is stored in the ManyEcoEvo:: package repository and its outputs are made available to users of the package when the library is loaded.

The full manuscript, including further analysis and presentation of results is written in Quarto (Allaire et al. 2024). The source code to reproduce the manuscript is hosted at https://github.com/egouldo/ManyAnalysts/ (Gould et al. 2024), and the rendered version of the source code may be viewed at https://egouldo.github.io/ManyAnalysts/. All R packages and their versions used in the production of the manuscript are listed in Table 7 at the end of this paper.

Results

Summary Statistics

In total, 173 analyst teams, comprising 246 analysts, contributed 182 usable analyses (compatible with our meta-analyses and provided with all information needed for inclusion) of the two datasets examined in this study which yielded 215 effects. Analysts produced 134 distinct effects that met our criteria for inclusion in at least one of our meta-analyses for the blue tit dataset. Analysts produced 81 distinct effects meeting our criteria for inclusion for the *Eucalyptus* dataset. Excluded analyses and effects either did not answer our specified biological questions, were submitted with insufficient information for inclusion in our meta-analyses, or were incompatible with production of our effect size(s). We expected cases of this final scenario (incompatible analyses), for instance we cannot extract a Z_r from random forest models, which is why we analyzed two distinct types of effects, Z_r and out-of-sample predictions. Some effects only provided sufficient information for a subset of analyses and were only included in that subset. For both datasets, most submitted analyses incorporated mixed effects. Submitted analyses of the blue tit dataset typically specified normal error and analyses of the *Eucalyptus* dataset typically specified a non-normal error distribution (Supplementary Material A, Table A.1).

For both datasets, the composition of models varied substantially in regards to the number of fixed and random effects, interaction terms, and the number of data points used, and these patterns differed somewhat between the blue tit and *Eucalyptus* analyses (See Supplementary Material A, Table A.2). Focusing on the models included in the Z_r analyses (because this is the larger sample), blue tit models included a similar number of fixed effects on average (mean 5.2 ± 2.92 SD, range: 1 to 19) as *Eucalyptus* models (mean 5.01 ± 3.83 SD, range: 1 to 13), but the standard deviation in number of fixed effects was somewhat larger in the *Eucalyptus* models. The average number of interaction terms was much larger for the blue tit models (mean 0.44 ± 1.11 SD, range: 0 to 10) than for the *Eucalyptus* models (mean 0.16 ± 0.65 SD, range: 0 to 5), but still under 0.5 for both, indicating that most models did not contain interaction terms. Blue tit models also contained more random effects (mean 3.53 ± 2.08 SD, range: 0 to 10) than *Eucalyptus* models (mean 1.41 ± 1.09 SD, range: 0 to 4). The maximum possible sample size in the blue tit dataset (3720 nestlings) was an order of magnitude larger than the maximum possible in the *Eucalyptus* dataset (351 plots), and the means and standard deviations of the sample size used to derive the effects eligible for our study were also an order of magnitude greater for the blue tit dataset (mean 2611.09 ± 937.48 SD, range: 76 to 76) relative to the *Eucalyptus* models (mean 298.43 ± 106.25 SD, range: 18 to 351). However, the standard deviation in sample size from the *Eucalyptus* models was heavily influenced by a few cases of dramatic sub-setting (described below). Approximately three quarters of *Eucalyptus* models used sample sizes within 3% of the maximum. In contrast, fewer than 20% of blue tit models relied on sample sizes within 3% of the maximum, and approximately 50% of blue tit models relied on sample sizes 29% or more below the maximum.

Analysts provided qualitative descriptions of the conclusions of their analyses. Each analysis team provided one conclusion per dataset. These conclusions could take into account the results of any formal analyses completed by the team as well as exploratory and visual analyses of the data. Here we summarize all qualitative responses, regardless of whether we had sufficient information to use the corresponding model results in our quantitative analyses below. We classified these conclusions into the categories summarized below (Table 1):

- *Mixed*: some evidence supporting a positive effect, some evidence supporting a negative effect
- *Conclusive negative*: negative relationship described without caveat
- *Qualified negative*: negative relationship but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive none*: analysts interpret the results as conclusive of no effect
- *Qualified none*: analysts describe finding no evidence of a relationship but they describe the potential for an undetected effect
- *Qualified positive*: positive relationship described but only in certain circumstances or where analysts express uncertainty in their result
- *Conclusive positive*: positive relationship described without caveat

For the blue tit dataset, most analysts concluded that there was negative relationship between measures of sibling competition and nestling growth, though half the teams expressed qualifications or described effects as mixed or absent. No analysts concluded that there was a positive relationship even though some individual effect sizes were positive, apparently because all analysts who produced effects indicating positive relationships also produced effects indicating negative relationships and therefore described their results as qualified, mixed, or absent. For the *Eucalyptus* dataset, there was a broader spread of conclusions with at least one analyst team providing conclusions consistent with each conclusion category. The most common conclusion for the *Eucalyptus* dataset was that there was no relationship between grass cover and *Eucalyptus* recruitment (either conclusive or qualified description of no relationship), but more than half the teams concluded that there were effects; negative, positive, or mixed.

Table 1: Tallies of analysts' qualitative answers to the research questions addressed by their analyses.

Dataset	Mixed	Negative Conclusive	Negative Qualified	None Conclusive	None Qualified	Positive Qualified	Positive Conclusive
blue tit	5	37	27	4	1	0	0
Eucalyptus	8	6	12	19	12	4	2

Distribution of effects

Effect sizes (Z_r)

Although the majority (118 of 131) of the usable *Zr* effects from the blue tit dataset found nestling growth decreased with sibling competition, and the meta-analytic mean $\bar{Z}_r$ (Fisher's transformation of the correlation coefficient) was convincingly negative (-0.35 ± 0.06 95%CI), there was substantial variability in the strength and the direction of this effect. Z_r ranged from -1.55 to 0.38, and approximately continuously from -0.93 to 0.19 (Figure 2a and Table 4), and of the 118 effects with negative slopes, 93 had confidence intervals excluding 0. Of the 13 with positive slopes indicating increased nestling growth in the presence of more siblings, 2 had confidence intervals excluding zero (Figure 2a).

Meta-analysis of the *Eucalyptus* dataset also showed substantial variability in the strength of effects as measured by Z_r, and unlike with the blue tits, a notable lack of consistency in the direction of effects (Figure 2b, Table 4). Z_r ranged from -4.47 (Supplementary Material A, Figure A.2), indicating a strong tendency for reduced *Eucalyptus* seedling success as grass cover increased, to 0.39, indicating the opposite. Although the range of reported effects skewed strongly negative, this was due to a small number of substantial outliers. Most values of Z_r were relatively small with values <|0.2| and the meta-analytic mean effect size was close to zero (-0.09 ± 0.12 95%CI). Of the 79 effects, fifty-three had confidence intervals overlapping zero, approximately a quarter (fifteen) crossed the traditional threshold of statistical significance indicating a negative relationship between grass cover and seedling success, and eleven crossed the significance threshold indicating a positive relationship between grass cover and seedling success (Figure 2b).

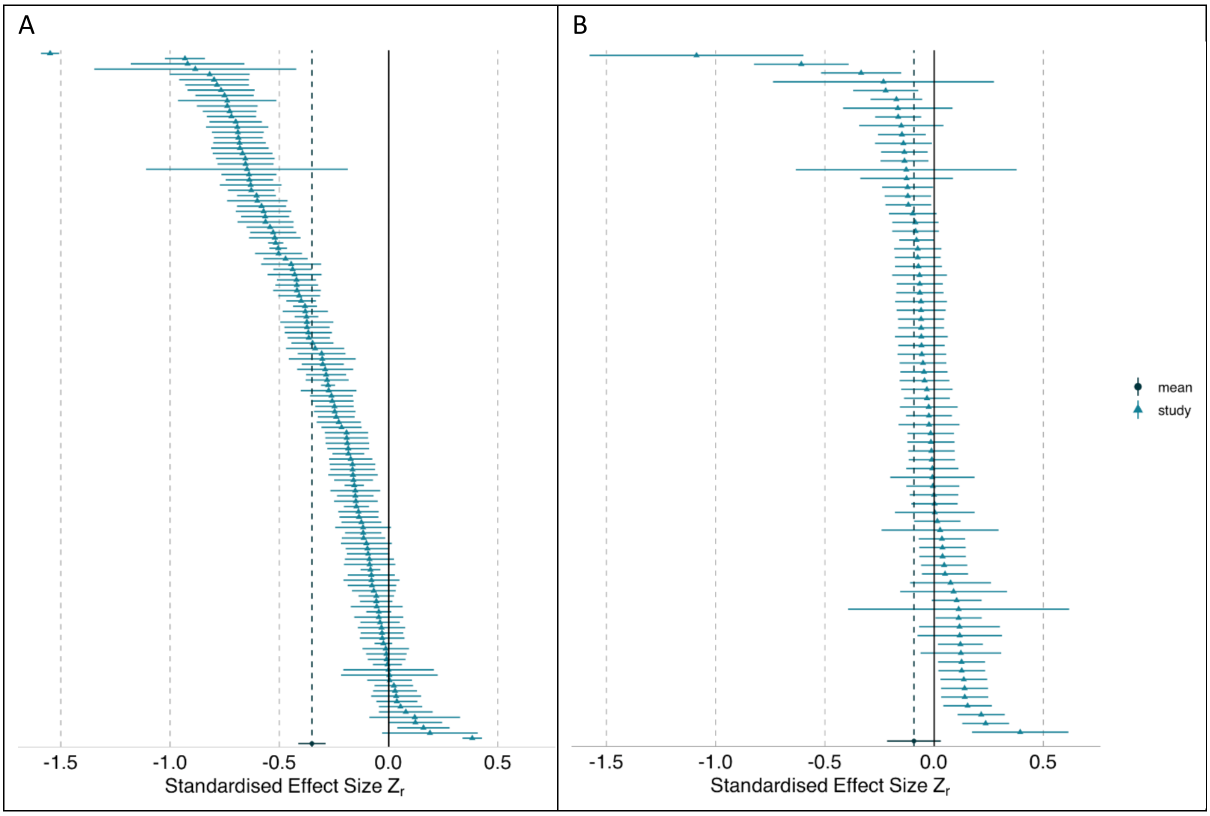

Figure 2: Forest plots of meta-analytic estimated standardized effect sizes (Z_r, blue triangles) and their 95% confidence intervals for each effect size included in the meta-analysis model. (A) Blue tit analyses: Points where Z_r are less than 0 indicate analyses that found a negative relationship between sibling number and nestling growth. (B) *Eucalyptus* analyses: Points where Z_r are less than 0 indicate a negative relationship between grass cover and Eucalyptus seedling success. The meta-analytic mean effect size is denoted by a black circle and a dashed vertical line, with error bars also representing the 95% confidence interval. The solid black vertical line demarcates effect size of 0, indicating no relationship between the test variable and the response variable. Note that the *Eucalyptus* plot omits one extreme outlier with the value of -4.47 (Supplementary Material A, Figure A.2) in order to standardize the x-axes on these two panels.

Out-of-sample predictions (y_i)

As with the effect size Z_r, we observed substantial variability in the size of out-of-sample predictions derived from the analysts' models. Blue tit predictions (Figure 3a), which were z-score-standardised to accommodate the use of different response variables, always ranged far in excess of one standard deviation. In the y_{25} scenario, model predictions ranged from -1.84 to 0.42 (a range of 2.68 standard deviations), in the y_{50} they ranged from -0.52 to 1.08 (a range of 1.63 standard deviations), and in the y_{75} scenario they ranged from -0.03 to 1.59 (a range of 1.9 standard deviations). As should be expected given the existence of both negative and positive Z_r values, all three out-of-sample scenarios produced both negative and positive predictions, although as with the Z_r values, there is a clear trend for scenarios with more siblings to be associated with smaller nestlings. This is supported by the meta-analytic means of these three sets of predictions which were -0.66 (95%CI -0.82–0.5) for the y_{25}, 0.34 (95%CI 0.2-0.48) for the y_{50}, and 0.67 (95%CI 0.57-0.77) for the y_{75}.

Eucalyptus out-of-sample predictions also varied substantially (Figure 3b), but because they were not z-score-standardised and are instead on the original count scale, the types of interpretations we can make differ. The predicted *Eucalyptus* seedling counts per 15 x 15 m plot for the y_{25} scenario ranged from 0.04 to 26.99, for the y_{50} scenario ranged from 0.04 to 44.34, and for the y_{75} scenario they ranged from 0.03 to 61.34. The meta-analytic mean predictions for these three scenarios were similar; 1.27 (95%CI 0.59-2.3) for the y_{25}, 2.92 (95%CI 0.98-3.89) for the y_{50}, and 2.92 (95%CI 1.59-4.9) for the y_{75} scenarios respectively.

