7 Deadly Sins in Educational Research

Katherine Picho, PhD
Anthony R. Artino Jr, PhD

Is coffee the new superfood or early death in a cup? It depends on where you look for supporting evidence. For example, our colleague recently gave up attempts to quit his coffee habit because findings from a new study suggested that daily doses boost longevity.[1] On the other hand, another recent study reported that antioxidants (of which coffee has plenty) are now believed to be related to cancer, and worse, may spread cancer faster in those with the disease.[2,3]

Such conflicting messages are not uncommon in biomedical research. Although there are many causes of such contradictory results, some of the inconsistencies result from bias in the original research. Research bias is often the product of questionable research practices and poor research design. In fact, errors in research design and analysis increase the probability of obtaining results that are misleading, exaggerated, or just plain wrong.

Concerns over the validity of scientific research have grown in recent years, with considerable evidence indicating that most published research findings in the biomedical sciences are false.[4] The major flaws that infect research studies—in education as well as biomedical science—often relate to small samples, small effects, and loosely defined and implemented research designs.[4] While many researchers expect that the scientific literature self-corrects over time, this is not always the case. Indeed, considering the "file drawer effect" (unpublished studies with negative outcomes) and the fact that replication remains an underappreciated and relatively uncommon enterprise,[5] self-correction of faulty results may be the exception, not the rule. In response to these challenges, this editorial highlights the most common educational research practices, particularly for quantitative studies, that lead researchers to report misleading, exaggerated, or entirely false findings. The intent of this article is to raise awareness and encourage medical education researchers to avoid the "7 deadly sins" in educational research (BOX).

DOI: http://dx.doi.org/10.4300/JGME-D-16-00332.1

Sins Committed Before Research

Sin #1: The Curse of the Handicapped Literature Review

Empirical research is the primary means of theory testing and development. It is also essential for testing practical interventions in authentic educational environments. The literature review is central to this process as it identifies existing strengths, weaknesses, and knowledge gaps in a particular field. The literature review informs key aspects of the research process (ie, research questions, design, and methods) and delineates boundaries within which inferences about findings can be discussed. Consequently, sins committed in the literature review process can have profound effects on every aspect of a study and thus negatively influence study quality.[6]

Unfortunately, researchers will often conduct partial reviews that are skewed in favor of their hypotheses. Even more common (and worse) is the practice of conducting the literature review after the study has been completed and the results are known. Such practices allow researchers to selectively use articles and revise hypotheses in support of their results. This is a problem because variation due to randomness, which is an expected part of scientific research, yields a fair number of spurious findings.[7] Reformulating hypotheses after results are known is not only a backward approach to the scientific method, but it also increases the likelihood of polluting the field of study with false conclusions based on spurious findings. Such practices could explain why some study findings fail to replicate.[8]

Sin #2: Inadequate Power

In quantitative studies, statistical tests help researchers make inferences about the nature and magnitude of the relationships between independent or predictor variables and outcomes. The extent to which conclusions about these inferences are deemed reasonable is sometimes referred to as *statistical conclusion validity*.[9] In the social sciences, many investigations focus on evaluating group differences on certain phenomena. However, there is always the risk that one could falsely find group differences where they do not exist in the population. This is called a type 1 error, or a false positive.[9] Type 1 errors can be minimized by

increasing the statistical power of a test, which is the probability of finding a statistically significant difference among groups when such a difference actually exists.[10] Statistical power values range from 0 (no power) to 1 (extremely high power). Although increasing power to extremely high values (eg, to a power of 1) might seem like a simple solution to drastically reduce the likelihood of obtaining a false positive, this approach has the unintended consequence of increasing the probability of obtaining a false negative, or a type 2 error.[9] Therefore, statistical power must walk a fine line between the 2 ends of the spectrum: high enough to detect true group differences without drastically increasing the risk of making a type 2 error. In educational research, the convention for optimum power is typically 0.8.[11]

Power is affected by sample size and the number of hypotheses being tested, among other factors. One study found that most studies in the social sciences, including psychology and education,[12] were underpowered. In psychology, the average power of studies was 0.35.[12] In medical education, it is not uncommon for quantitative studies to be conducted with sample sizes as low as 20, 15, or even 10 participants. Therefore, it is likely that many medical education research studies are insufficiently powered to detect true differences among groups.

