Ecología Austral 31:065-070 Abril 2021
Asociación Argentina de Ecología

Reproducibilidad y replicabilidad en la investigación en ciencias naturales: ¿Hay una crisis?

David E. Gorla

Instituto de Diversidad y Ecología Animal (Conicet - Universidad Nacional de Córdoba).

Resumen. Suponemos que por la propiedad de autocorrección del conocimiento científico, todo estudio publicado que contenga errores será revisado por alguien que replicará aquel estudio y, eventualmente, publicará la corrección. De este modo el proceso descartará errores y conservará aciertos para construir el conocimiento científico aceptado como válido. Sin embargo, debido a prácticas cuestionables de investigación el proceso no funciona tal como se espera y la situación ha llevado a una crisis de reproducibilidad y replicabilidad (aun cuando hay quienes niegan tal crisis). Este trabajo considera, en general, el estado de situación y presenta un ejemplo local (al menos como prueba de concepto) que muestra que las prácticas cuestionables de investigación que causan problemas de reproducibilidad y replicabilidad en investigaciones científicas son más prevalentes que lo que suponemos. Concluyo con un llamado a crear espacios para reflexionar sobre el problema y enumero algunas ideas para reducir el impacto de sus causas, de importancia particular para la formación de nuevos investigadores.

[Palabras clave: prácticas cuestionables de investigación, causas de no-replicabilidad]

Abstract. Reproducibility and replicability in natural science research: Is there a crisis? We assume that because of the self-correction nature of the scientific knowledge, every published study will be revised by somebody that will replicate it and, eventually, publish a correction. Therefore, the process will discard errors and keep successes to build the established scientific knowledge. However, because of questionable research practices, the process does not work as well as expected, and has driven a reproducibility and replicability crisis (although some will deny such crisis). This document considers in general the state-of-the-art of the problem and presents a local example (at least as a proof of concept) showing that questionable research practices are more prevalent than we think. I conclude calling for spaces to reflect on the problem and I suggest ideas to reduce the impact of its causes, especially important for the training of young researchers.

[Keywords: questionable research practices, non-replicability causes]

Introducción

Desde que Ioannidis (2005) escribió el muy citado artículo (8597 veces según Google Scholar en julio 2020) sobre por qué la mayoría de los hallazgos científicos son falsos, e insistió 11 años más tarde al afirmar que la mayor parte de la investigación clínica no es útil (Ioannidis 2016), el debate acerca de los problemas de reproducibilidad y replicabilidad de los estudios científicos se intensificó. Durante los últimos 10 años se publicó una profusa bibliografía, y muchas agencias científicas gubernamentales y no gubernamentales priorizaron el problema en sus agendas institucionales. La Academia de Ciencias de los Estados Unidos de Norteamérica, por ejemplo, emitió recientemente un documento fijando posición (NAS 2019). Al mismo tiempo surgieron numerosas iniciativas que proponen el abordaje del problema desde el nivel individual hasta el institucional, incluyendo la industria de las publicaciones científicas (en especial luego de la aparición de los llamados *journals depredadores* y sonados casos como el del artículo "Get me off your fucking mailing list" [Mazieres and Kohler 2014]).

Reproducibilidad y replicabilidad (y muchos otros términos emparentados) se usan en diferentes contextos y con diferentes interpretaciones. En el documento de la Academia de Ciencias de los Estados Unidos (NAS 2019) se proponen definiciones para ambos términos. La reproducibilidad es la propiedad por la que se obtienen resultados consistentes utilizando los mismos datos, pasos computacionales, métodos, código y condiciones de análisis. Por su parte, la replicabilidad hace referencia a la obtención de resultados consistentes en diferentes estudios que apuntan a responder la misma pregunta científica, cada uno de los cuales obtuvo sus propios datos. Suponemos que el conocimiento científico se autocorrige y que, aunque algunos resultados publicados no serán replicables, esos resultados no serán sumados al cuerpo de conocimiento pues se

Editora asociada: María Semmartin
✉ david.gorla@conicet.gov.ar

Recibido: 2 de Noviembre de 2019
Aceptado: 9 de Octubre de 2020

espera que los estudios posteriores demuestren la falsedad del primero por aquello de que "algo es cierto hasta que alguien demuestre lo contrario". Una buena parte del problema actual es que la fracción de estudios publicados no replicables es abrumadoramente elevada. Lo no replicable erróneo cuesta recursos a los sistemas científicos, aunque no todo es blanco y negro, y lo no replicable no necesariamente es siempre incorrecto.

