Letter to the Editor

Errors or Irreproducibility in Effect Size Calculations and Incomplete Reporting of Results in “Systematic Review of the Effects of Blueberry on Cognitive Performance as We Age”

Colby J. Vorland, PhD^{1,•} Paola P. Matthey-Mora, MPH² Luis M. Mestre, MS² Xiwei Chen, MS² Stephanie L. Dickinson, MS² Andrew W. Brown, PhD¹ and David B. Allison, PhD^{2,*}

¹Department of Applied Health Science, Indiana University School of Public Health-Bloomington. ²Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington.

*Address correspondence to: David B. Allison, PhD, Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN. E-mail: allison@iu.edu

Dear Editor,

We read the article “Systematic Review of the Effects of Blueberry on Cognitive Performance as We Age” by Hein and colleagues (1) (hereafter “the review”). We are unable to reproduce some of the published effect sizes, and despite the label as a systematic review in the article’s title, it does not follow standard protocols for systematic review conduct or reporting.

We attempted to recalculate the 21 total effect sizes reported in the tables in the review (Table 1) and encountered insufficient information from the publications to reproduce 10 of them. All of the largest effect sizes reported were unable to be reproduced from the original publications. Of the 21, we were able to exactly reproduce 2 of them, closely reproduce 3, and 4 were not close when recalculated. Eight of the effect sizes were from studies of authors of the review and have not been previously peer reviewed, and the data and results are not available for independent verification. Further, we note that two of the comparisons by Krikorian and colleagues (2010) (2), among the largest effect sizes included in the review, are actually within-group comparisons, which are invalid for between-group inferences of effects of blueberry on cognitive performance (3).

An accepted reporting standard for systematic reviews is the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement (4), which requires a systematic reporting of study outcomes, minimizing the likelihood of a bias in the presentation of the literature to readers. Within tables 2, 3, and 4 in the review, only results below or near $p = .05$ are shown in the “Key Findings” column, but most studies reviewed included many outcomes and statistical comparisons that resulted in $p > .05$. While

some of these other comparisons are discussed narratively within the text, the discussion is not comprehensive, which is a key purpose of systematic reviews. To underscore this point with an example from the review, Whyte and colleagues (2016) (5) involved four cognitive tests after consumption of freeze-dried blueberries at 15 g, 30 g, or vehicle control. For each cognitive test, each group was tested at baseline, 1.15, 3, and 6 hours. In Table 2 of the review, three p -values $< .05$ from ANOVA models from this study are noted. In neither the table nor in the text does the review emphasize that most comparisons yielded no differences for blueberries at either dose compared to vehicle. We count on the order of 200 reported means among all cognitive tests from which only these few between group differences are highlighted. Further, the comparisons that were < 0.05 were each at different timepoints, and two were the 30 g blueberry dose and one the 15 g dose, and they were each within different measures, revealing no consistent effects across time, biological gradient, nor test. A plausible explanation for these inconsistent findings could be that the many comparisons produced some findings that favor blueberries that are type 1 errors due to multiple testing. The extent of multiple comparisons within and between studies is not currently obvious to readers of the review.

Finally, we discovered a lack of systematic review guideline adherence and errors in study descriptions in the review. According to the PRISMA statement, which has adopted the Cochrane systematic review definition (6) details from the review are missing to fulfill the checklist criteria of a reproducible systematic review. While reading the review, we observed that multiple items were not reported: the exact search queries used in each database (criteria #8), the search dates and dates of coverage for each database

Table 1. Summary of our attempt at reproducing the effect sizes reported in the studies included in Hein et al.

Article (ref # in Hein et al.)	Outcome	Effect Size Calculated by Hein et al. authors? ^a	Reported Cohen's <i>d</i>	Our Calculated Cohen's <i>d</i> ^b	Match ^c
Schrager et al. (36)	DTAG (step errors)	No	1.16	NC	NA
Whyte et al. (41)	RAVLT: word recog (12 wk)	Yes	0.578	NC	NA
	Corsi Blocks: total sequences	No	0.289	NC	NA
Miller et al. (35)	TST: switch cost	No	0.629	NC	NA
	CVLT: repetition errors	No	0.759	0.758	Yes
Barfoot et al. (30)	AVLT: total acquisition performance	Yes	0.425	0.482	Close
	AVLT: short delay recall	Yes	0.405	0.414	Close
	MANT: reaction time	Yes	0.175	0.062	No
Boespflug et al. (32)	fMRI (left inferior parietal gyrus)	No	1.82	NC	NA
	fMRI (left precentral gyrus)	No	1.94	NC	NA
Krikorian et al. (33)	V-PAL: across visits	No	1.78	Invalid within group comparison	NA
	V-PAL: vs placebo	No	0.96	NC	NA
	CVLT: word recall across visits	No	1.18	Invalid within group comparison	NA
Whyte et al. (27)	AVLT: delayed recall	Yes	0.904	NC	NA ^d
	AVLT: proactive interference	Yes	0.883	0.601	No
Whyte et al. (28)	AVLT: final acquisition at 1.15h	Yes	0.908	NC	NA
	AVLT: delayed word recognition at 6h	Yes	0.245	0.598	No
	MFT: incongruent trial accuracy at 3h	No	0.201	0.606	No
Whyte et al. (29)	MANT: reaction time	No	0.94	NC	NA
McNamara et al. (34)	DEX: cognitive symptoms	No	0.68	0.657	Close
	HVLT: memory discrimination	No	0.68	0.677	Yes

Note: AVLT = auditory verbal learning task; CVLT = California verbal learning test; DEX = dysexecutive questionnaire; DTAG = dual-task adaptive gait; fMRI = functional magnetic resonance imaging; HVLT = Hopkins verbal learning test; MANT = modified attention network task; MFT = modified Flanker task; RAVLT = Rey's auditory verbal learning test; TST = task-switching test; V-PAL = verbal paired associate learning.