Power is also affected by the magnitude of the expected effect, such as the size of the differences between 2 groups. Hence, in a given study, low power may stem from small samples and small effects or a combination of both.[13] In addition to missing a true difference between groups, low power also reduces the likelihood that a statistically significant result represents a true effect rather than a spurious finding.[13] Both of these issues weaken the reliability of findings in a given field. The former may lead to prematurely discarding hypotheses that might advance understanding, and the latter, to spurious findings that cannot be replicated.

A power analysis should be conducted prior to data collection to avoid these negative consequences. Besides increasing sample size, power can be increased by improving experimental design efficiency, such as through the use of equal cell sample sizes; matching participants; measuring covariates a priori; and correcting for covariates in subsequent analyses.

Sin #3: Ignoring the Importance of Measurement

Measurement error weakens the relationship between 2 variables and can also strengthen (or weaken) the relationships among 3 or more variables.[9] Using measures that have not been tested, or employing those that have poor psychometric properties, only serves to add more "noise" to the results and potentially taints the field with contradictory or implausible findings.[14]

Measurement problems can stem from measurement tools (eg, questionnaires) that underrepresent or overrepresent the construct under study. When a measurement tool is too narrow (eg, in the case of single-item measures), then it likely excludes important aspects of the construct and thus fails to capture the true nature of the phenomenon of interest.[14] Measurement problems also occur when the outcome variables (eg, test scores, clerkship grades) are too easy or too difficult. Tasks that are extremely easy or difficult lead to ceiling and floor effects, respectively, which weaken correlations and bias results.

BOX Checklist of Recommendations for Responsible Research Conduct

- Conduct a thorough literature review
- Specify hypotheses a priori based on literature review
- Enlist the help of a statistician prior to study design
- Select research designs appropriate to the research questions
- Conduct a power analysis based on research design and literature
- Select measures with evidence of reliability and validity for the intended purpose
- Avoid using single-item measures of complex constructs (eg, motivation, confidence, satisfaction, resilience)
- Before analysis, check to make sure statistical assumptions for the analytic technique have been met
- If assumptions are violated, take steps to remedy those violations and report these steps in the manuscript
- If outliers are removed, report this practice and provide a rationale for removal
- Conduct statistical analyses appropriate to the research questions
- Avoid testing hypotheses that were not specified a priori
- Report descriptive statistics, including means and standard deviations
- Report effect sizes and confidence intervals around effect sizes
- Report nonsignificant results along with statistically significant findings

Sins Committed During Research

Sin #4: Using the Wrong Statistical Tool

Scholars have written much about the sins related to statistical analyses in research. The most common involve not checking (or reporting) whether the data meet assumptions of the statistical technique being used. Perhaps the most frequently violated assump-

tion is the assumption that observations are independent. Related to this specific violation is the mistake of treating nondependent data as if they were independent (eg, treating data from 20 participants that are measured 3 times as if data are from 60 participants).[15]

The violation of such statistical assumptions has the effect of artificially inflating type 1 errors (false positives), which leads to more statistically significant results than warranted. This outcome threatens the validity of inferences that can be made from statistically significant results and can also result in replication failure. To avoid this pitfall, researchers should verify that their data meet the assumptions of the data analytic technique they intend to use. When statistical assumptions are violated, one should take steps to remedy the problem (eg, transforming nonnormal data) or use alternate statistical techniques that are robust to these violations (eg, nonparametric statistics for continuous data that do not follow a normal distribution). Moreover, it can be helpful to consult a statistician early in the research process; such a practice is critical to finding the right statistical tool for the job.

