Seed Science Research

cambridge.org/ssr

Your *P*-values are significant (or not), so what … now what?

Héctor E. Pérez

Department of Environmental Horticulture, University of Florida, 2047 IFAS Research Drive, Gainesville, FL 32611, USA

Technical Update

Cite this article: Pérez HE (2024). Your *P*-values are significant (or not), so what … now what? *Seed Science Research* **34**, 86–89. https://doi.org/10.1017/S0960258524000035

Received: 25 October 2023
Revised: 26 January 2024
Accepted: 31 January 2024
First published online: 16 February 2024

Keywords:
d family; effect size index; magnitude of effect; *r* family; statistical hypothesis testing; strength of association

Corresponding author:
Héctor E. Pérez
Email: heperez@ufl.edu

Abstract

Statistical significance, or lack thereof, is often erroneously interpreted as a measure of the magnitude of effects, correlations between variables or practical relevance of research results. However, calculated *P*-values do not provide any information of this sort. Alternatively, effect sizes as measured by effect size indices provide complementary information to results of statistical hypothesis testing that is crucial and necessary to fully interpret data and then draw meaningful conclusions. Effect size indices have been used extensively for decades in the medical, psychological and social sciences but have received scant attention in the plant sciences. This Technical Update focuses on (1) raising awareness of these important statistical tools for seed science research, (2) providing additional resources useful for incorporating effect sizes into research programmes and (3) encouraging further applications of these tools in our discipline.

Introduction

Consider the hypothetical information presented in Figure 1. Data like this are often followed by enthusiastic statements that observed responses were 'very highly significantly different' (Fig. 1A) or dispirited assertions that differences were 'not statically significant' (Fig. 1B) when assessed against a theoretical level of statistical significance such as 0.05. Researchers then interpret these findings as evidence for large (Fig. 1A) or no (Fig. 1B) effects of the independent variables on response variables. Perhaps you have witnessed such examples at recent meetings or in publications. However, do comparisons of *P*-values against cut-off values (e.g., $\alpha = 0.05$) grant researchers the ability to make claims about the magnitude of differences, the strength of association between variables, or the practical relevance of a study? No! They do not (Nickerson, 2000; Ellis, 2010; Aarts et al., 2014; Nuzzo, 2014; Greenland et al., 2016; Wasserstein and Lazar, 2016; Wasserstein et al., 2019).

Unfortunately, the problem of *P*-value misinterpretation is widespread and chronic. Authors link this problem to: (1) pervasive misunderstandings regarding the fundamentals of statistical hypothesis testing; (2) conflating the original intent of *P*-values as a test of evidence against a null with the later application of *P*-values in evidence-based decision-making frameworks and (3) shortcomings of statistical training programmes (Nickerson, 2000; Nuzzo, 2014; Greenland et al., 2016; Pernet, 2016; Wasserstein and Lazar, 2016). Regardless, consensus exists that the use of results from statistical hypothesis testing alone, especially when viewed through the lens of significance or non-significance, distorts conclusions (Nuzzo, 2014; Greenland et al., 2016; Wasserstein and Lazar, 2016; Kimmel et al., 2023). As Wasserstein and Lazar (2016) stated: 'Statistical significance is not equivalent to scientific, human, or economic significance' [p. 132]. In this brief paper, I plan to raise the awareness of effect size measurements as necessary statistical tools; provide resources for further consideration and encourage more widespread use of effect sizes in the seed science literature.

A reminder of the information *P*-values provide

Fundamentally, a *P*-value represents the probability of observing a summary statistic (e.g., a mean difference between two groups) that is equal to or more extreme than the sample statistic given a specific statistical model (Wasserstein and Lazar, 2016). In more tangible terms, a *P*-value represents a measure of compatibility between observed data and the expected data if all assumptions of a test model (e.g., the null hypothesis) were correct. The smaller the *P*-value the more likely that observed data are unusual compared to the test model. Alternatively, the larger the *P*-value the more likely that observed data are not unusual compared to the test model. That's it! This is *all* the information a *P*-value provides the researcher – nothing else (Nuzzo, 2014; Greenland et al., 2016; Wasserstein and Lazar, 2016).