Causas de no-replicabilidad

Los resultados de una investigación científica pueden no ser replicables por causas positivas o negativas. Las causas positivas incluyen la complejidad del sistema bajo estudio, la comprensión del número y relaciones entre variables dentro del sistema bajo estudio, la posibilidad de controlar variables, el cociente ruido/señal, entre otras. Entre las causas negativas se incluyen diseños de estudios deficientes debido a problemas tales como no reconocer o ajustar sesgos conocidos, no seguir buenas prácticas de aleatorización, obtener una baja potencia debido a tamaños muestrales inadecuados, confundir datos durante su manipulación, fallar en la caracterización y tener en cuenta incertidumbres conocidas, presentar una descripción metodológica incompleta. También son causas negativas los errores debidos al desconocimiento de buenas prácticas, a la publicación selectiva de resultados 'significativos' en detrimento de los resultados 'no significativos', errores simples, investigación descuidada, sesgo inconsciente hacia resultados específicos, incentivos espurios (por ejemplo basados en el número de artículos publicados y cantidad de proyectos aprobados), inferencia estadística inapropiada (e.g., *P-hacking*, *cherry-picking*, *HARKing*) y fraude (Tabla 1).

Tabla 1. Prácticas cuestionables de inferencia estadística que frecuentemente producen falta de replicabilidad.

Table 1. Questionable statistical inference practices that often result in lack of replicability.

Práctica	Definición
P-hacking y *cherry-picking*	Práctica de colectar, seleccionar o analizar datos hasta que se encuentre un resultado estadísticamente significativo; analizar muchas relaciones diferentes y sólo reportar aquellas para las que $P<0.05$
HARKing (del inglés *Hypothesis After Results are Known*)	Investigación confirmatoria que construye las hipótesis después que los datos fueran colectados y luego usa los mismos datos como evidencia para apoyar la hipótesis

Si bien no es un problema nuevo, el uso inapropiado de los *P*-valores (probabilidad de rechazar la hipótesis nula cuando es cierta) para indicar significación estadística está especialmente presente en la discusión durante la última década y está asociado a los problemas de replicabilidad (NAS 2016). Tan fuerte fue el impacto de las discusiones que la Asociación de Estadística de los Estados Unidos emitió una declaración acerca de los *P*-valores en un editorial (Wasserstein and Lazar 2016), y más recientemente, The American Statistician publicó un número especial titulado "Moving to a World Beyond $P<0.05$", incluyendo 45 artículos que elaboran consideraciones acerca de la cuestión (Wasserstein et al. 2019). Durante 2019, la revista Nature publicó un artículo firmado por más de 800 autores con el pedido de "jubilar la significancia estadística", afirmando que en la medida en que la significancia estadística se use menos, el pensamiento estadístico se usará más (Amrhein et al. 2019). ¿Debemos subirnos a la ola y tirar el *P*-valor a la basura? No, pero no lo podemos usar para tomar decisiones que deberían ser tomadas sobre la base de la biología y no de la estadística.

La publicación selectiva es una de las causas que más afectan a la replicabilidad. Debido a que las revistas científicas buscan 'la novedad', raramente publican artículos que informan resultados no significativos o que replican otros estudios. La selección de los autores o la decisión editorial, muchas veces orientada por una afición por usar una técnica numérica sofisticada (snobismo metodológico), puede producir un sesgo hacia la adopción de la creencia de que el efecto de un factor es significativo, cuando tal vez la mayoría de los estudios muestren lo contrario. Como lo único que aparece en la bibliografía son estudios que produjeron pruebas 'significativas', dicho sesgo en el 'estado del arte' queda establecido. El fenómeno ya recibió un nombre propio y se llama *Efecto chrysalis*, que explica cómo "resultados iniciales poco claros se transforman en elegantes artículos" (O'Boyle et al. 2014).