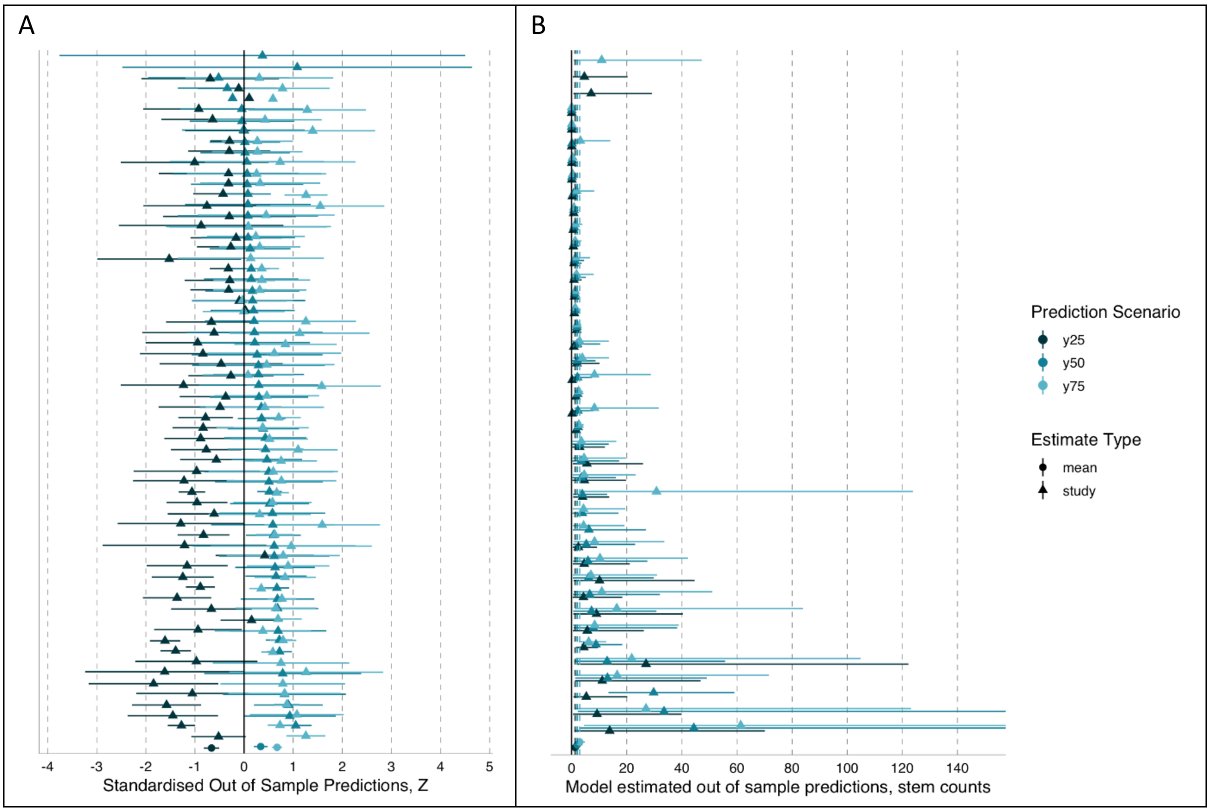

Figure 3: Forest plot of meta-analytic estimated out-of-sample predictions. A) Standardized (z-score) blue tit out-of-sample predictions, y_i. B) response-scale (stem counts) *Eucalyptus* out-of-sample predictions. Triangles represent individual estimates. Circles represent the meta-analytic mean for each prediction scenario. Dark-blue points correspond to y_{25} scenario, medium-blue points correspond to the y_{50} scenario, while light blue points correspond to the y_{75} scenario. Error bars are 95% confidence intervals. Note that, for the *Eucalyptus* analysis, outliers (observations more than 3 SD above the mean) have been removed prior to model fitting and do not appear on this figure. The x-axis is truncated to approximately 140, and thus some error bars are incomplete. See Supplementary Material B, Figure B.6 for full figure.

Quantifying heterogeneity

Effect sizes (Z_r)

We quantified both absolute (τ^2) and relative (I^2) heterogeneity resulting from analytical variation. Both measures suggest that substantial variability among effect sizes was attributable to the analytical decisions of analysts.

The total absolute level of variance beyond what would typically be expected due to sampling error, τ^2 (Table 2), among all usable blue tit effects was 0.08 and for *Eucalyptus* effects was 0.27. This is similar to or exceeding the median value (0.105) of τ^2 found across 31 recent meta-analyses (calculated from the data in Yang et al. 2023). The similarity of our observed values to values from meta-analyses of different studies based on different data suggest the potential for a large portion of heterogeneity to arise from analytical decisions. For further discussion of interpretation of τ^2 in our study, please consult discussion of post hoc analyses below.

Table 2: Heterogeneity in the estimated effects Z_r for meta-analyses of: the full dataset, as well as from post hoc analyses wherein analyses with outliers are removed, analyses with effects from analysis teams with at least one "unpublishable" rating are excluded, analyses receiving at least one "major revisions" rating or worse excluded, analyses from teams with at least one analyst self-rated as "highly proficient" or "expert" in statistical analysis are included, and (blue tit only) analyses that did not included the pair of highly collinear predictors together. τ^2_{Team} is the absolute heterogeneity for the random effect Team. $\tau^2_{Effect\ ID}$ is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{Team,\ Effect\ ID}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences.

Dataset	N_{Obs}	τ^2_{Total}	τ^2_{Team}	$\tau^2_{Effect\ ID}$	I^2_{Total}	I^2_{Team}	$I^2_{Team,\ Effect\ ID}$
All Analyses							
Eucalyptus	79	0.27	0.02	0.25	98.59%	6.89%	91.70%
blue tit	131	0.08	0.03	0.05	97.61%	36.71%	60.90%
Blue tit analyses containing highly collinear predictors removed							
blue tit	117	0.07	0.04	0.03	96.92%	58.18%	38.75%
All analyses, outliers removed							
Eucalyptus	75	0.01	0.00	0.01	66.19%	19.25%	46.94%
blue tit	127	0.07	0.04	0.02	96.84%	64.63%	32.21%
Analyses receiving at least one 'Unpublishable' rating removed							
Eucalyptus	55	0.01	0.01	0.01	79.74%	28.31%	51.43%
blue tit	109	0.08	0.03	0.05	97.52%	35.68%	61.84%
Analyses receiving at least one 'Unpublishable' and or 'Major Revisions' rating removed							
Eucalyptus	13	0.03	0.03	0.00	88.91%	88.91%	0.00%
blue tit	32	0.14	0.01	0.13	98.72%	5.17%	93.55%
Analyses from teams with highly proficient or expert data analysts							
Eucalyptus	34	0.58	0.02	0.56	99.41%	3.47%	95.94%
blue tit	89	0.09	0.03	0.06	97.91%	31.43%	66.49%

In our analyses, I^2 is a plausible index of how much more variability among effect sizes we have observed, as a proportion, than we would have observed if sampling error were driving variability. We discuss our interpretation of I^2 further in the methods, but in short, it is a useful metric for comparison to values from published meta-analyses and provides a plausible value for how much heterogeneity could arise in a normal meta-analysis with similar sample sizes due to analytical variability alone. In our study, total I^2 for the blue tit Z_r estimates was extremely large, at 97.61%, as was the *Eucalyptus* estimate (98.59% Table 2).

Although the overall I^2 values were similar for both *Eucalyptus* and blue tit analyses, the relative composition of that heterogeneity differed. For both datasets, the majority of heterogeneity in Z_r was driven by differences among effects as opposed to differences among teams, though this was more prominent for the *Eucalyptus* dataset, where nearly all of the total heterogeneity was driven by differences among effects (91.7%) as opposed to differences among teams (6.89%) (Table 2).

Out-of-sample predictions (y_i)

We observed substantial heterogeneity among out-of-sample estimates, but the pattern differed somewhat from the Z_r values (Table 3). Among the blue tit predictions, I^2 ranged from medium-high for the y_{25} scenario (68.54) to low (27.9) for the y_{75} scenario. Among the *Eucalyptus* predictions, I^2 values were uniformly high (>82%). For both datasets, most of the existing heterogeneity among predicted values was attributable to among-team differences, with the exception of the y_{50} analysis of the *Eucalyptus* dataset. We are limited in our interpretation of τ^2 for these estimates because, unlike for the Z_r estimates, we have no benchmark for comparison with other meta-analyses.

Table 3: Heterogeneity among the out-of-sample predictions y_i for both blue tit and *Eucalyptus* datasets. τ^2_{Team} is the absolute heterogeneity for the random effect Team. $T^2_{Effect\ ID}$ is the absolute heterogeneity for the random effect Effect ID nested under Team. Effect ID is the unique identifier assigned to each individual statistical effect submitted by an analysis team. We nested Effect ID within analysis team identity (Team) because analysis teams often submitted >1 statistical effect, either because they considered >1 model or because they derived >1 effect per model, especially when a model contained a factor with multiple levels that produced >1 contrast. τ^2_{Total} is the total absolute heterogeneity. I^2_{Total} is the proportional heterogeneity; the proportion of the variance among effects not attributable to sampling error, I^2_{Team} is the subset of the proportional heterogeneity due to differences among Teams and $I^2_{Team,Effect\ ID}$ is subset of the proportional heterogeneity attributable to among-Effect ID differences.

Prediction Scenario	N_{Obs}	T_{Total}	T^2_{Team}	$T^2_{Effect\ ID}$	I^2_{Total}	I^2_{Team}	$I^2_{Team,\ Effect\ ID}$
blue tit							
y25	63	0.23	0.11	0.03	68.54%	53.43%	15.11%
y50	60	0.23	0.06	0.00	50%	46.29%	3.71%
y75	63	0.23	0.02	0.00	27.9%	27.89%	0.01%
Eucalyptus							
y25	38	5.75	1.48	0.68	86.93%	59.54%	27.39%
y50	38	5.75	1.32	0.83	89.63%	55%	34.64%
y75	38	5.75	1.03	0.41	80.19%	57.41%	22.78%

Post-hoc analysis: Exploring outlier characteristics and the effect of outlier removal on heterogeneity

Effect sizes (Z_r)

The outlier *Eucalyptus Zr* values were striking and merited special examination. The three negative outliers had very low sample sizes that were based on either small subsets of the dataset or, in one case, extreme aggregation of data. The outliers associated with small subsets had sample sizes (n= 117, 90, 18) that were less than half of the total possible sample size of 351. The case of extreme aggregation involved averaging all values within each of the 351 sites in the dataset.

Surprisingly, both the largest and smallest effect sizes in the blue tit analyses (Figure 2a) come from the same analyst (anonymous ID: 'Adelong'), with identical models in terms of the explanatory variable structure, but with different response variables. However, the radical change in effect was primarily due to collinearity with covariates. The primary predictor variable (brood count after manipulation) was accompanied by several collinear variables, including the highly collinear (correlation of 0.89 Supplementary Material D, Figure D.2) covariate (brood count at day 14) in both analyses. In the analysis of nestling weight, brood count after manipulation showed a strong positive partial correlation with weight after controlling for brood count at day 14 and treatment category (increased, decreased, unmanipulated). In that same analysis, the most collinear covariate (the day 14 count) had a negative partial correlation with weight. In the analysis with tarsus length as the response variable, these partial correlations were almost identical in absolute magnitude, but reversed in sign and so brood count after manipulation was now the collinear predictor with the negative relationship. The two models were therefore very similar, but the two collinear predictors simply switched roles, presumably because a subtle difference in the distribution of weight and tarsus length data.

When we dropped the *Eucalyptus* outliers, I^2 decreased from high (98.59 %), using Higgins' (Higgins et al. 2003) suggested benchmark, to between moderate and high (66.19 %, Table 2). However, more notably, τ^2 dropped from 0.27 to 0.01, indicating that, once outliers were excluded, the observed variation in effects was similar to what we would expect if sampling error were driving the differences among effects (since τ^2 is the variance beyond that driven by sampling error). The interpretation of this value of τ^2 in the context of our many-analyst study is somewhat different than a typical meta-analysis, however, since in our study (especially for *Eucalyptus*, where most analyses used almost exactly the same data points), there is almost no role for sampling error in driving the observed differences among the estimates. Thus, rather than concluding that the variability we observed among estimates (after removing outliers) was due only to sampling error (because τ^2 became small: 10% of the median from Yang et al. 2023), we instead conclude that the observed variability, which must be due to the divergent choices of analysts rather than sampling error, is approximately of the same magnitude as what we would have expected if, instead, sampling error, and not analytical heterogeneity, were at work. Conversely, dropping outliers from the set of blue tit effects did not meaningfully reduce I^2, and only modestly reduced τ^2 (Table 2). Thus, effects at the extremes of the distribution were much stronger contributors to total heterogeneity for effects from analyses of the *Eucalyptus* than for the blue tit dataset.