Crucially, notice how *P*-values provide no information regarding the magnitude of differences or the level of association between variables. Nuzzo (2014), Greenland et al. (2016),

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

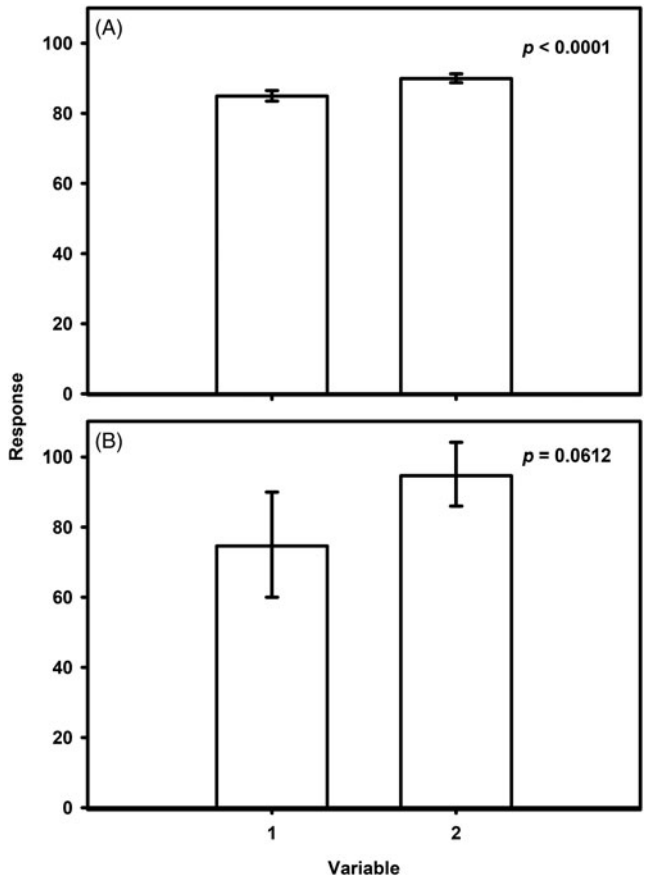


Figure 1. Hypothetical experimental results of seed biology experiments displaying responses that are considered (A) highly and (B) not significantly different according to the common statistical cut-off value of $\alpha = 0.05$.

Wasserstein and Lazar (2016), and Wasserstein et al. (2019) offer more complete descriptions of *P*-value misinterpretations, provide an excellent refresher on what *P*-values do and do not represent, and explain what not do with *P*-values while offering meaningful actions researchers can take in the context of statistical analyses.

The power of effect size indexes

It is important to note that sample size affects *P*-values. For instance, *P*-values typically decrease as sample size increases due to the impact of random error reduction. Moreover, variability decreases and measurements become more precise in large samples. Such improvements facilitate the detection of smaller differences (Cohen, 1988; Ellis, 2010; Greenland et al., 2016; Wasserstein and Lazar, 2016). This means that trivial differences or associations may be deemed statistically significant (e.g., Fig. 1A) if the sample size is large enough or measurements are highly precise. The reverse is also true. Non-trivial differences may show up as not statistically significant in studies with small sample sizes or imprecise measurements (Cohen, 1988; Ellis, 2010; Greenland et al., 2016; Wasserstein and Lazar, 2016). Alternatively, effect size indices are independent of sample size (Cohen, 1988; Ellis, 2010).

So, what are effect size indices? Effect size indices are statistics that quantify the magnitude of differences between treatment groups or experimental conditions and correlations between variables (Cohen, 1988; Ellis, 2010). Researchers may be familiar with some types of indices (Kallogjeri and Piccirillo, 2023) but may not have interpreted these as effect sizes. For example, the odds ratio, which is often calculated in connection with logistic regression, computes the odds of an event (e.g., fungal contamination) occurring in one group (e.g., seeds treated with fungicide A) compared to another group (e.g., seeds treated with fungicide B). Let's say a subsequent analysis yields an odds ratio equal to 1.86. This means that the odds of fungal contamination in seeds treated with fungicide A is 86% higher than the odds for fungal contamination in seeds treated with fungicide B. Similarly, the hazard ratio (HR), which is associated with regression-based time-to-event analyses in seed biology (McNair et al., 2012; Pérez and Kettner, 2013; Genna et al., 2015; Adegbola and Pérez, 2016; Genna and Pérez, 2016; Pérez and Kane, 2017; Tyler et al., 2017; Campbell-Martínez et al., 2019; Pérez and Chumana, 2020), represents the ratio of estimated hazard rates (i.e., likelihood of germination) between different covariate values (e.g., doses of a germination-stimulating chemical; treated vs. control) over a unit of time (Allison, 2010). Consider an experiment from a germination perspective where a group of seeds received an increasing dose of a germination inhibitor. In this case, the calculated HR equals 0.95. Applying the formula $100 \cdot (\mathrm{HR} - 1)$ yields the percent change in hazard for each 1-unit increase in the germination inhibitor dose. Therefore, the likelihood of germination decreases by 5% for each 1-unit increase of inhibitor. Other types indices, such as Hedges' *g*, Cramér's *V*, or eta^2 (η^2), may be less familiar, given the large number (around 70) of indices that exist (Cohen, 1988; Kirk, 2003; Ellis, 2010).