Una buena parte de la razón por la que el problema apareció como tema de discusión está vinculada a las muy bajas tasas de replicabilidad de ensayos clínicos, hecho

que promovió el cuestionamiento de la industria farmacéutica sobre la calidad de la investigación académica. Las evaluaciones de replicabilidad en estas áreas mostraron que hasta un 70% de los resultados publicados no se pudieron replicar, tanto por investigadores intentando replicar hallazgos de otros como por investigadores tratando de reproducir sus propios hallazgos publicados (Baker et al. 2016). Los problemas de reproducibilidad y replicabilidad fueron y son motivo de profunda discusión en la investigación en psicología, tal como muestran los visitados videos de Cummings (2013, 2017) "Dance of the *P*-values" y "Significance roulette", y la decisión del Journal of Basic and Applied Social Psychology, que prohibió la publicación de métodos basados en pruebas de significación de la hipótesis nula. ¿Debería esa situación gatillar alarmas para evaluar el estado del problema en ámbitos más afines de nuestras investigaciones en ciencias naturales? La reproducibilidad puede sonar al principio como una tarea trivial, pero la experiencia mostró que no es siempre fácil alcanzar esto que parece un estándar mínimo (NAS 2019).

Tomé contacto con el problema a partir de la discusión sobre el uso de los *P*-valores y la llamada crisis de las pruebas de significación de la hipótesis nula (NHST). El uso inapropiado de los *P*-valores es una de las causas de no replicabilidad y frecuentemente es consecuencia de un comportamiento automatizado que toma como umbral de decisión sobre la validez de una hipótesis el 'mágico' valor de probabilidad 0.05 para construir dos universos en los que viven y mueren las hipótesis nulas (H0). Un valor de P=0.049 significa que 'tranquilamente' se debe rechazar la H0 (en especial si esa decisión coincide con lo que uno espera que suceda), sin importar otro 'detalle' (que habitualmente tiene que ver con mecanismos explicativos). Incluso para justificar el rechazo de una H0 no es extraño leer que P=0.051 está en el borde de la significación, o que se aproxima a la significación. ¿Cómo sabe uno que no se está alejando raudamente de la significación? La ejecución de múltiples pruebas de hipótesis debería llevar a corregir los niveles de significación, como la corrección de Bonferroni, que indica que hay que dividir el nivel de significación deseado por el número de hipótesis que se prueban. Es decir que si uno 'probara' 10 hipótesis debería usar un valor de P=0.005 (0.05/10).

Pero, tal como argumentan Leek y Peng (2015), los problemas asociados con el mal uso de los *P*-valores son apenas la punta del iceberg del problema. El problema completo está vinculado con el llamado ciclo de la investigación científica (del inglés *scientific workflow*), que comienza con la identificación de un problema y, eventualmente, termina en una publicación, pasando por la secuencia pregunta-objetivo/hipótesis-diseño (selección de métodos de análisis)-colecta de datos-análisis de datos-interpretación-comunicación.

Muchas de las iniciativas que abordan los problemas de replicabilidad en la investigación están enmarcadas dentro del movimiento Ciencia Abierta. Por ejemplo, el Center for Open Science (http://cos.io) (que mantiene la iniciativa del *Open Science Framework* [http://osf.io]) o la *UK Reproducibility Network* (http://www.bristol.ac.uk/psychology/research/ukrn/), y propuestas vinculadas (Munafo et al. 2017; Steward-Lowndes et al. 2017).

¿Están todas las disciplinas igualmente afectadas por los problemas de replicabilidad? Sí, con variaciones. Una encuesta entre 1576 lectores de Nature mostró que los físicos y los químicos fueron quienes declararon una mayor replicabilidad en sus estudios, y que entre el 50 y 60% de los consultados declaró problemas para reproducir sus propios estudios en cualquiera de las disciplinas (química, física, biología, ciencias de la tierra y ambiente, medicina) (Baker et al. 2016). Ante el requerimiento de opinión de colegas en mi ámbito de trabajo, un argumento frecuente refirió a las dificultades de evaluar la replicabilidad en estudios en ecología debido a la complejidad de los sistemas en estudio. Aunque no es fácil evaluar la replicabilidad en estudios ecológicos, un artículo reciente (Fraser et al. 2018) muestra que existe una elevada prevalencia en el uso de prácticas cuestionables de investigación en trabajos sobre ecología y evolución, que aumentan el número de falsos positivos en la literatura (=informar efectos significativos cuando no existen), lo cual disminuye la replicabilidad.