Table 4: Estimated mean value of the standardised correlation coefficient, $\bar{Z}_r$, along with its standard error and 95% confidence intervals. We re-computed the meta-analysis for different post hoc subsets of the data: All eligible effects, removal of effects from blue tit analyses that contained a pair of highly collinear predictor variables, removal of effects from analysis teams that received at least one

peer rating of “deeply flawed and unpublishable”, removal of any effects from analysis teams that received at least one peer rating of either “deeply flawed and unpublishable” or “publishable with major revisions”,, inclusion of only effects from analysis teams that included at least one member who rated themselves as “highly proficient” or “expert” at conducting statistical analyses in their research area.

Dataset	$\hat{\mu}$	$SE[\widehat{\mu}]$	95% CI	statistic	p
All analyses					
Eucalyptus	−0.09	0.06	[−0.22,0.03]	−1.47	0.14
blue tit	−0.35	0.03	[−0.41,−0.29]	−11.02	<0.001
Blue tit analyses containing highly collinear predictors removed					
blue tit	−0.36	0.03	[−0.42,−0.29]	−10.97	<0.001
All analyses, outliers removed					
Eucalyptus	−0.03	0.01	[−0.06,0.00]	−2.23	0.026
blue tit	−0.36	0.03	[−0.42,−0.30]	−11.48	<0.001
Analyses receiving at least one 'Unpublishable' rating removed					
Eucalyptus	−0.02	0.02	[−0.07,0.02]	−1.15	0.3
blue tit	−0.36	0.03	[−0.43,−0.30]	−10.82	<0.001
Analyses receiving at least one 'Unpublishable' and or 'Major Revisions' rating removed					
Eucalyptus	−0.04	0.05	[−0.15,0.07]	−0.77	0.4
blue tit	−0.37	0.07	[−0.51,−0.23]	−5.34	<0.001
Analyses from teams with highly proficient or expert data analysts					
Eucalyptus	−0.17	0.13	[−0.43,0.10]	−1.24	0.2
blue tit	−0.36	0.04	[−0.44,−0.28]	−8.93	<0.001

Out-of-sample predictions (y_i)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller and the pattern of outliers differed.

Post hoc analysis: Exploring the effect of removing analyses with poor peer ratings on heterogeneity

Effect sizes (Z_r)

Removing poorly rated analyses had limited impact on the meta-analytic means (Supplementary Material B, Figure B.3). For the *Eucalyptus* dataset, the meta-analytic mean shifted from -0.09 to -0.02 when effects from analyses rated as unpublishable were removed, and to -0.04 when effects from analyses rated, at least once, as unpublishable or requiring major revisions were removed. Further, the confidence intervals for all of these means overlapped each of the other means (Table 4). We saw similar patterns for the blue tit dataset, with only small shifts in the meta-analytic mean, and confidence intervals of all three means overlapping each other mean (Table 4). Refitting the meta-analysis with a fixed effect for categorical ratings also showed no indication of differences in group meta-analytic means due to peer ratings (Supplementary Material B, Figure B.1).

For the blue tit dataset, removing poorly-rated analyses led to only negligible changes in I^2_{Total} and relatively minor impacts on τ^2. However, for the *Eucalyptus* dataset, removing poorly-rated analyses led to notable reductions in I^2_{Total} and substantial reductions in τ^2. When including all analyses, the *Eucalyptus* I^2_{Total} was 98.59% and τ^2 was 0.27, but eliminating analyses with ratings of "unpublishable" reduced I^2_{Total} to 79.74% and τ^2 to 0.01, and removing also those analyses "needing major revisions" left I^2_{Total} at 88.91% and τ^2 at 0.03 (Table 2). Additionally, the allocations of I^2 to the team versus individual effect were altered for both blue tit and *Eucalyptus* meta-analyses by removing poorly-rated analyses, but in different ways. For blue tit meta-analysis, between a third and two-thirds of the total I^2 was attributable to among-team variance in most analyses until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost all remaining heterogeneity was attributable to among-effect differences. In contrast, for *Eucalyptus* meta-analysis, the among-team component of I^2 was less than third until both analyses rated "unpublishable" and analyses rated in need of "major revision" were eliminated, in which case almost 90% of heterogeneity was attributable to differences among teams.

Out-of-sample predictions (y_i)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller and our ability to interpret heterogeneity values for these analyses was limited

Post hoc analysis: Exploring the effect of including only analyses conducted by analysis teams with at least one member self-rated as "highly proficient" or "expert" in conducting statistical analyses in their research area

Effect sizes (Z_r)

Including only analyses conducted by teams that contained at least one member who rated themselves as "highly proficient" or "expert" in conducting the relevant statistical methods had negligible impacts on the meta-analytic means (Table 4), the distribution of Z_r effects (Supplementary Material B, Figure B.4), or heterogeneity estimates (Table 2), which remained extremely high.

Out-of-sample predictions (y_i)

We did not conduct these post hoc analyses on the out-of-sample predictions as the number of eligible effects was smaller.

Post hoc analysis: Exploring the effect of excluding estimates of Z_r in which we had reduced confidence

As described in our addendum to the methods, we identified a subset of estimates of Z_r in which we had less confidence because of features of the submitted degrees of freedom. Excluding these effects in which we had lower confidence had minimal impact on the meta-analytic mean and the estimates of total I^2 and τ^2 for both blue tit and *Eucalyptus* meta-analyses, regardless of whether outliers were also excluded (Supplementary Material B, Table B.1).

Post hoc analysis: Exploring the effect of excluding effects from blue tit models that contained two highly collinear predictors

Effect sizes (Z_r)

Excluding effects from blue tit models that contained the two highly collinear predictors (brood count after manipulation and brood count at day 14) had negligible impacts on the meta-analytic means (Table 4), the distribution of Z_r effects (Supplementary Material B, Figure B.5), or heterogeneity estimates (Table 2), which remained high.

Out-of-sample predictions

Inclusion of collinear predictors does not harm model prediction, and so we did not conduct these post hoc analyses.

Explaining Variation in Deviation Scores

None of the pre-registered predictors explained substantial variation in deviation among submitted statistical effects from the meta-analytic mean (Table 5, Table 6).

Table 5: Summary metrics for registered models seeking to explain deviation (Box-Cox transformed absolute deviation scores) from $\bar{Z}_r$ as a function of Sorensen's Index, categorical peer ratings, and continuous peer ratings for blue tit and *Eucalyptus* analyses, and as a function of the presence or absence of random effects (in the analyst's models) for *Eucalyptus* analyses. We report coefficient of determination, R^2, for our models including only fixed effects as predictors of deviation, and we report $R^2_{Conditional}$, $R^2_{Marginal}$ and the intra-class correlation (ICC) from our models that included both fixed and random effects. For all our models, we calculated the residual standard deviation σ and root mean squared error (RMSE).

Dataset	NObs	R^2	$R^2_{Conditional}$	$R^2_{Marginal}$	ICC	σ	RMSE
Deviation explained by categorical ratings							
Eucalyptus	346		0.13	0.01	0.12	1.06	1.02
blue tit	473		0.09	7.47×10^{-3}	0.08	0.5	0.48
Deviation explained by continuous ratings							
Eucalyptus	346		0.12	7.44×10^{-3}	0.11	1.06	1.03
blue tit	473		0.09	3.44×10^{-3}	0.09	0.5	0.48
Deviation explained by Sorensen's index							
Eucalyptus	79	1.84×10^{-4}				1.12	1.1
blue tit	131	6.32×10^{-3}				0.51	0.51
Deviation explained by inclusion of random effects							
Eucalyptus	79	8.75×10^{-8}				1.12	1.1

Table 6: Parameter estimates from models of Box-Cox transformed deviation scores from $\bar{Z}_r$ as a function of continuous and categorical peer ratings, Sorensen scores, and the inclusion of random effects. Standard Errors (SE), 95% confidence intervals (95% CI) are reported for all estimates, while *t* values, degrees of freedom and p-values are presented for fixed-effects. Note that positive parameter estimates mean that as the predictor variable increases, so does the absolute value of the deviation from the meta-analytic mean.

Parameter	Random effect	Coefficient	SE	95% CI	t	*df*	p
Deviation explained by inclusion of random effects - *Eucalyptus*							
(Intercept)		-2.53	0.27	[-3.06, -1.99]	-9.31	77	<0.001
Mixed model		0.00	0.31	[-0.60, 0.60]	0.00	77	>0.9
Deviation explained by Sorensen’s index - *Eucalyptus*							
(Intercept)		-2.65	1.05	[-4.70, -0.60]	-2.53	77	0.011
Mean Sorensen's index		0.18	1.51	[-2.78, 3.14]	0.12	77	>0.9
Deviation explained by Sorensen’s index - blue tit							
(Intercept)		-1.53	0.28	[-2.08, -0.98]	-5.42	129	<0.001
Mean Sorensen's index		0.42	0.47	[-0.49, 1.34]	0.91	129	0.4
Deviation explained by continuous ratings - *Eucalyptus*							
(Intercept)		-2.23	0.23	[-2.69, -1.78]	-9.65	342	<0.001
RateAnalysis		-0.004	0	[-0.011, 0]	-1.44	342	0.15
SD (Intercept)	Reviewer ID	0.37	0.09	[0.24, 0.60]			
SD (Observations)	Residual	1.06	0.04	[0.98, 1.15]			
Deviation explained by continuous ratings - blue tit							
(Intercept)		-1.16	0.11	[-1.37, -0.94]	-10.60	469	<0.001
RateAnalysis		-0.002	0	[-0.004, 0]	-1.22	469	0.2
SD (Intercept)	Reviewer ID	0.16	0.03	[0.10,0.24]			
SD (Observations)	Residual	0.5	0.02	[0.46,0.53]			
Deviation explained by categorical ratings - *Eucalyptus*							
(Intercept)		-2.66	0.27	[-3.18, -2.13]	-9.97	340	<0.001
Publishable with major revision		0.29	0.29	[-0.27, 0.85]	1.02	340	0.3
Publishable with minor revision		0.01	0.28	[-0.54, 0.56]	0.04	340	>0.9
Publishable as is		0.05	0.31	[-0.55, 0.66]	0.17	340	0.9
SD (Intercept)	Reviewer ID	0.39	0.09	[0.25, 0.61]			
SD (Observations)	Residual	1.06	0.04	[0.98, 1.15]			
Deviation explained by categorical ratings - blue tit							
(Intercept)		-1.11	0.11	[-1.33, -0.89]	-9.91	467	<0.001
Publishable with major revision		-0.19	0.12	[-0.42, 0.04]	-1.62	467	0.10
Publishable with minor revision		-0.19	0.12	[-0.42, 0.04]	-1.65	467	0.10
Publishable as is		-0.13	0.13	[-0.39, 0.12]	-1.02	467	0.3
SD (Intercept)	Reviewer ID	0.15	0.04	[0.10, 0.24]			
SD (Observations)	Residual	0.5	0.02	[0.46, 0.53]			

Deviation scores as explained by reviewer ratings

Effect sizes (Z_r)

We obtained reviews from 153 reviewers who reviewed analyses for a mean of 3.27 (range 1 - 11) analysis teams. Analyses of the blue tit dataset received a total of 240 reviews, each was reviewed by a mean of 3.87 (SD 0.71, range 3-5) reviewers. Analyses of the *Eucalyptus* dataset received a total of 178 reviews, each was reviewed by a mean of 4.24 (SD 0.79, range 3-6) reviewers. We tested for inter-rater-reliability (IRR) to examine how similarly reviewers reviewed each analysis and found approximately no agreement among reviewers. When considering continuous ratings, IRR was 0.01, and for categorical ratings, IRR was -0.14.

Many of the models of deviation as a function of peer ratings faced issues of failure to converge or singularity due to sparse design matrices with our pre-registered random effects (Effect ID and Reviewer ID) (see Supplementary Material C). These issues persisted after increasing the tolerance and changing the optimizer. For both *Eucalyptus* and blue tit datasets, models with continuous ratings as a predictor were singular when both pre-registered random effects were included.