Effect size indices fall into the *d* or *r* families. Indices in the *d* family measure differences between groups. Indices of the *r* family measure associations between variables (Ellis, 2010; Kallogjeri and Piccirillo, 2023). Ellis (2010, see table 1.1) goes on to subdivide the *d* family into indices that compare groups on dichotomous outcomes (e.g., odds ratio) and those that compare groups on continuous outcomes (e.g., Cohen's *d*). Likewise, the *r* family is divided into indices assessing correlation (e.g., Cramér's *V*) or the proportion of variance (e.g., η^2).

Selecting a suitable effect size index requires the consideration of several factors (Ellis, 2010; Kallogjeri and Piccirillo, 2023). For example, researchers should consider the research problem under investigation. This helps to identify study aims, target outcomes, data structure, measurement methods and study design. Next, researchers define whether outcomes or dependent variables are categorical, continuous or time-to-event in nature. Finally, researchers describe the type of analysis being conducted such as correlations, regressions, multivariate analysis or analysis of variance (ANOVA) with multiple groups. Researchers with this information in hand will find it easier to determine which index to use when referring to helpful tabulated resources (Ellis, 2010, see table 1.2) or decision trees (Kallogjeri and Piccirillo, 2023).

With the proper effect size index selected, researchers can then move on to analyses and interpretation. But first, consider these cautions. Different indexes will provide different measurement scales corresponding to what constitutes small, medium or large effects (or association). For example, depending on the scientific discipline, a Pearson's correlation coefficient (*r*) value of 0.25 could be considered as a small association between variables of interest. However, a Cohen's *d* value of 0.25 can be deemed a medium effect size (Aarts et al., 2014). Therefore, it is challenging to compare indices that use different effect size criteria unless

https://doi.org/10.1017/S0960258524000035 Published online by Cambridge University Press

Table 1. List of additional resources related to effect sizes

Resource	URL	Notes
Book/Article		
Ellis P (2010) *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*	https://www.cambridge.org/core/books/essential-guide-to-effect-sizes/72C26CA99366A19CAC4EF5B16AE3297F	Excellent resource and straightforward primer on effect sizes
Kallogjeri D. &Piccirillo JF (2023) A simple guide to effect size measures	https://jamanetwork.com/journals/jamaotolaryngology/fullarticle/2802363	Decision tree for selecting effect size indices is extremely helpful
Software		
R – effect size package	https://cran.r-project.org/web/packages/effectsize/vignettes/effectsize.html https://easystats.github.io/effectsize/	
SAS		
EFFECTSIZE option (PROC GLM)	https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_glm_details22.htm	
MEASURES option (PROC FREQ)	https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_freq_details22.htm	
ODDSRATIO option (PROC GLIMMIX, LOGISTIC);	https://documentation.sas.com/doc/en/statcdc/14.2/statug/statug_logistic_syntax26.htm; https://documentation.sas.com/doc/en/statug/15.2/statug_glimmix_details49.htm	
HAZARDRATIO option (PROC PHREG)	https://documentation.sas.com/doc/en/statug/15.2/statug_phreg_syntax13.htm	Also see: (Savarese and Patetta 2010)
G*Power	https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower	Freeware that combines effect size with power analysis and facilitates the calculation of appropriate sample sizes
Websites		
Wikipedia	https://en.wikipedia.org/wiki/Effect_size	
Wikiversity	https://en.wikiversity.org/wiki/Effect_size	
Wikiversity	https://en.wikiversity.org/wiki/Effect_size/Data_analysis_tutorial	
Statistics by Jim	https://statisticsbyjim.com/basics/effect-sizes-statistics/	
Effect size calculators	https://lbecker.uccs.edu/	Many more on-line effect size calculators exist
Computation of effect sizes	https://www.psychometrica.de/effect_size.html	
Effect size calculator	https://www.cem.org/effect-size-calculator	
Statistics Kingdom	https://www.statskingdom.com/effect-size-calculator.html	
MOTE Effect Size Calculator	https://www.aggieerin.com/shiny-server/	
Statistics in Research	https://statsinresearch.com/links-to-useful-sites/effect-size-and-power-calculators.html	