Es interesante el hecho de que en nuestros seminarios internos de discusión sobre reproducibilidad y replicabilidad celebrados recientemente, estos conceptos se asocian más a etéreas cuestiones epistemológicas que al quehacer cotidiano de un investigador.

Investigadores jóvenes consideran que pensar sobre o resolver estos problemas es un 'lujo' que pueden darse investigadores maduros (*sensu* Farji-Brener and Ruggiero 2010), que ya tienen su carrera asegurada. Sin embargo, la presión por la publicación a las que están sometidos los investigadores, sumada a la falta de reflexión sobre inferencia estadística, aumenta las chances de falta de replicabilidad en algunos estudios.

Un ejemplo con sabor local

Entre 2017 y 2018 se presentaron 92 tesinas de grado en la Escuela de Biología de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de Córdoba. La tesina es un trabajo de investigación científica original, dentro del área de las Ciencias Biológicas (Reglamento Tesinas Res 136 HCD 2013); debe ser aprobada para obtener el título de grado. Con el objeto de indagar acerca de las características de dichas tesinas en el contexto de ciencia abierta envié una encuesta estructurada a los 94 directores de tesinas (Tabla 2). De las 31 respuestas recibidas (33%), 93% (29/31) de las tesinas colectaron datos cuantitativos, lo que sugiere un papel importante del análisis cuantitativo en la interpretación de los resultados. La mitad (16/31) fue declarada como una mezcla de estudios exploratorios (=descriptivos) y confirmatorios (=inferenciales), y sólo 16% (5/31) se declaró como exclusivamente exploratorio.

Aunque 65% (20/31) declaró adhesión al concepto de datos abiertos, sólo 3% (1/31) declaró que los datos están disponibles en el dominio público. La Universidad Nacional de Córdoba tiene una Oficina de Conocimiento Abierto (http://oca.uc.edu.ar), aunque es de reciente creación. El 7% (2/29) declaró que tuvo dificultades para re-usar los datos de la tesina. Aunque no hay precisiones sobre la naturaleza de las dificultades, los dos casos podrían representar problemas de reproducibilidad computacional, en la que no es posible reconstruir la secuencia de pasos usados en el análisis para obtener los resultados originales.

El 93% (29/31) de los directores continúa desarrollando estudios en la línea de la tesina dirigida entre 2017-2018. Sin embargo, sólo 32% (10/31) declara haber realizado una publicación con base en la tesina (o enviado a su publicación).

De las 29 respuestas que indicaron que en la tesina se usó alguna técnica estadística que selecciona variables según su nivel de significación (por ejemplo: regresión por pasos [Wittingham et al. 2006]), 24% (7/29) respondió positivamente (*P-hacking*). El 24% (7/29) declaró que en el estudio se obtuvieron resultados estadísticamente no significativos que no se incluyeron en la tesina (publicación selectiva). El 21% (6/29) cambió a otro tipo de análisis estadístico después de que el análisis original falló en mostrar significación estadística (*cherry-picking*).

Aunque la encuesta es muy limitada en estructura y tamaño de muestra, sugiere que la situación percibida en publicaciones de circulación internacional tiene un correlato en nuestro ámbito local. Dada la importancia de estos aspectos sobre la formación de nuevos científicos y sobre el conocimiento producido en este contexto sería muy conveniente que instituciones académicas convocaran a una discusión abierta sobre las cuestiones planteadas. Por mi lado, confieso que todos los errores que se mencionan en este texto los cometí al menos una vez (posiblemente más de una vez).

Existen acciones que se pueden realizar a diferentes niveles con el objeto de minimizar los problemas vinculados a la falta de reproducibilidad y replicabilidad. Las acciones van desde lo individual a lo institucional. Sin pretender una cobertura exhaustiva de posibilidades, incluyo algunas realizables, al menos como para abrir el debate. 1) Revisar comportamientos automatizados en el análisis de datos. Por ejemplo, seleccionar variables en modelos estadísticos con base en umbrales de significación $P<0.05$. Más que dividir el mundo por el valor 0.05, mejor informar el *P*-valor resultante y no depositar en la significación la decisión de seleccionar variables para construir un modelo. Es mucho más productivo (aunque requiera mayor esfuerzo intelectual) usar el raciocinio biológico que el raciocinio estadístico binario para esas tareas. Frecuentemente, más que concluir que un factor es 'significativo', conviene informar el efecto del factor a través de un intervalo de confianza. Se debe reconsiderar el mirar a *P*-valor como criterio de verdad. 2) Alentar el pensamiento crítico para la construcción de múltiples hipótesis alternativas que se transformen en sendos modelos estadísticos que puedan contrastarse

Tabla 2. Preguntas incluidas en la encuesta sobre tesinas en biología de la Universidad Nacional de Córdoba y síntesis de las respuestas.