When using both categorical and continuous ratings as predictors, only models converged and allowed 95% confidence intervals to be calculated when specifying Reviewer ID as a random effect. The categorical ratings model had a R^2_C of 0.09 and a R^2_M of 0.01, the continuous ratings model had a R^2_C of 0.09 and a R^2_M of 0.01 for the blue tit dataset and a R^2_C of 0.12 and a R^2_M of 0.01 for the *Eucalyptus* dataset. Neither continuous or categorical reviewer ratings of the analyses meaningfully predicted deviance from the meta-analytic mean (Table 6, Figure 4). We re-ran the multi-level meta-analysis with a fixed effect for the categorical publishability ratings and found no difference in mean standardised effect sizes among publishability ratings (Supplementary Material B, Figure B.1).

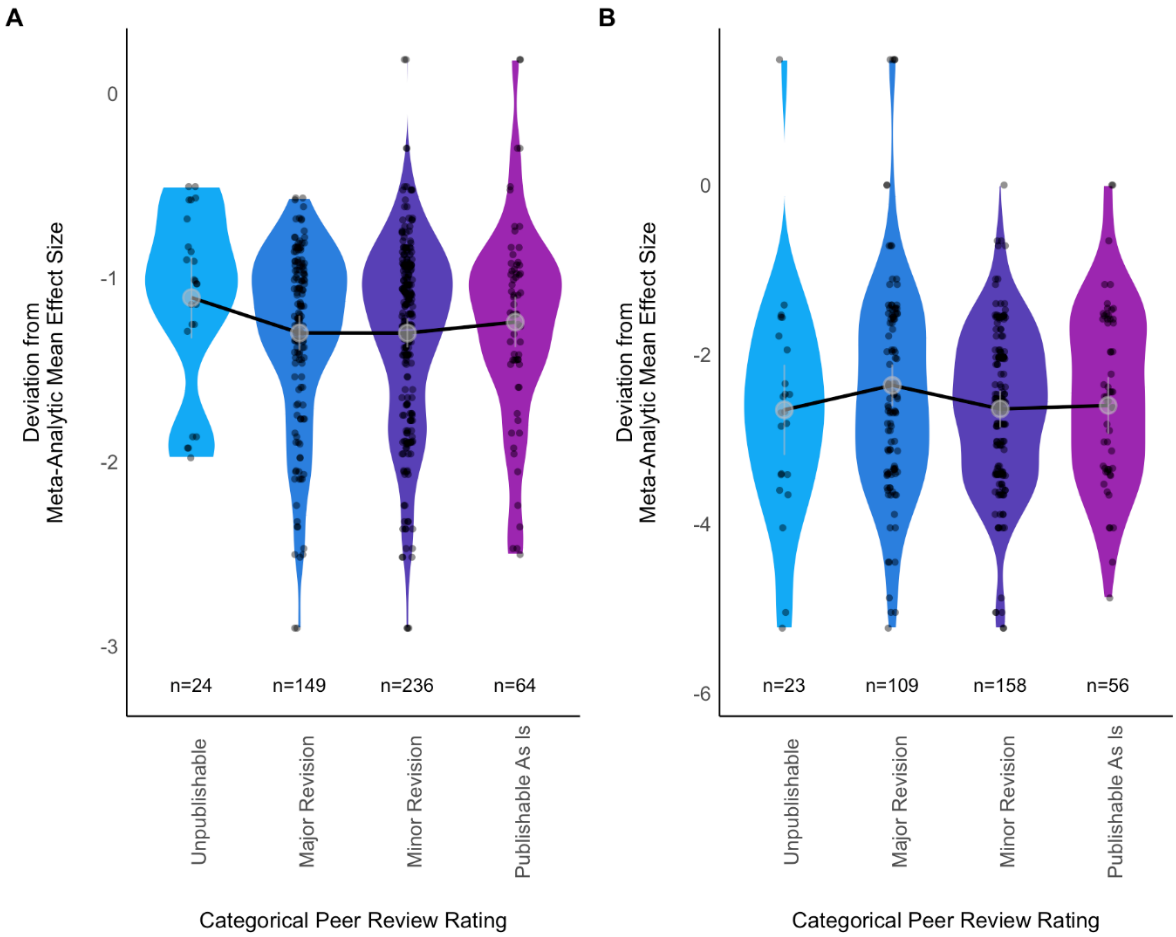

Figure 4: Violin plot of Box-Cox transformed deviation from meta-analytic mean $\bar{Z}_r$ as a function of categorical peer rating. Grey points for each rating group denote model-estimated marginal mean deviation, and error bars denote 95%CI of the estimate. **A** Blue tit dataset, **B** *Eucalyptus* dataset.

Out-of-sample predictions (y_i)

Some models of the influence of reviewer ratings on out-of-sample predictions (y_i) had issues with convergence and singularity of fit (see Supplementary Material C, Table C.3) and those models that converged and were not singular showed no strong relationship (Supplementary Material C, Figure C.2, Supplementary Material C, Figure C.3), as with the Z_r analyses.

Deviation scores as explained by the distinctiveness of variables in each analysis

Effect sizes (Z_r)

We employed Sorensen's index to calculate the distinctiveness of the set of predictor variables used in each model (Figure 5). The mean Sorensen's score for blue tit analyses was 0.59 (SD: 0.1, range 0.43-0.86), and for *Eucalyptus* analyses was 0.69 (SD: 0.08, range 0.55-0.98).

We found no meaningful relationship between distinctiveness of variables selected and deviation from the meta-analytic mean (Table 6, Figure 5) for either blue tit (mean 0.42, 95%CI -0.49,1.34) or *Eucalyptus* effects (mean 0.18, 95%CI -2.78,3.14).

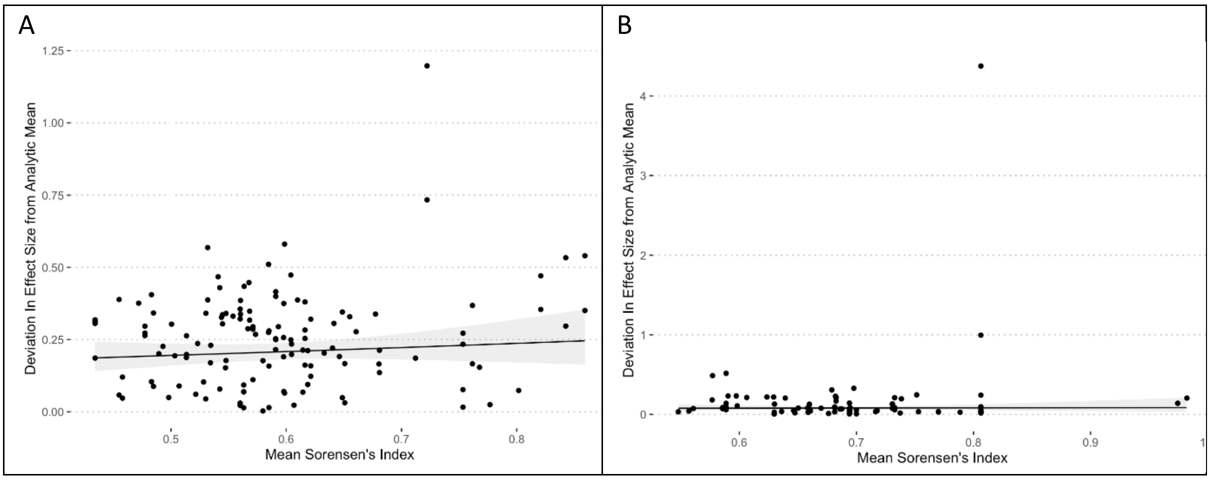

Figure 5: Fitted model of the Box-Cox-transformed deviation score (deviation in effect size from meta-analytic mean) as a function of the mean Sorensen's index showing distinctiveness of the set of predictor variables. Grey ribbons on predicted values are 95% CI's. A) blue tit dataset, B) *Eucalyptus* dataset.

Out-of-sample predictions (y_i)

As with the Z_r estimates, we did not observe any convincing relationships between deviation scores of out-of-sample predictions and Sorensen's index values (see Supplementary Material C4.1).

Deviation scores as explained by the inclusion of random effects

Effect sizes (Z_r)

There were only three blue tit analyses that did not include random effects, which is below the pre-registered threshold for fitting a model of the Box-Cox transformed deviation from the meta-analytic mean as a function of whether the analysis included random-effects. However, 17 *Eucalyptus* analyses included only fixed effects, which crossed our pre-registered threshold.

Consequently, we performed this analysis for the *Eucalyptus* dataset only. There was no relationship between random-effect inclusion and deviation from meta-analytic mean among the *Eucalyptus* analyses (Table 6, Figure 6).

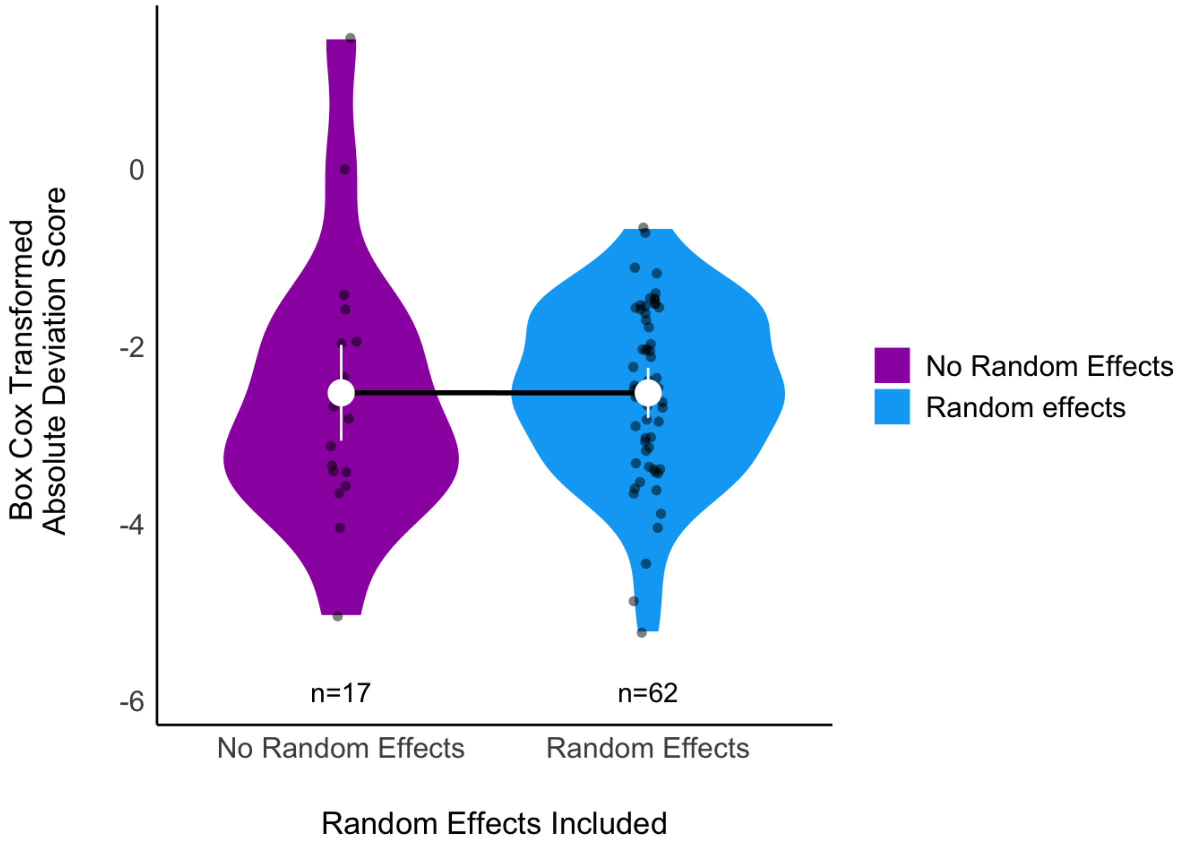

Figure 6: Violin plot of mean Box-Cox transformed deviation from meta-analytic mean as a function of random-effects inclusion in *Eucalyptus* analyses. White point for each group of analyses denotes model-estimated marginal mean deviation, and error bars denote 95% CI of the estimate.

Out-of-sample predictions (y_i)

As with the Z_r estimates, we did not examine the possibility of a relationship between the inclusion of random effects and the deviation scores of the blue tit out-of-sample predictions. When we examined the possibility of this relationship for the *Eucalyptus* effects, we found consistent evidence of somewhat higher Box-Cox-transformed deviation values for models including a random effect, meaning the models including random effects averaged slightly higher deviation from the meta-analytic means (Supplementary Material C, Figure C.5).