Sin #5: Merciless Torture of Data and Other Questionable Analysis Practices

Questionable research practices are prevalent in the social sciences, and medical education is not immune to these problems. Although data fabrication constitutes the extreme end of a continuum, there is evidence that other questionable practices are rampant. Examples of such practices include reporting only results that align with one's hypotheses ("cherry picking"), relaxing statistical significant thresholds to fit results, using 1-sided *t* tests but failing to mention this in the research report, and wrongly rounding *P* values upward or downward to fit with a hypothesis (eg, reporting $P = .04$, when the actual *P* value is .049).[16]

Another popular yet questionable practice is fishing, which refers to mining data for statistically significant findings that do not stem from prespecified hypotheses.[9] Fishing increases type 1 error rates and artificially inflates statistical significance. Indeed, it would be a sin to restructure an entire study around findings from a fishing expedition, especially since these findings are more likely to be a product of chance than the result of actual differences in the population. Although findings based on fishing expeditions and other questionable practices generally work to the advantage of the researcher (ie, they improve the chances of reaching a statistically significant result and getting published), they ultimately hurt rather than advance knowledge.

Sins Committed After Research

Sin #6: Slavery to the *P* Value

The most commonly applied and accepted approach to statistical inference in the social sciences is null hypothesis significance testing,[17] where a researcher's hypothesis about group differences on a given construct is tested against the null hypothesis: there are no differences.[18] Generally, statistical analyses generate a score that reflects mean group differences for a variable, accompanied by test statistics (*t* ratios, chi-square analyses, etc) and a probability value (*P* value). *P* values represent the probability of obtaining the observed group difference or a more extreme result if said difference did not exist in the population from which the data were sampled.[19] To determine statistical significance, *P* values corresponding to .05 (or less than .05) are usually selected as being indicative of a statistically significant group difference.

Although a useful tool, *P* values are not very informative. First, a statistically significant result (ie, rejecting the null hypothesis) does not in any way confirm the researcher's hypotheses, although most times it is falsely perceived and interpreted as such.[20,21] Second, extremely large sample sizes (eg, in the thousands) will magnify small group differences; the result may be statistically significant yet practically unimportant due to tiny differences. In educational research, large sample sizes are rare but occasionally are seen when large databases are available (eg, specialty board scores). Researchers should focus on supplementing *P* value statistics with more informative and practical metrics like effect sizes and confidence intervals around effect sizes. Although such metrics have been underreported,[22–24] recent efforts are moving research practices in this direction.[12] In fact, many journals now require that these metrics be provided in all quantitative research papers.[25,26]

Sin #7: Lack of Transparency in Reporting Results and Maintaining Raw Data

Although author concerns about word count limits or lack of statistical sophistication may cause inadequate reporting, such practices also serve to cover up questionable research practices. For example, authors sometimes include basic information about descriptive statistics (eg, means) but fail to include standard deviations. To advance medical education, it is critical that authors maintain a high level of transparency in reporting results and retain the

integrity of their raw data for later analysis by other investigators (eg, data warehousing and data sharing repositories). Correct reporting and transparency of statistical analyses are important because statistical results from articles are used in meta-analyses. Thus, errors of reporting in primary level studies can lead to errors and bias in meta-analytic findings as well. Researchers should strive to provide full information on basic descriptive statistics (sample sizes, means, and standard deviations) and exact *P* values, regardless of whether or not they are significant. Last but not least, researchers should fully disclose all of their statistical analyses.

Summary

High-quality educational research is essential for producing generalizable results that can inform medical education. Although questionable research practices can be found in educational research papers, basic steps can prevent these "sins." After a study has been published it is quite difficult to determine if, when, and how the findings were influenced by questionable research practices; thus, a proactive approach is best. If spurious findings do find their way into the literature, the consequence is a knowledge base rooted in misleading, exaggerated, or entirely false findings. By avoiding the 7 deadly sins described here, medical education researchers will be in better positions to produce high-quality results that advance the field.