Table 2. Questions included in the survey on dissertations in biology at the National University of Cordoba and synthesis of the answers.

1. Fecha de presentación de la tesina	2017 = 8	2018/2019 = 23	
2. ¿Calificaría el trabajo de la tesina como exploratorio (apuntó a describir, generar hipótesis), como confirmatorio (apuntó a verificar hipótesis) o mezcla de ambos? (opciones: exploratorio, confirmatorio, ambos)	Exploratorio: 5	Confirmatorio: 10	Ambos: 10
3. ¿Se colectaron datos cuantitativos para hacer la tesina?	Sí: 29	No: 2	
4. ¿Conserva los datos originales cuantitativos crudos?	Sí: 30	No: 1	
5. Una vez terminada, la tesina puede ser archivada y olvidada. O puede continuar viva si es preparada para su publicación, o si sus resultados son guardados a la espera de resultados adicionales que sirvan para elaborar una publicación. Si la tesina que dirigió fue preparada para su publicación y/o guardada a la espera de datos adicionales, encontró inconvenientes para obtener los mismos resultados que los obtenidos en la tesina? A veces los nuevos datos no coinciden con los viejos datos, y/o los análisis que se realizaron oportunamente no pudieron repetirse, o se concluyó que los análisis hechos no eran los más adecuados, u otra/s razón/es.	Nunca volvió a mirar los datos: 1	Volvió a usarlos sin inconvenientes: 28	Intentó usarlos, pero encontró inconvenientes para obtener los mismos resultados reportados en la tesina: 2
6. ¿Se usó en la tesina alguna técnica estadística que selecciona variables según su nivel de significación (por ejemplo: regresión por pasos)? ¿Cuál/es? (respuesta abierta)	Sí: 7	No: 24	
7. ¿Se obtuvieron en el estudio resultados estadísticamente no significativos que no se incluyeron en la tesina?	Sí: 7	No: 24	
8. Si realizó un análisis cuantitativo, ¿Tuvo que cambiar a otro tipo de análisis estadístico después de que el análisis original falló en mostrar significación estadística?	Sí: 6	No: 25	
9. ¿Adhiere Ud. al concepto de Datos Abiertos? (datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen - http://opendatahandbook.org/guide/es/what-is-open-data/)	Sí: 20	No: 11	
10. ¿Están los datos cuantitativos crudos disponibles en el dominio público?	Sí: 1	No: 30	
11. ¿Continúa Ud. con la línea de trabajo vinculada a la tesina?	Sí: 29	No: 2	
12. ¿La tesina fue publicada?	Sí : 10	No: 21	

usando alternativas al *P*, tal como el criterio de información de Akaike (AIC). 3) Reforzar el entrenamiento para la elaboración de preguntas, hipótesis, diseño de estudios observacionales y experimentales, análisis de datos, y sobre la forma de presentar los análisis y sus interpretaciones. La estadística que se enseña en el grado es la desarrollada en la década de 1950. Si bien sigue siendo válida, los contenidos y estrategias didácticas distan de lo recomendado en GAISE (2016). La tradición de dictar Diseño de Experimentos, derivada de la experimentación en la escuela agronómica, deja un enorme hueco en la formación para la realización de estudios observacionales, típicos por ejemplo en estudios ecológico-ambientales. 4) Adoptar políticas editoriales para dejar a disposición los datos originales de una publicación, como ya están promoviendo las revistas de mayor jerarquía. 5) Alentar el entrenamiento para escribir el conjunto de pasos realizados para obtener un resultado (por ejemplo, usando scripts de R), con el objeto de facilitar la reproducibilidad computacional, y dejarlos a disposición junto con la publicación de los artículos. 6) Procurar esfuerzo institucional para abordar los problemas de reproducibilidad y replicabilidad en ciencia, incluyendo el entrenamiento en trabajo colaborativo, entrenamiento para identificar y declarar explícitamente investigación exploratoria/

confirmatoria, discusión sobre buenas prácticas en investigación, entrenamiento en *scientific workflows* y en *data+code sharing*.