^aYes: effect sizes previously reported in original publications of included studies. No: effect sizes were calculated by authors of Hein et al., who performed the original studies.

^bNC: Not Calculable. Insufficient information in the original paper to calculate the Cohen's *d*. Calculations are described in more detail at <https://osf.io/9rxya/>.

^cQualitative interpretation of how closely our calculations match those of Hein et al. NA: Not Applicable. Yes: results are exactly replicated or within rounding error. Close: results deviate within a range that we posit could potentially be explained by differences in calculation procedures (eg, pooling, assumed equal variance or sample size, imputation of correlations). No: values deviate substantially.

^dEffect size was reproduced when ignoring group dependency from the crossover design, thus the reported value may not be correct.

(#7), whether study screening was performed in duplicate (#10) and how many studies were screened and excluded (#17). In addition, risk of bias assessments within and across studies (#15, 19, and 22) should be included to formally assess the quality and certainty of the research in a standardized manner. Indeed, a recent analysis of the studies included in the review is suggestive of publication bias and/or other questionable research practices (7). Further, 3 of the 11 studies employ crossover designs, which are appropriately described within the text, but the authors mislabel some designs in the Tables and in the discussion: "... *all but two studies (39, 45) employed a double-blind crossover, placebo-controlled design....*"

The combination of irreproducible effect size calculations, selective reporting of effects, and general errors in systematic review methodology result in a misrepresentation of the strength of evidence about blueberries and cognitive performance. We encourage the authors to share their data and calculations and to correct this article.

Funding

This study was supported in part by the Gordon and Betty Moore Foundation and National Institutes of Health (NIH) grants U24AG056053, P30AG050886, and R25HL124208. The opinions expressed are those of the authors and do not necessarily represent those of the NIH or any other organization.

Conflict of Interest

D.B.A. has received personal payments or promises for same from: American Society for Nutrition; American Statistical Association; Biofortis; California Walnut Commission; Columbia University; Fish & Richardson, P.C.; Frontiers Publishing; Henry Stewart Talks; IKEA; Indiana University; Laura and John Arnold Foundation; Johns Hopkins University; Law Offices of Ronald Marron; MD Anderson Cancer Center; Medical College of Wisconsin; National Institutes of Health (NIH); Sage Publishing; The Obesity Society; Tomasik, Kotin & Kasserman LLC; University of Alabama at Birmingham; University of Miami; Nestle; WW (formerly Weight Watchers International, LLC). Donations to a foundation have been made on his behalf by the Northharvest Bean Growers Association. D.B.A. is an unpaid member of the International Life Sciences Institute North America Board of Trustees. D.B.A.'s institution, Indiana University, has received funds to support his research or educational activities from: NIH; Alliance for Potato Research and Education; American Federation for Aging Research; Dairy Management Inc; Herbalife; Laura and John Arnold Foundation; National Cattlemen's Beef Association, Oxford University Press, the Sloan Foundation, The Gordan and Betty Moore Foundation, and numerous other for-profit and nonprofit organizations to support the work of the School of Public Health and the university more broadly. D.B.A.'s prior institution, the University of Alabama at Birmingham, received gifts, contracts, and grants from other organizations including the Coca-Cola Company, Pepsi, and Dr. Pepper/Snapple. In the last 12 months, A.W.B. has received travel expenses from the University of Louisville and grants through his institution from Dairy Management, Inc. and the National Cattlemen's Beef Association. He has been involved in research for which his institution or colleagues have received grants

from the Gordon and Betty Moore Foundation, NIH/NHLBI, NIH/NIA, NIH/NIDDK, and Sloan Foundation. Other authors report no disclosures.

References

1. Hein S, Whyte AR, Wood E, Rodriguez-Mateos A, Williams CM. Systematic review of the effects of blueberry on cognitive performance as we age. *J Gerontol A Biol Sci Med Sci*. 2019;74:984–995. doi: [10.1093/gerona/glz082](https://doi.org/10.1093/gerona/glz082)
2. Krikorian R, Shidler MD, Nash TA, et al. Blueberry supplementation improves memory in older adults. *J Agric Food Chem*. 2010;58:3996–4000. doi: [10.1021/jf9029332](https://doi.org/10.1021/jf9029332)
3. Bland JM, Altman DG. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *Am J Clin Nutr*. 2015;102:991–994. doi: [10.3945/ajcn.115.119768](https://doi.org/10.3945/ajcn.115.119768)
4. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med*. 2009;151:264–9, W64. doi: [10.7326/0003-4819-151-4-200908180-00135](https://doi.org/10.7326/0003-4819-151-4-200908180-00135)
5. Whyte AR, Schafer G, Williams CM. Cognitive effects following acute wild blueberry supplementation in 7- to 10-year-old children. *Eur J Nutr*. 2016;55:2151–2162. doi: [10.1007/s00394-015-1029-4](https://doi.org/10.1007/s00394-015-1029-4)
6. Higgins J, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 [updated March 2011]. Chichester (UK): John Wiley & Sons; 2011.
7. Brydges C, Gaeta L. Blueberries and cognitive ability: a test of publication bias and questionable research practices. *PsyArXiv*. 2019.