index conversion formulas are available. In some cases, it may not be possible to convert between indices. Additionally, the criteria for effect sizes of a specific index (e.g., Cohen's d) may not necessarily be applicable across disciplines. A small effect in seed science may not be the same as a small effect in medical research. Consequently, interpretations of effect sizes should be discipline-specific (Cohen, 1988; Ellis, 2010; Brydges, 2019). Finally, remember to report confidence intervals associated with the calculated effect size index. This provides a measure of precision of the effect size estimate and represents good statistical practice (Greenland et al., 2016; Wasserstein and Lazar, 2016; Wasserstein et al., 2019; Kallogjeri and Piccirillo, 2023).

Researchers in the medical and social sciences have been applying effect size indices in their analyses for decades. Such a robust body of analyses often leads to the standardization and contextualization of small, medium and large effects within a discipline (Cohen, 1988; Ellis, 2010). Alternatively, apart from ecology, the utilization of effect sizes in many plant-related disciplines including seed science has been negligible (Sileshi, 2012). An important outcome is that the standardization of small, medium and large effects for some indices will be absent. To remedy this, Cohen (1988) cautiously suggested using criteria outlined in his publication when no discipline-specific criteria exist. For example, values of Cohen's $d = 0.2$, 0.5, and 0.8 represent benchmarks for small, medium and large effect sizes. But the use of various general criteria offered by Cohen (1988) must be tempered with the researcher's experience and wisdom. For instance, a five-percentage point difference in a laboratory germination test (e.g., 93 vs. 98%) for lettuce seeds may turn out to be a small effect. Nonetheless, when scaled to the field level, this difference

https://doi.org/10.1017/S0960258524000035 Published online by Cambridge University Press

can have a substantial impact since the success of a lettuce crop may rely on each sown seed producing a harvestable head. So, context is essential when interpreting effect sizes. Otherwise, criteria such as small, medium or large may remain ambiguous (Cohen, 1988; Ellis, 2010; Carey et al., 2023).

More information on effect sizes

Reporting and interpretation of effect sizes in the seed science literature is rare (Sileshi, 2012); suggesting that effect sizes represent new concepts for our discipline. If the topic of effect sizes is new to you, then a good place to start is with easy to digest reading materials (Table 1). Fortunately, most statistical analysis programmes have the capacity to calculate many effect size indices (Table 1). These programmes also tend to provide adequate documentation explaining available indices. If statistical programmes are unavailable or inaccessible, then various websites provide applications to calculate effect sizes (Table 1). Similarly, calculations for several effect sizes are straightforward (Cohen, 1988; Ellis, 2010) and can easily be computed in a spreadsheet or by hand if necessary (Table 1).

Concluding remarks

Effect size indices are powerful tools crucial for extending our results beyond mere statistical significance. Moreover, effect sizes are important to ensure that our studies are properly powered rather than underpowered (Cohen, 1988; Ellis, 2010; Nuzzo, 2014; Greenland et al., 2016, Brydges, 2019; Kimmel et al., 2023; also see Table 1). Effect size indices are simple to apply. More importantly, the information these indices yield contributes to more impactful conclusions relevant to broader audiences while moving a discipline forward. Therefore, I strongly encourage the use of effect sizes in future seed science research.

Funding statement

This work received no specific grant from any funding agency, commercial or non-profit sectors.

Competing interests. The author declares none.

References

Aarts S, van den Akker M and Winkens B (2014) The importance of effect sizes. *European Journal of General Practice* **20**, 61–64. https://doi.org/10.3109/13814788.2013.818655

Adegbola Y and Pérez H (2016) Extensive desiccation and ageing stress tolerance characterize *Gaillardia pulchella* (Asteraceae) seeds. *HortScience* **51**, 159–163.

Allison P (2010) *Survival Analysis using SAS®: A Practical Guide*, 2nd Edn. Cary, NC: SAS Institute Inc.

Brydges CR (2019) Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging* **3**, 1–8. https://doi.org/10.1093/geroni/igz036

Campbell-Martínez G, Thetford M, Miller D and Pérez H (2019) Seedling emergence of *Lupinus diffusus* in response to abrasion in an electric seed scarifier. *Native Plants Journal* **20**, 14–24.