Multivariate Analysis Effect size (Z_r) and out-of-sample predictions (y_i)

Like the univariate models, the multivariate models did a poor job of explaining deviations from the meta-analytic mean. Because we pre-registered a multivariate model that contained collinear predictors that produce results which are not readily interpretable, we present these models in the supplement. We also had difficulty with convergence and singularity for multivariate models of out-of-sample (y_i) result, and had to adjust which random effects we included (Supplementary Material C, Table C.8). However, no multivariate analyses of *Eucalyptus* out-of-sample results avoided problems of convergence or singularity, no matter which random effects we included (Supplementary Material C, Table C.8). We therefore present no multivariate *Eucalyptus* y_i models.

We present parameter estimates from multivariate Z_r models for both datasets (Supplementary Material C, Table C.6, Table C.7) and from y_i models from the blue tit dataset (Supplementary Material C, Table C.10, Table C.9). We include interpretation of the results from these models in the supplement, but the results do not change the interpretations we present above based on the univariate analyses.

Discussion

When a large pool of ecologists and evolutionary biologists analyzed the same two datasets to answer the corresponding two research questions, they produced substantially heterogeneous sets of answers. Although the variability in analytical outcomes was high for both datasets, the patterns of this variability differed distinctly between them. For the blue tit dataset, there was nearly continuous variability across a wide range of Z_r values. In contrast, for the *Eucalyptus* dataset, there was less variability across most of the range, but more striking outliers at the tails. Among out-of-sample predictions, there was again almost continuous variation across a wide range (2 SD) among blue tit estimates. For *Eucalyptus*, out-of-sample predictions were also notably variable, with about half the predicted stem count values at <2 but the other half being much larger, and ranging to nearly 40 stems per 15 m x 15 m plot. We investigated several hypotheses for drivers of this variability within datasets, but found little support for any of these. Most notably, even when we excluded analyses that had received one or more poor peer reviews, the heterogeneity in results largely persisted. Regardless of what drives the variability, the existence of such dramatically heterogeneous results when ecologists and evolutionary biologists seek to answer the same questions with the same data should trigger conversations about how ecologists and evolutionary biologists analyze data and interpret the results of their own analyses and those of others in the literature (e.g., Silberzahn et al. 2018; Simonsohn, Simmons, and Nelson 2020; Auspurg and Brüderl 2021; Breznau et al. 2022).

Our observation of substantial heterogeneity due to analytical decisions is consistent with a small earlier study in ecology (Stanton-Geddes, de Freitas and de Sales Dambros 2014) and a growing body of work from the quantitative social sciences (e.g., Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Huntington-Klein et al. 2021; Schweinsberg et al. 2021; Breznau et al. 2022; Coretta et al. 2023). In these studies, when volunteers from the discipline analyzed the same data, they produced a worryingly diverse set of answers to a pre-set question. This diversity included a wide range of effect sizes, and in most cases, even involved effects in opposite directions. Thus, our result should not be viewed as an anomalous outcome from two particular datasets, but instead as evidence from additional disciplines regarding the heterogeneity that can emerge from analyses of complex datasets to answer questions in probabilistic science. Not only is our major observation consistent with other studies, it is, itself, robust because it derived primarily from simple forest plots that we produced based on a small set of decisions that were mostly registered before data gathering and which conform to widely accepted meta-analytic practices.

Unlike the strong pattern we observed in the forest plots, our other analyses, both registered and post hoc, produced either inconsistent patterns, weak patterns, or the absence of patterns. Our registered analyses found that deviations from the meta-analytic mean by individual effect sizes ($\overline{Zr}$) or the predicted values of the dependent variable ($\bar{y}$) were poorly explained by our hypothesized predictors: peer rating of each analysis team's method section, a measurement of the distinctiveness of the set of predictor variables included in each analysis, or whether the model included random effects. However, in our post hoc analyses, we found that dropping analyses identified as unpublishable or in need of major revision by at least one reviewer modestly reduced the observed heterogeneity among the Z_r outcomes, but only for *Eucalyptus* analyses, apparently because this led to the dropping of the major outlier. This limited role for peer review in explaining the variability in our results should be interpreted cautiously because the inter-rater reliability among peer reviewers was extremely low, and at least some analyses that appeared flawed to us were not marked as flawed by reviewers. Thus, it seems that the peer reviews we received were of mixed quality, possibly due to lack of expertise or lack of care on the part of some reviewers. However, the hypothesis that poor quality analyses drove a substantial portion of the heterogeneity we observed was also contradicted by our observation that analysts' self-declared statistical expertise appeared unrelated to heterogeneity. When we retained only analyses from teams including at least one member with high self-declared levels of expertise, heterogeneity among effect sizes remained high. Thus, our results suggest lack of statistical expertise is not the primary factor responsible for the heterogeneity we observed, although further work is merited before rejecting a role for statistical expertise. Besides variability in expertise, it is also possible that the volunteer analysts varied in the effort they invested, and low effort presumably drove at least some heterogeneity in results. However, analysts often submitted thoughtful and extensive code, tables, figures, and textual explanation and interpretations, which is evidence of substantial investment. Further, we are confident that low effort alone is an insufficient explanation for the heterogeneity we observed because we have worked with these datasets ourselves, and we know from experience that there are countless plausible modeling alternatives that can produce a diversity of effects. Additionally, heterogeneity in analytical outcomes differed notably between datasets, and there is no reason to expect that one set of analysts took this project less seriously than the other. Returning to our exploratory analyses, not surprisingly, simply dropping outlier values of Z_r for *Eucalyptus* analyses, which had more extreme outliers, led to less observable heterogeneity in the forest plots, and also reductions in our quantitative measures of heterogeneity. We did not observe a similar effect in the blue tit dataset because that dataset had outliers that were much less extreme and instead had more variability across the core of the distribution.

Our major observations raise two broad questions; why was the variability among results so high, and why did the pattern of variability differ between our two datasets. One important and plausible answer to the first question is that much of the heterogeneity derives from the lack of a precise relationship between the two biological research questions we posed and the data we provided. This lack of a precise relationship between data and question creates many opportunities for different model specifications, and so may inevitably lead to varied analytical outcomes (Auspurg and Brüderl 2021). However, we believe that the research questions we posed are consistent with the kinds of research question that ecologists and evolutionary biologists typically work from. When designing the two biological research questions, we deliberately sought to represent the level of specificity we typically see in these disciplines. This level of specificity is evident when we look at the research questions posed by some recent meta-analyses in these fields:

- "how [does] urbanisation impact mean phenotypic values and phenotypic variation … [in] paired urban and non-urban comparisons of avian life-history traits" (Capilla-Lasheras et al. 2022)

- "[what are] the effects of ocean acidification on the crustacean exoskeleton, assessing both exoskeletal ion content (calcium and magnesium) and functional properties (biomechanical resistance and cuticle thickness)" (Siegel et al. 2022)

- "[what is] the extent to which restoration affects both the mean and variability of biodiversity outcomes … [in] terrestrial restoration" (Atkinson et al. 2022)

- “[does] drought stress [have] a negative, positive, or null effect on aphid fitness” (Leybourne et al. 2021)
- “[what is] the influence of nitrogen-fixing trees on soil nitrous oxide emissions” (Kou-Giesbrecht and Menge 2021)

There is not a single precise answer to any of these questions, nor to the questions we posed to analysts in our study. And this lack of single clear answers will obviously continue to cause uncertainty since ecologists and evolutionary biologists conceive of the different answers from the different statistical models as all being answers to the same general question. A possible response would be a call to avoid these general questions in favor of much more precise alternatives (Auspurg and Brüderl 2021). However, the research community rewards researchers who pose broad questions (Simons, Shoda, and Lindsay 2017), and so researchers are unlikely to narrow their scope without a change in incentives. Further, we suspect that even if individual studies specified narrow research questions, other scientists would group these more narrow questions into broader categories, for instance in meta-analyses, because it is these broader and more general questions that often interest the research community.

Although variability in statistical outcomes among analysts may be inevitable, our results raise questions about why this variability differed between our two datasets. We are particularly interested in the differences in the distribution of Z_r since the distributions of out-of-sample predictions were on different scales for the two datasets, thus limiting the value of comparisons. The forest plots of Z_r from our two datasets showed distinct patterns, and these differences are consistent with several alternative hypotheses. The results submitted by analysts of the *Eucalyptus* dataset showed a small average (close to zero) with most estimates also close to zero (± 0.2), though about a third far enough above or below zero to cross the traditional threshold of statistical significance. There were a small number of striking outliers that were very far from zero. In contrast, the results submitted by analysts of the blue tit dataset showed an average much further from zero (- 0.35) and a much greater spread in the core distribution of estimates across the range of Z_r values (± 0.5 from the mean), with few modest outliers. So, why was there more spread in effect sizes (across the estimates that are not outliers) in the blue tit analyses relative to the *Eucalyptus* analyses?

One possible explanation for the lower heterogeneity among most *Eucalyptus* Z_r effects is that weak relationships may limit the opportunities for heterogeneity in analytical outcome. Some evidence for this idea comes from two sets of “many labs” studies in psychology (Klein et al. 2014, 2018). In these studies, many independent lab groups each replicated a large set of studies, including, for each study, the experiment, data collection, and statistical analyses. These studies showed that, when the meta-analytic mean across the replications from different labs was small, there was much less heterogeneity among the outcomes than when the mean effect sizes were large (Klein et al. 2014, 2018). Of course, a weak average effect size would not prevent divergent effects in all circumstances. As we saw with the *Eucalyptus* analyses, taking a radically smaller subset of the data can lead to dramatically divergent effect sizes even when the mean with the full dataset is close to zero.

Our observation that dramatic sub-setting in the *Eucalyptus* dataset was associated with correspondingly dramatic divergence in effect sizes leads us towards another hypothesis to explain the differences in heterogeneity between the *Eucalyptus* and blue tit analysis sets. It may be that when analysts often divide a dataset into subsets, the result will be greater heterogeneity in analytical outcome for that dataset. Although we saw sub-setting associated with dramatic outliers in the *Eucalyptus* dataset, nearly all other analyses of *Eucalyptus* data used close to the same set of 351 samples, and as we saw, these effects did not vary substantially. However, analysts often analyzed only a subset of the blue tit data, and as we observed, sample sizes were much more variable among blue tit effects, and the effects themselves were also much more variable. Important to note here is that subsets of data may differ from each other for biological reasons, but they may also differ due to sampling error. Sampling error is a function of sample size, and sub-samples are, by definition, smaller samples, and so more subject to variability in effects due to sampling error (Jennions et al. 2013).

Other features of datasets are also plausible candidates for driving heterogeneity in analytical outcomes, including features of covariates. In particular, relationships between covariates and the response variable as well as relationships between covariates and the primary independent variable (collinearity) can strongly influence the modeled relationship between the independent variable of interest and the dependent variable (Morrissey and Ruxton 2018; Dormann et al. 2013). Therefore, inclusion or exclusion of these covariates can drive heterogeneity in effect sizes (Z_r). Also, as we saw with the two most extreme Z_r values from the blue tit analyses, in multivariate models with collinear predictors, extreme effects can emerge when estimating partial correlation coefficients due to high collinearity, and conclusions can differ dramatically depending on which relationship receives the researcher's attention. Therefore, differences between datasets in the presence of strong and/or collinear covariates could influence the differences in heterogeneity in results among those datasets.

Although it is too early in the many-analyst research program to conclude which analytical decisions or which features of datasets are the most important drivers of heterogeneity in analytical outcomes, we must still grapple with the possibility that analytical outcomes may vary substantially based on the choices we make as analysts. If we assume that, at least sometimes, different analysts will produce dramatically different statistical outcomes, what should we do as ecologists and evolutionary biologists? We review some ideas below.

The easiest path forward after learning about this analytical heterogeneity would be simply to continue with "business as usual", where researchers report results from a small number of statistical models. A case could be made for this path based on our results. For instance, among the blue tit analyses, the precise values of the estimated Z_r effects varied substantially, but the average effect was convincingly different from zero, and a majority of individual effects (84%) were in the same direction. Arguably, many ecologists and evolutionary biologists appear primarily interested in the direction of a given effect and the corresponding p-value (Fidler et al. 2006), and so the variability we observed when analyzing the blue tit dataset may not worry these researchers. Similarly, most effects from the *Eucalyptus* analyses were relatively close to zero, and about two-thirds of these effects did not cross the traditional threshold of statistical significance. Therefore, a large proportion of people analyzing these data would conclude that there was no effect, and this is consistent with what we might conclude from the meta-analysis.