References

1. Ding M, Satija A, Bhupathiraju SN, et al. Association of coffee consumption with total and cause-specific mortality in 3 large prospective cohorts. *Circulation*. 2015;132(24):2305–2315.
2. Le Gal K, Ibrahim MX, Wiel C, et al. Antioxidants can increase melanoma metastasis in mice. *Sci Transl Med*. 2015;7(308):308re8.
3. Sayin VI, Ibrahim MX, Larsson E, et al. Antioxidants accelerate lung cancer progression in mice. *Sci Transl Med*. 2014;6(221):221ra15.
4. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
5. Artino AR Jr. Why don't we conduct replication studies in medical education? [letter]. *Med Educ*. 2013;47(7):746–747.
6. Maggio LA, Sewell JL, Artino AR Jr. The literature review: a foundation for high-quality medical educational research. *J Grad Med Educ*. 2016;8(3):297–303.
7. Baumeister RF, Leary MR. Writing narrative literature reviews. *Rev Gen Psychol*. 1997;1(3):311–320.
8. Open Science Collaboration. Psychology. Estimating the reproducibility of psychological science. *Science*. 2015;349(6251):aac4716.
9. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin; 2002.
10. Coladarci T, Cobb CD, Minium EW, et al. *Fundamentals of Statistical Reasoning in Education*. 2nd ed. Hoboken, NJ: Wiley & Sons; 2008.
11. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155–159.
12. Bakker M, van Dijk A, Wicherts JM. The rules of the game called psychological science. *Perspect Psychol Sci*. 2012;7(6):543–554.
13. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365–376.
14. Artino AR Jr, La Rochelle JS, DeZee KJ, et al. Developing questionnaires for educational research: AMEE guide no. 87. *Med Teach*. 2014;36(6):463–474.
15. Leppink J. Data analysis in medical education research: a multilevel perspective. *Perspect Med Educ*. 2015;4(1):14–24.
16. Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. *Behav Res Methods*. 2011;43(3):666–678.
17. Gigerenzer G, Swijtink Z, Porter T, et al. *The Empire of Chance: How Probability Changed Science and Everyday Life*. New York, NY: Cambridge University Press; 1997.
18. Levine TR, Weber R, Hullett C, et al. A critical assessment of null hypothesis significance testing in quantitative communication research. *Hum Commun Res*. 2008;34(2):171–187.
19. Gigerenza G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were too afraid to ask. In: Kaplan D, ed. *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage; 2004:391–408.
20. Gill J. The insignificance of null hypothesis significance testing. *Political Res Q*. 1999;52(3):647–674.
21. Rodgers JL. The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *Am Psychol*. 2010;65(1):1–12.
22. Cumming G, Fidler F, Leonard M, et al. Statistical reform in psychology: is anything changing? *Psychol Sci*. 2007;18(3):230–232.
23. Hoekstra R, Finch S, Kiers HA, et al. Probability as certainty: dichotomous thinking and the misuse of *P* values. *Psychon Bull Rev*. 2006;13(6):1033–1037.
24. Vacha-Haase T, Nilsson JE, Reetz DR, et al. Reporting practices and APA editorial policies regarding statistical

significance and effect size. *Theory Psychol*. 2000;10(3):413–425.
25. Sullivan GM, Feinn R. Using effect size—or why the *P* value is not enough. *J Grad Med Educ*. 2012;4(3):279–282.
26. Sullivan GM. FAQs about effect size. *J Grad Med Educ*. 2012;4(3):283–284.

Katherine Picho, PhD, is Research Assistant Professor, Department of Medicine, Uniformed Services University of the Health Sciences; and **Anthony R. Artino Jr, PhD,** is Deputy Editor, *Journal of Graduate Medical Education*, and Professor and Deputy Director for Graduate Programs in Health Professions Education, Department of Medicine, Uniformed Services University of the Health Sciences.

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Uniformed Services University of the Health Sciences, the Department of the Navy, the Department of Defense, or the US government.

Corresponding author: Katherine Picho, PhD, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814-4799, katherine.picho.ctr@usuhs.edu