Carey EG, Ridler I, Ford TJ and Stringaris A (2023) Editorial perspective: when is a 'small effect' actually large and impactful? *Journal of Child Psychology and Psychiatry* **64**, 1643–1647. https://doi.org/10.1111/jcpp.13817

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*. New York: Lawrence Erlbaum Associates.

Ellis P (2010) *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, UK: Cambridge University Press.

Genna N and Pérez H (2016) Mass-based germination dynamics of *Rudbeckia mollis* (Asteraceae) seeds following thermal and ageing stress. *Seed Science Research* **26**, 231–244. https://doi.org/10.1017/S0960258516000180

Genna N, Kane M and Pérez H (2015) Simultaneous assessment of germination and infection dose-responses in fungicide-treated seeds with non- and semiparametric statistical analyses. *Seed Science and Technology* **43**, 1–19. https://doi.org/10.15258/sst.2015.43.2.13

Greenland S, Senn S, Rothman K, Carlin J, Poole C, Goodman SN and Altman D (2016) Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* **31**, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Kallogjeri D and Piccirillo JF (2023) A simple guide to effect size measures. *JAMA Otolaryngology-Head & Neck Surgery* **149**, 447–451. https://doi.org/10.1001/jamaoto.2023.0159

Kimmel K, Avolio ML and Ferraro PJ (2023) Empirical evidence of widespread exaggeration bias and selective reporting in ecology. *Nature Ecology & Evolution* **7**, 1525–1536. https://doi.org/10.1038/s41559-023-02144-3

Kirk R (2003) The importance of effect magnitude. In Davis S (ed.), *Handbook of Research Methods in Experimental Psychology*, Malden, MA: Blackwell Publishing, pp. 83–105.

McNair J, Sunkara A and Frobish D (2012) How to analyse seed germination data using statistical time-to-event analysis: non-parametric and semi-parametric methods. *Seed Science Research* **22**, 77–95. https://doi.org/10.1017/S096025851100054

Nickerson R (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* **5**, 241–301.

Nuzzo R (2014) Scientific method: statistical errors. *Nature* **506**, 150–152. https://doi.org/10.1038/506150a

Pérez H and Chumana L (2020) Enhancing conservation of a globally imperiled rockland herb (*Linum arenicola*) through assessments of seed functional traits and multi-dimensional germination niche breadths. *Plants* **9**, 1493. https://doi.org/10.3390/plants9111493

Pérez H and Kane M (2017) Different plant provenance same seed tolerance to abiotic stress: implications for ex situ germplasm conservation of a widely distributed coastal dune grass (*Uniola paniculata* L.). *Plant Growth Regulation* **82**, 123–137. https://doi.org/10.1007/s10725-016-0244-1

Pérez H and Kettner K (2013) Characterizing *Ipomopsis rubra* (Polemoniaceae) germination under various thermal scenarios with non-parametric and semi-parametric statistical methods. *Planta* **238**, 771–784. https://doi.org/10.1007/s00425-013-1935-8

Pernet C (2016) Null hypothesis significance testing: a short tutorial. *F1000 Research* **4**, 621. https://doi.org/https://doi.org/10.12688/f1000research.6963.5

Savarese P and Patetta M (2010) An overview of the CLASS, CONTRAST, and HAZARDRATIO statements in the SAS® 9.2 PHREG procedure. In *Proceedings of the SAS Global Forum 2010 Conference*. Cary, NC: SAS Institute, pp. 1–23. Available at https://support.sas.com/resources/papers/proceedings10/253-2010.pdf

Sileshi G (2012) A critique of current trends in the statistical analysis of seed germination and viability data. *Seed Science Research* **22**, 145–159. https://doi.org/10.1017/S0960258512000025

Tyler T, Adams C, MacDonald G and Pérez H (2017) Florida ecotype Elliot's lovegrass (*Eragrostis elliottii*) germination testing for use in non-optimal restoration sites: the role of season and seed vigor. *Native Plants Journal* **18**, 114–125.

Wasserstein RL and Lazar NA (2016) The ASA's Statement on *p*-values: context, process, and purpose. *The American Statistician* **70**, 129–131.

Wasserstein R, Schrim A and Lazar N (2019) Moving to a world beyond '$p<0.05$'. *The American Statistician* **73**, 1–19. https://doi.org/10.1080/00031305.2019.1583913

https://doi.org/10.1017/S0960258524000035 Published online by Cambridge University Press