However, we find the counter arguments to "business as usual" to be compelling. For blue tits, there were a substantial minority of calculated effects that would be interpreted by many biologists as indicating the absence of an effect (28%), and there were three traditionally 'significant' effects in the opposite direction to the average. The qualitative conclusions of analysts also reflected substantial variability, with fully half of teams drawing a conclusion distinct from the one we draw from the distribution as a whole. These teams with different conclusions were either uncertain about the negative relationship between competition and nestling growth, or they concluded that effects were mixed or absent. For the *Eucalyptus* analyses, this issue is more concerning. Around two-thirds of effects had confidence intervals overlapping zero, and of the third of analyses with confidence intervals excluding zero, almost half were positive, and the rest were negative. Accordingly, the qualitative conclusions of the *Eucalyptus* teams were spread across the full range of possibilities. But, as we describe in the next paragraph, even this striking lack of consensus may be much less of a problem than what could emerge as scientists select which results to publish.

A potentially larger argument against “business as usual” is that it provides the raw material for biasing the literature. When different model specifications readily lead to different results, analysts may be tempted to report the result that appears most interesting, or that is most consistent with expectation (Gelman and Loken 2013; Forstmeier, Wagenmakers and Parker 2017). There is growing evidence that researchers in ecology and evolutionary biology often report a biased subset of the results they produce (Deressa et al. 2023; Kimmel, Avolio and Ferraro 2023), and that this bias exaggerates the average size of effects in the published literature between 30 and 150% (Yang et al. 2023; Parker and Yang 2023). The bias then accumulates in meta-analyses, apparently more than doubling the rate of conclusions of “statistical significance” in published meta-analyses above what would have been found in the absence of bias (Yang et al. 2023). Thus, “business as usual” does not just create noisy results, it helps create systematically misleading results.

If we move away from “business as usual”, where do we go? Many obvious options involve multiple analyses per dataset. For instance, there is the traditional robustness or sensitivity check (e.g., Pei et al. 2020; Briga and Verhulst 2021), in which the researcher presents several alternative versions of an analysis to demonstrate that the result is ‘robust’ (Lu and White 2014). Unfortunately, robustness checks are at risk of the same potential biases of reporting found in other studies (Silberzahn et al. 2018), especially given the relatively few models typically presented. However, these risks could be minimized by running more models and doing so with a pre-registration or registered report. Another option is model averaging. Averages across models often perform well (e.g. Taylor and Taylor 2023), and in some forms this may be a relatively simple solution. Model averaging, as most often practiced in ecology and evolutionary biology, involves first identifying a small suite of candidate models (see Burnham and Anderson 2002), then using Akaike weights, based on Akaike’s Information Criterion (AIC), to calculate weighted averages for parameter estimates from those models. As with typical robustness checks, the small number of models limits the exploration of specification space, but examining a larger number of models could become the norm. However, there are more concerning limitations. The largest of these limitations is that averaging regression coefficients is problematic when models differ in interaction terms or collinear variables (Cade 2015). Additionally, weighting by AIC may often be inconsistent with our modelling goals. AIC balances the trade-off between model complexity and predictive ability, but penalizing models for complexity may not be suited for testing hypotheses about causation (Arif and MacNeil 2022). So, AIC may often not offer the weight we want to use, and we may also not wish to just generate an average at all. Instead, if we hope to understand an extensive universe of possible modelling outcomes, we could conduct a multiverse analysis, possibly with a specification curve (Simonsohn, Simmons, and Nelson 2015, 2020). This could mean running hundreds or thousands of models (or more!) to examine the distribution of possible effects, and to see how different model specification choices map onto these effects. However, exploring large areas of specification space may come at the cost of including biologically implausible specifications. Thus, we expect a trade-off, and attempts to limit models to the most biologically plausible may become increasingly difficult in proportion to the number of variables and modeling choices. To make selecting plausible models easier, one could recruit multiple analysts to design one or a few plausible specifications each as with our ‘many analyst’ study (Silberzahn et al. 2018). An alternative that may be more labor intensive for the primary analyst, but which may lead to a more plausible set of models, could involve hypothesizing about causal pathways with DAGs [directed acyclic graphs; Arif and MacNeil (2023)] to constrain the model set. As with other options outlined above, generating model specifications with DAGs could be partnered with pre-registration to hinder bias from undisclosed data dredging.

Responses to heterogeneity in analysis outcomes need not be limited to simply conducting more analyses, especially if it turns out that analysis quality drives some of the observed heterogeneity. As we noted above, we cannot yet rule out the possibility that insufficient statistical expertise or poor-quality analyses might drive some portion of the heterogeneity we observed. Improving the quality of analyses might be accomplished with a deliberate increase in investment in statistical education. Many ecology and evolutionary biology students learn their statistical practice informally, with many ecology doctoral programs in the USA not requiring a statistics course (Touchon and McCoy 2016), and no formal courses of any kind included in doctoral degrees in most other countries. In cases where formal investment in statistical education is lacking, informal resources, such as guidelines and checklists, may help researchers avoid common mistakes. However, unless following guidelines or checklists is enforced for publication, the adherence to guidelines is patchy. For example, despite the publication of guidelines for conducting meta-analyses in ecology, the quality of meta-analyses did not improve substantially over time (Koricheva and Gurevitch 2014). Even in medical research where adherence to guidelines such as the PRISMA standards for systematic reviews and meta-analyses is more highly valued, adherence is often poor (Page and Moher 2017).

Although we have reviewed a variety of potential responses to the existence of variability in analytical outcomes, we certainly do not wish to imply that this is a comprehensive set of possible responses. Nor do we wish to imply that the opinions we have expressed about these options are correct. Determining how the disciplines of ecology and evolutionary biology should respond to knowledge of the variability in analytical outcome will benefit from the contribution and discussion of ideas from across these disciplines. We look forward to learning from these discussions and to seeing how these disciplines ultimately respond.

Conclusions

Overall, our results suggest to us that, where there is a diverse set of plausible analysis options, no single analysis should be considered a complete or reliable answer to a research question. Further, because of the evidence that ecologists and evolutionary biologists often present a biased subset of the analyses they conduct (Deressa et al. 2023; Yang et al. 2023; Kimmel, Avolio and Ferraro 2023), we do not expect that even a collection of different effect sizes from different studies will accurately represent the true distribution of effects (Yang et al. 2023). Therefore, we believe that an increased level of skepticism of the outcomes of single analyses, or even single meta-analyses, is warranted going forward. We recognize that some researchers have long maintained a healthy level of skepticism of individual studies as part of sound and practical scientific practice, and it is possible that those researchers will be neither surprised nor concerned by our results. However, we doubt that many researchers are sufficiently aware of the potential problems of analytical flexibility to be appropriately skeptical. We hope that our work leads to conversations in ecology, evolutionary biology, and other disciplines about how best to contend with heterogeneity in results that is attributable to analytical decisions.

Declarations

Ethics, consent and permissions

We obtained permission to conduct this research from the Whitman College Institutional Review Board (IRB). As part of this permission, the IRB approved the consent form (https://osf.io/xyp68/) that all participants completed prior to joining the study. The authors declare that they have no competing interests.

Availability of data and materials

All materials and data are archived and hosted on the OSF at https://osf.io/mn5aj/, including survey instruments and analyst / reviewer consent forms. The Evolutionary Ecology Data and Ecology and Conservation Data provided to analysts are available at https://osf.io/34fzc/ and https://osf.io/t76uy/ respectively. Data has been anonymised, and the non-anonymised data is stored on the project OSF within private components accessible to the lead authors.

We built an R package, ManyEcoEvo to conduct the analyses described in this study (Gould et al. 2023), which can be downloaded from https://github.com/egouldo/ManyEcoEvo/ to reproduce our analyses or replicate the analyses described here using alternate datasets. Data cleaning and preparation of analysis-data, as well as the analysis, is conducted in R (R Core Team 2024) reproducibly using the targets package (Landau 2021). This data and analysis pipeline is stored in the ManyEcoEvo package repository and its outputs are made available to users of the package when the library is loaded.

The full manuscript, including further analysis and presentation of results is written in Quarto (J. J. Allaire et al. 2024). The source code to reproduce the manuscript is hosted at https://github.com/egouldo/ManyAnalysts/, and the rendered version of the source code may be viewed at https://egouldo.github.io/ManyAnalysts/. All R packages and their versions used in the production of this manuscript are listed in the session info at Section 6.6.

Competing interests

The authors declare that they have no competing interests

Funding

EG’s contributions were supported by an Australian Government Research Training Program Scholarship, AIMOS top-up scholarship (2022) and Melbourne Centre of Data Science Doctoral Academy Fellowship (2021). FF’s contributions were supported by ARC Future Fellowship FT150100297.

Author’s contributions

HF, THP and FF conceptualized the project. PV provided raw data for *Eucalyptus* analyses and SG and THP provided raw data for blue tit analyses. DGH, HF and THP prepared surveys for collecting participating analysts and reviewer’s data. EG, HF, THP, PV, SN and FF planned the analyses of the data provided by our analysts and reviewers, EG, HF, and THP curated the data, EG and HF wrote the software code to implement the analyses and prepare data visualisations. EG ensured that analyses were documented and reproducible. THP and HF administered the project, including coordinating with analysts and reviewers. FF provided funding for the project. THP, HF, and EG wrote the manuscript. Authors listed alphabetically contributed analyses of the primary datasets or reviews of analyses. All authors read and approved the final manuscript.

References

Aczel, Balazs, Barnabas Szaszi, Gustav Nilsonne, Olmo R van den Akker, Casper J Albers, Marcel ALM van Assen, Jojanneke A Bastiaansen, et al. 2021. “Consensus-Based Guidance for Conducting and Reporting Multi-Analyst Studies.” *eLife* 10 (November). https://doi.org/10.7554/elife.72185.

Allaire, J. J., Charles Teague, Carlos Scheidegger, Yihui Xie, and Christophe Dervieux. 2024. “Quarto.” https://doi.org/10.5281/zenodo.5960048.

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2024. *rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Arif, S., & M. Aaron MacNeil. 2022. “Predictive models aren’t for causal inference.” *Ecology Letters* 25(8), 1741–1745. https://doi.org/10.1111/ele.14033

Arif, Suchinta, and M. Aaron MacNeil. 2023. “Applying the Structural Causal Model Framework for Observational Causal Inference in Ecology.” *Ecological Monographs* 93 (1): e1554. https://doi.org/https://doi.org/10.1002/ecm.1554.

Arnold, Jeffrey B. 2024. *ggthemes: Extra Themes, Scales and Geoms for “ggplot2”*. https://jrnold.github.io/ggthemes/.

Atkinson, Joe, Lars A. Brudvig, Max Mallen-Cooper, Shinichi Nakagawa, Angela T. Moles, and Stephen P. Bonser. 2022. “Terrestrial Ecosystem Restoration Increases Biodiversity and Reduces Its Variability, but Not to Reference Levels: A Global Meta-Analysis.” *Ecology Letters* 25 (7): 1725–37. https://doi.org/https://doi.org/10.1111/ele.14025.

Auspurg, Katrin, and Josef Brüderl. 2021. “Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the ‘Many Analysts, One Data Set’ Project.” *Socius* 7: 23780231211024421. https://doi.org/10.1177/23780231211024421.

Bartoń, Kamil. 2023. *MuMIn: Multi-Model Inference*.

Baselga, Andres, David Orme, Sebastien Villeger, Julien De Bortoli, Fabien Leprieur, Maxime Logez, Sara Martinez-Santalla, et al. 2023. *betapart: Partitioning Beta Diversity into Turnover and Nestedness Components*. https://CRAN.R-project.org/package=betapart.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *2015* 67 (1): 48. https://doi.org/10.18637/jss.v067.i01.

Blake, Kevin. 2022. *NatParksPalettes: Color Palettes Inspired by National Parks*. https://github.com/kevinsblake/NatParksPalettes.

Bolker, Ben, David Robinson, Dieter Menne, Jonah Gabry, Paul Buerkner, Chrisopher Hau, William Petry, et al. 2024. *broom.mixed: Tidying Methods for Mixed Models*. https://github.com/bbolker/broom.mixed.

Borenstein, Michael, Julian P. T. Higgins, Larry Hedges, and Hannah Rothstein. 2017. “Basics of Meta-Analysis: I^2 Is Not an Absolute Measure of Heterogeneity.” *Research Synthesis Methods* 8: 5–18. https://doi.org/10.1002/jrsm.1230.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. “Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams.” *Nature* 582 (7810): 84–88.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung H. V. Nguyen, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, et al. 2022. “Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty.” *Proceedings of the National Academy of Sciences* 119 (44): e2203150119. https://doi.org/10.1073/pnas.2203150119.

Briga, Michael, and Simon Verhulst. 2021. “Mosaic Metabolic Ageing: Basal and Standard Metabolic Rates Age in Opposite Directions and Independent of Environmental Quality, Sex and Life Span in a Passerine.” *Functional Ecology* 35 (5): 1055–68. https://doi.org/https://doi.org/10.1111/1365-2435.13785.

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. “glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling.” *The R Journal* 9 (2): 378–400. https://doi.org/10.32614/RJ-2017-066.