Estas consideraciones pretenden llamar la atención sobre aspectos que no parecen estar actualmente integrados a la agenda de discusión académica, al menos en las ciencias naturales en la Argentina. Lejos están de realizar una evaluación sobre la calidad, honestidad o integridad de los colegas en esta parte del mundo.

AGRADECIMIENTOS. A E. Bucher, S. Catalá, J. DiRienzo, A. Farji-Brener, L. Galetto, R. Gurtler, D. Gurvich, J. Navarro y J. Polop, que con sus observaciones críticas ayudaron a mejorar la claridad del mensaje.

REFERENCIAS

Amrhein, V., S. Greenland, B. McShane, et al. 2019. Retire statistical significance. Nature **567**:305-307.

Baker, M., D. Penny. 2016. Is there a reproducibility crisis in science? Nature **533:**452-454. https://doi.org/10.1038/533452a.

Cummings, G. 2013. Dance of the p-values. URL: https://www.youtube.com/watch?v=5OL1RqHrZQ8.

Cummings, G. 2017. Significance roulette. URL: https://www.youtube.com/watch?v=OcJImS16jR4.

Farji-Brener, A., and A. Ruggiero. 2010. ¿Impulsividad o paciencia? ¿Qué estimula y qué selecciona el sistema científico argentino? Ecología Austral **20**:307-314.

Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. 2018. Questionable research practices in ecology and evolution. PLOS One 13(7):e0200303. https://doi.org/10.1371/journal.pone.0200303.

GAISE. 2016. College Report ASA Revision Committee, Guidelines for Assessment and Instruction in Statistics. Education College Report 2016. URL: http://www.amstat.org/education/gaise.

Ioannidis, J. 2005. Why most published research findings are false. PLoS Med **2(8)**:e124. https://doi.org/10.1371/journal.pmed.0020124.

Ioannidis, J. 2016. Why Most Clinical Research Is Not Useful. PLOS Med **13(6)**:e1002049. https://doi.org/10.1371/journal.pmed.1002049.

Leek, J. T., and R. D. Peng. 2015. P values are just the tip of the iceberg. Nature **520**:612. https://doi.org/10.1038/520612a.

Mazieres, M., and E. Kohler. 2014. Get me off Your Fucking Mailing List (accepted, unpublished). http://www.scs.stanford.edu/~dm/home/papers/remove.pdf.

Munafo, M. R., B. A. Nose, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. du Sert, U. Simmondson, E. Wagenmakers, J. J. Ware, and J. P. A. Ioannidis. 2017. A manifesto for reproducible science. Nat Hum Behav **1**, 0021. https://doi.org/10.1038/s41562-016-0021.

NAS (National Academies of Sciences, Engineering, and Medicine). 2016. Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop. Washington, DC. The National Academies Press. https://doi.org/10.17226/21915.

NAS (National Academies of Sciences, Engineering, and Medicine). 2019. Reproducibility and Replicability in Sciences. Washington DC. The National Academic Press. https://doi.org/10.17226/25303.

O'Boyle, E. H., G. C. Banks, E. González-Mulé. 2014. The Chrysalis effect: how ugly initial results metamorphosize into beautiful articles. Journal of Management **20(10)**:1-24. https://doi.org/10.5465/ambpp.2013.43.

Steward-Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern. 2017. Our path to better science in less time using open data science tools.

Nat Ecol Evol **1(6)**:160. https://doi.org/10.1038/s41559-017-0160.

Wasserstein, R. L., and N. A. Lazar. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician **70(2)**:129-133. https://doi.org/10.1080/00031305.2016.1154108.

Wasserstein, R. L., A. L. Schirm, N. A. Lazar. 2019. Moving to a world beyond *P*<0.05. The American Statistician **73(1)**:1-19. Special Issue Statistical Inference in the 21st Century: A World beyond P<0.05. https://doi.org/10.1080/00031305.2019.1583913.

Wittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology **75**:1182-1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x.