Buck, Robert J., John Fieberg, and Daniel J. Larkin. 2022. “The use of weighted averages of Hedges' d in meta-analysis: Is it worth it?” *Methods in Ecology and Evolution* 13 (5): 1093—1105. https://doi.org/10.1111/2041-210X.13818.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach*. Book. 2nd ed. New York: Springer-Verlag. https://doi.org/10.1007/b97636.

Cade, Brian S. 2015. “Model Averaging and Muddled Multimodel Inferences.” *Ecology* 96 (9): 2370–82. http://www.jstor.org.ezproxy.whitman.edu/stable/24702343.

Capilla-Lasheras, Pablo, Megan J. Thompson, Alfredo Sánchez-Tójar, Yacob Haddou, Claire J. Branston, Denis Réale, Anne Charmantier, and Davide M. Dominoni. 2022. “A Global Meta-Analysis Reveals Higher Variation in Breeding Phenology in Urban Birds Than in Their Non-Urban Neighbours.” *Ecology Letters* 25 (11): 2552–70. https://doi.org/10.1111/ele.14099.

Coretta, Stefano, Joseph V. Casillas, Simon Roessig, Michael Franke, Byron Ahn, Ali H. Al-Hoorie, Jalal Al-Tamimi, et al. 2023. “Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses.” *Advances in Methods and Practices in Psychological Science* 6 (3): 25152459231162567. https://doi.org/10.1177/25152459231162567.

Dancho, Matt, and Davis Vaughan. 2023. *Timetk: A Tool Kit for Working with Time Series*. https://CRAN.R-project.org/package=timetk.

DeKogel, C. H. 1997. “Long-Term Effects of Brood Size Manipulation on Morphological Development and Sex-Specific Mortality of Offspring.” *Journal of Animal Ecology* 66 (2): 167–78. <Go to ISI>://WOS:A1997WQ19600003.

Deressa, Teshome, David Stern, Jaco Vangronsveld, Jan Minx, Sebastien Lizin, Robert Malina, and Stephan Bruns. 2023. “More Than Half of Statistically Significant Research Findings in the Environmental Sciences Are Actually Not.” *EcoEvoRxiv*. https://doi.org/https://doi.org/10.32942/X24G6Z.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, et al. 2013. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” *Ecography* 36 (1): 27–46. https://doi.org/https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Fanelli, Daniele, Rodrigo Costas, and John P. A. Ioannidis. 2017. “Meta-Assessment of Bias in Science.” *Proceedings of the National Academy of Sciences* 114: 3714–19. https://doi.org/10.1073/pnas.1618569114.

Fanelli, Daniele, and John P. A. Ioannidis. 2013. “US Studies May Overestimate Effect Sizes in Softer Research.” *Proceedings of the National Academy of Sciences* 110 (37): 15031–36. https://doi.org/10.1073/pnas.1302997110.

Fidler, Fiona, Mark A. Burgman, Geoff Cumming, Robert Buttrose, and Neil Thomason. 2006. “Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology.” *Conservation Biology* 20 (5): 1539–44. https://doi.org/10.1111/j.1523-1739.2006.00525.x.

Fidler, Fiona, Yung En Chee, Bonnie C. Wintle, Mark A. Burgman, Michael A. McCarthy, and Ascelin Gordon. 2017. “Metaresearch for Evaluating Reproducibility in Ecology and Evolution.” *BioScience* 67 (3): 282–89. https://doi.org/10.1093/biosci/biw159.

Firke, Sam. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Forstmeier, Wolfgang, Eric-Jan Wagenmakers, and T. H. Parker. 2017. “Detecting and Avoiding Likely False-Positive Findings – a Practical Guide.” *Biological Reviews* 92: 1941–68. https://doi.org/10.1111/brv.12315.

Fraser, Hannah, Tim Parker, Shinichi Nakagawa, Ashley Barnett, and Fiona Fidler. 2018. “Questionable Research Practices in Ecology and Evolution.” *PLOS ONE* 13 (7): e0200303. https://doi.org/10.1371/journal.pone.0200303.

Gamer, Matthias, Jim Lemon, and Ian Fellows Puspendra Singh. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. https://www.r-project.org.

Gelman, Andrew, and Eric Loken. 2013. “The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No ‘Fishing Expedition’ or ‘p-Hacking’ and the Research Hypothesis Was Posited Ahead of Time.” *Department of Statistics, Columbia University*.

Gelman, Andrew, and David Weakliem. 2009. “Of Beauty, Sex, and Power.” *American Scientist* 97: 310–16.

Gould, Elliot, Hannah S. Fraser, Shinichi Nakagawa, and Timothy H. Parker. 2023. “ManyEcoEvo: Meta-Analyse Data from ManyAnalyst Style Studies.” Zenodo. https://doi.org/10.5281/zenodo.10046153.

Gould, Ellliot, Hannah S. Fraser, Shinichi Nakagawa, Timothy H. Parker. 2024. egouldo/ManyAnalysts: Manuscript Source Code for 'Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology.' Zenodo. https://doi.org/10.5281/zenodo.13850927. Version 2.0.2.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. “Multimodel Inference in Ecology and Evolution: Challenges and Solutions.” *Journal of Evolutionary Biology* 24 (4): 699–711. https://doi.org/doi:10.1111/j.1420-9101.2010.02210.x.

Harrell Jr, Frank E. 2024. *Hmisc: Harrell Miscellaneous*. https://hbiostat.org/R/Hmisc/.

Hester, Jim, Lionel Henry, Kirill Müller, Kevin Ushey, Hadley Wickham, and Winston Chang. 2024. *withr: Run Code "With" Temporarily Modified Global State*. https://withr.r-lib.org.

Higgins, Julian P T, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. 2003. "Measuring Inconsistency in Meta-Analyses." *BMJ* 327 (7414): 557–60. https://doi.org/10.1136/bmj.327.7414.557.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–60. https://doi.org/https://doi.org/10.1111/ecin.12992.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2024. *gt: Easily Create Presentation-Ready Display Tables*. https://gt.rstudio.com.

Jennions, M. D., C. J. Lortie, M. S. Rosenberg, and H. R. Rothstein. 2013. "Publication and Related Biases." Book Section. In *Handbook of Meta-Analysis in Ecology and Evolution*, edited by J. Koricheva, J. Gurevitch, and K. Mengersen, 207–36. Princeton, USA: Princeton University Press.

Kassambara, Alboukadel. 2023. *ggpubr: "ggplot2" Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/.

Kimmel, Kaitlin, Meghan L. Avolio, and Paul J. Ferraro. 2023. "Empirical Evidence of Widespread Exaggeration Bias and Selective Reporting in Ecology." *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-023-02144-3.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams Jr., Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. "Investigating Variation in Replicability: A "Many Labs" Replication Project." *Social Psychology* 45 (3): 142–52. https://doi.org/10.1027/1864-9335/a000178.

Klein, Richard A., Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, et al. 2018. "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1 (4): 443–90. https://doi.org/10.1177/2515245918810225.

Knight, K. 2000. *Mathematical Statistics*. Book. New York: Chapman; Hall.

Koricheva, Julia, and Jessica Gurevitch. 2014. "Uses and Misuses of Meta-Analysis in Plant Ecology." *Journal of Ecology* 102 (4): 828–44. https://doi.org/https://doi.org/10.1111/1365-2745.12224.

Kou-Giesbrecht, Sian, and Duncan N. L. Menge. 2021. "Nitrogen-Fixing Trees Increase Soil Nitrous Oxide Emissions: A Meta-Analysis." *Ecology* 102 (8): e03415. https://doi.org/https://doi.org/10.1002/ecy.3415.

Kuhn, Max, and Hannah Frick. 2022. *multilevelmod: Model Wrappers for Multi-Level Models*. https://github.com/tidymodels/multilevelmod.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. “lmerTest Package: Tests in Linear Mixed Effects Models.” *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13.

Landau, William Michael. 2021. “The Targets r Package: A Dynamic Make-Like Function-Oriented Pipeline Toolkit for Reproducibility and High-Performance Computing.” *Journal of Open Source Software* 6 (57): 2959. https://doi.org/10.21105/joss.02959.

Leybourne, Daniel J., Katharine F. Preedy, Tracy A. Valentine, Jorunn I. B. Bos, and Alison J. Karley. 2021. “Drought Has Negative Consequences on Aphid Fitness and Plant Vigor: Insights from a Meta-Analysis.” *Ecology and Evolution* 11 (17): 11915–29. https://doi.org/https://doi.org/10.1002/ece3.7957.

Lu, Xun, and Halbert White. 2014. “Robustness Checks and Robustness Tests in Applied Economics.” *Journal of Econometrics* 178: 194–206. https://doi.org/https://doi.org/10.1016/j.jeconom.2013.08.016.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. “Extracting, Computing and Exploring the Parameters of Statistical Models Using R.” *Journal of Open Source Software* 5 (53): 2445. https://doi.org/10.21105/joss.02445.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. “performance: An R Package for Assessment, Comparison and Testing of Statistical Models.” *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

Lüdecke, Daniel, Indrajeet Patil, Mattan S. Ben-Shachar, Brenton M. Wiernik, Philip Waggoner, and Dominique Makowski. 2021. “see: An R Package for Visualizing Statistical Models.” *Journal of Open Source Software* 6 (64): 3393. https://doi.org/10.21105/joss.03393.

Luke, S. G. 2017. “Evaluating Significance in Linear Mixed-Effects Models in r.” *Behavior Research Methods* 49 (4): 1494–1502.

Makowski, Dominique, Mattan S. Ben-Shachar, Indrajeet Patil, and Daniel Lüdecke. 2020. “Estimation of Model-Based Predictions, Contrasts and Means.” *CRAN*. https://github.com/easystats/modelbased.

Masur, Philipp K., and Michael Scharkow. 2020. “specr: Conducting and Visualizing Specification Curve Analyses (Version 1.0.0).” https://CRAN.R-project.org/package=specr.

Meschiari, Stefano. 2022. *Latex2exp: Use LaTeX Expressions in Plots*. https://www.stefanom.io/latex2exp/.

Miles, C. 2008. “Testing Market-Based Instruments for Conservation in Northern Victoria.” Book Section. In *Biodiversity: Integrating Conservation and Production: Case Studies from Australian Farms, Forests and Fisheries*, edited by T. Norton, T. Lefroy, K. Bailey, and G. Unwin, 133–46. Melbourne, Australia: CSIRO Publishing.

Millard, Steven P. 2013. *EnvStats: An r Package for Environmental Statistics*. New York: Springer. https://www.springer.com.

Molina, Isabel, and Yolanda Marhuenda. 2015. “sae: An R Package for Small Area Estimation.” *The R Journal* 7 (1): 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf.

Morrissey, Michael B., and Graeme D. Ruxton. 2018. “Multiple Regression Is Not Multiple Regressions: The Meaning of Multiple Regression and the Non-Problem of Collinearity.” *Philosophy, Theory, and Practice in Biology* 10 (3). https://doi.org/10.3998/ptpbio.16039257.0010.003.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Nakagawa, Shinichi, and Innes C. Cuthill. 2007. “Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists.” *Biological Reviews* 82 (4): 591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x.

Nakagawa, Shinichi, Malgorzata Lagisz, Michael D. Jennions, Julia Koricheva, Daniel W. A. Noble, Timothy H. Parker, Alfredo Sánchez-Tójar, Yefeng Yang, and Rose E. O’Dea. 2022. “Methods for Testing Publication Bias in Ecological and Evolutionary Meta-Analyses.” *Methods in Ecology and Evolution* 13 (1): 4–21. https://doi.org/https://doi.org/10.1111/2041-210X.13724.

Nakagawa, Shinichi, Malgorzata Lagisz, Rose E. O’Dea, Patrice Pottier, Joanna Rutkowska, Alistair M. Senior, Yefeng Yang, and Daniel W. A. Noble. 2023. “orchaRd 2.0: An r Package for Visualizing Meta-Analyses with Orchard Plots.” *EcoEvoRxiv* 12: 4–12. https://doi.org/https://doi.org/10.32942/X2QC7K.

Nakagawa, Shinichi, Yefeng Yang, Erin L. Macartney, Rebecca Spake, and Malgorzata Lagisz. 2023. “Quantitative Evidence Synthesis: A Practical Guide on Meta-Analysis, Meta-Regression, and Publication Bias Tests for Environmental Sciences.” *Environmental Evidence* 12 (1): 8. https://doi.org/10.1186/s13750-023-00301-6.

Nakagawa, S., D. W. Noble, A. M. Senior, and M. Lagisz. 2017. “Meta-Evaluation of Meta-Analysis: Ten Appraisal Questions for Biologists.” *BMC Biology* 15 (1): 18. https://doi.org/10.1186/s12915-017-0357-7.

Nicolaus, M., S. P. M. Michler, R. Ubels, M. van der Velde, J. Komdeur, C. Both, and J. M. Tinbergen. 2009. “Sex-Specific Effects of Altered Competition on Nestling Growth and Survival: An Experimental Manipulation of Brood Size and Sex Ratio.” *Journal of Animal Ecology* 78 (2): 414–26. https://doi.org/10.1111/j.1365-2656.2008.01505.x.

Noble, Daniel W. A., Malgorzata Lagisz, Rose E. O’Dea, and Shinichi Nakagawa. 2017. “Nonindependence and Sensitivity Analyses in Ecological and Evolutionary Meta-Analyses.” *Molecular Ecology* 26 (9): 2410–25. https://doi.org/10.1111/mec.14031.

O’Hara, Robert B., and D. Johan Kotze. 2010. “Do Not Log-Transform Count Data.” *Methods in Ecology and Evolution* 1 (2): 118–22. https://doi.org/10.1111/j.2041-210x.2010.00021.x.

Open Science Collaboration. 2015. “Estimating the Reproducibility of Psychological Science.” *Science* 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716.

Page, Matthew J., and David Moher. 2017. “Evaluations of the Uptake and Impact of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement and Extensions: A Scoping Review.” *Systematic Reviews* 6 (1): 263. https://doi.org/10.1186/s13643-017-0663-8.

Parker, Timothy H., Wolfgang Forstmeier, Julia Koricheva, Fiona Fidler, Jarrod D. Hadfield, Yung En Chee, Clint D. Kelly, Jessica Gurevitch, and Shinichi Nakagawa. 2016. “Transparency in Ecology and Evolution: Real Problems, Real Solutions.” *Trends in Ecology & Evolution* 31 (9): 711–19. https://doi.org/10.1016/j.tree.2016.07.002.

Parker, Timothy H., and Yefeng Yang. 2023. "Exaggerated Effects in Ecology." *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-023-02156-z.

Pedersen, Thomas Lin. 2024. *patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com.

Pei, Yifan, Wolfgang Forstmeier, Daiping Wang, Katrin Martin, Joanna Rutkowska, and Bart Kempenaers. 2020. "Proximate Causes of Infertility and Embryo Mortality in Captive Zebra Finches." *The American Naturalist* 196 (5): 577–96. https://doi.org/10.1086/710956.

Qiu, Yixuan. 2024. *showtext: Using Fonts More Easily in r Graphs*. https://github.com/yixuan/showtext.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosenberg, M. S. 2013. "Moment and Least-Squares Based Approaches to Metaanalytic Inference." Book Section. In *Handbook of Meta-Analysis in Ecology and Evolution*, edited by J. Koricheva, J. Gurevitch, and K. Mengersen, 108–24. Princeton, USA: Princeton University Press.

Royle, N. J., I. R. Hartley, I. P. F. Owens, and G. A. Parker. 1999. "Sibling Competition and the Evolution of Growth Rates in Birds." *Proceedings of the Royal Society B-Biological Sciences* 266 (1422): 923–32. https://doi.org/10.1098/rspb.1999.0725.

Scheinin, Ilari, Maria Kalimeri, Vilma Jagerroos, Juuso Parkkinen, Emmi Tikkanen, Peter Würtz, and Antti Kangas. 2020. *ggforestplot: Forestplots of Measures of Effects and Their Confidence Intervals*. https://github.com/NightingaleHealth/ggforestplot.

Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2024. *GGally: Extension to "ggplot2"*. https://ggobi.github.io/ggally/.

Schweinsberg, M., M. Feldman, N. Staub, O. R. van den Akker, R. C. M. van Aert, Malm van Assen, Y. Liu, et al. 2021. "Same Data, Different Conclusions: Radical Dispersion in Empirical Results When Independent Analysts Operationalize and Test the Same Hypothesis." *Organizational Behavior and Human Decision Processes* 165: 228–49. https://doi.org/10.1016/j.obhdp.2021.02.003.

Senior, Alistair M., Catherine E. Grueber, Tsukushi Kamiya, Malgorzata Lagisz, Katie O'Dwyer, Eduardo S. A. Santos, and Shinichi Nakagawa. 2016. "Heterogeneity in Ecological and Evolutionary Meta-Analyses: Its Magnitude and Implications." *Ecology* 97 (12): 3293–99. https://doi.org/10.1002/ecy.1591.

Shavit, A., and Aaron M. Ellison. 2017. *Stepping in the Same River Twice: Replication in Biological Research*. Edited Book. New Haven, Connecticut, USA: Yale University Press.

Siegel, Kyle R., Muskanjot Kaur, A. Calvin Grigal, Rebecca A. Metzler, and Gary H. Dickinson. 2022. "Meta-Analysis Suggests Negative, but pCO2-Specific, Effects of Ocean Acidification on the Structural and Functional Properties of Crustacean Biomaterials." *Ecology and Evolution* 12 (6): e8922. https://doi.org/https://doi.org/10.1002/ece3.8922.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56. https://doi.org/10.1177/2515245917747646.

Silge, Julia, and David Robinson. 2016. “tidytext: Text Mining and Analysis Using Tidy Data Principles in r.” *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Simons, Daniel J., Yuichi Shoda, and D. Stephen Lindsay. 2017. “Constraints on Generality (COG): A Proposed Addition to All Empirical Papers.” *Perspectives on Psychological Science*. https://doi.org/10.1177/174569161770863.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2015. “Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications.” Manuscript. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998 .

———. 2020. “Specification Curve Analysis.” *Nature Human Behaviour* 4 (11): 1208–14. https://doi.org/10.1038/s41562-020-0912-z.

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. “Reproducible Summary Tables with the Gtsummary Package.” *The R Journal* 13: 570–80. https://doi.org/10.32614/RJ-2021-053.

Slowikowski, Kamil. 2024. *ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2”*. https://ggrepel.slowkow.com/.

Stanton-Geddes, John, Cintia Gomes de Freitas, and Cristian de Sales Dambros. 2014. “In Defense of p Values: Comment on the Statistical Methods Actually Used by Ecologists.” *Ecology* 95 (3): 637–42. https://doi.org/https://doi.org/10.1890/13-1156.1.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. “Increasing Transparency Through a Multiverse Analysis.” *Perspectives on Psychological Science* 11 (5): 702–12. https://doi.org/10.1177/1745691616658637.

Taylor, James W., and Kathryn S. Taylor. 2023. “Combining Probabilistic Forecasts of COVID-19 Mortality in the United States.” *European Journal of Operational Research* 304 (1): 25–41. https://doi.org/https://doi.org/10.1016/j.ejor.2021.06.044.

Tierney, Nicholas, and Dianne Cook. 2023. “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations.” *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

Touchon, Justin C., and Michael W. McCoy. 2016. “The Mismatch Between Current Statistical Practice and Doctoral Training in Ecology.” *Ecosphere* 7 (8): e01394. https://doi.org/https://doi.org/10.1002/ecs2.1394.

Ushey, Kevin, and Hadley Wickham. 2023. *renv: Project Environments*. https://rstudio.github.io/renv/.

van den Brand, Teun. 2024. *Ggh4x: Hacks for “ggplot2”*. https://github.com/teunbrand/ggh4x.

Vander Werf, Eric. 1992. “Lack’s Clutch Size Hypothesis: An Examination of the Evidence Using Meta-Analysis.” *Ecology* 73 (5): 1699–1705. https://doi.org/10.2307/1940021.

Ver Hoef, Jay M. 2012. “Who Invented the Delta Method?” *The American Statistician* 66 (2): 124–27. https://doi.org/10.1080/00031305.2012.687494.

Verhulst, S., M. J. Holveck, and K. Riebel. 2006. "Long-Term Effects of Manipulated Natal Brood Size on Metabolic Rate in Zebra Finches." *Biology Letters* 2 (3): 478–80. https://doi.org/10.1098/rsbl.2006.0496.

Vesk, P. A., W. K. Morris, W. McCallum, R. Apted, and C. Miles. 2016. "Processes of Woodland Eucalypt Regeneration: Lessons from the Bush Returns Trial." *Proceedings of the Royal Society of Victoria* 128: 54–63.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the metafor Package." *Journal of Statistical Software* 36 (3): 1–48. https://doi.org/10.18637/jss.v036.i03.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *devtools: Tools to Make Developing r Packages Easier*. https://devtools.r-lib.org/.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *scales: Scale Functions for Visualization*. https://scales.r-lib.org.

Wilke, Claus O. 2024. *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"*. https://wilkelab.org/cowplot/.

Xie, Yihui. 2024a. *knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2024b. *xfun: Supporting Functions for Packages Maintained by "Yihui Xie"*. https://github.com/yihui/xfun.

Yang, Yefeng, Alfredo Sánchez-Tójar, Rose E. O'Dea, Daniel W. A. Noble, Julia Koricheva, Michael D. Jennions, Timothy H. Parker, Malgorzata Lagisz, and Shinichi Nakagawa. 2023. "Publication Bias Impacts on Effect Size, Statistical Power, and Magnitude (Type m) and Sign (Type s) Errors in Ecology and Evolutionary Biology." *BMC Biology* 21 (1): 71. https://doi.org/10.1186/s12915-022-01485-y.

Zeileis, Achim, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. 2020. "colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes." *Journal of Statistical Software* 96 (1): 1–49. https://doi.org/10.18637/jss.v096.i01.

R Package References and Session Information

Table 7: R packages used to generate this manuscript. Please see the ManyEcoEvo:: package for a full list of packages used in the analysis pipeline.

Package	Version	Citation
base	4.4.0	R Core Team (2024)
betapart	1.6	Baselga et al. (2023)
broom.mixed	0.2.9.5	Bolker et al. (2024)
colorspace	2.1.0	Zeileis et al. (2020)
cowplot	1.1.3	Wilke (2024)

devtools	2.4.5	Wickham et al. (2022)
EnvStats	2.8.1	Millard (2013)
GGally	2.2.1	Schloerke et al. (2024)
ggforestplot	0.1.0	Scheinin et al. (2020)
ggh4x	0.2.8	van den Brand (2024)
ggpubr	0.6.0	Kassambara (2023)
ggrepel	0.9.5	Slowikowski (2024)
ggthemes	5.1.0	Arnold (2024)
glmmTMB	1.1.8	Brooks et al. (2017)
gt	0.10.1	Iannone et al. (2024)
gtsummary	1.7.2	Sjoberg et al. (2021)
here	1.0.1	Müller (2020)
Hmisc	5.1.2	Harrell Jr (2024)
irr	0.84.1	Gamer, Lemon, and Singh (2019)
janitor	2.2.0	Firke (2023)
knitr	1.46	Xie (2024a)
latex2exp	0.9.6	Meschiari (2022)
lme4	1.1.35.3	Bates et al. (2015)
ManyEcoEvo	2.7.6	Gould et al. (2023)
metafor	4.6.0	Viechtbauer (2010)
modelbased	0.8.7	Makowski et al. (2020)
multilevelmod	1.0.0	Kuhn and Frick (2022)
MuMIn	1.47.5	Bartoń (2023)
naniar	1.1.0	Tierney and Cook (2023)
NatParksPalettes	0.2.0	Blake (2022)
orchaRd	2	Nakagawa, Lagisz, et al. (2023)
parameters	0.21.7	Lüdecke et al. (2020)
patchwork	1.2.0	Pedersen (2024)
performance	0.11.0	Lüdecke, Ben-Shachar, et al. (2021)
renv	1.0.2	Ushey and Wickham (2023)
rmarkdown	2.27	Allaire et al. (2024)
sae	1.3	Molina and Marhuenda (2015)
scales	1.3.0	Wickham, Pedersen, and Seidel (2023)
see	0.8.4	Lüdecke, Patil, et al. (2021)
showtext	0.9.7	Qiu (2024)
specr	1.0.0	Masur and Scharkow (2020)
targets	1.7.0	Landau (2021)
tidymodels	1.1.1	Kuhn and Wickham (2020)
tidytext	0.4.2	Silge and Robinson (2016)
tidyverse	2.0.0	Wickham et al. (2019)
withr	3.0.0	Hester et al. (2024)
xfun	0.44	Xie (2024b)

1982

```
— Session info ————————————————————————————————————————————————————
 setting  value
 version  R version 4.4.0 (2024-04-24)
 os       macOS Ventura 13.6.9
 system   aarch64, darwin20
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       Australia/Melbourne
 date     2024-09-17
 pandoc   3.1.12.2 @ /opt/homebrew/bin/ (via rmarkdown)
```