Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Sport and Health Science 1 (2012) 9–11

www.jshs.org.cn

Opinion

Sadly, the earth is still round ($p < 0.05$)

Weimo Zhu

Department of Kinesiology & Community Health, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 2 February 2012; revised 4 February 2012; accepted 6 February 2012

Looking through any exercise science journals today, in fact any science journals including many top Science Citation Index (SCI) journals, one can easily find examples of the wide-spread "$p < 0.05$/significance" abuse phenomenon, i.e., if the p value from a statistical/hypothesis test is less than 0.05 (or 0.01 sometimes), a conclusion that "the results/findings are significant" is then drawn. The abuse is so severe that it is already seriously threatening the integrity of scientific inquiry.

Why is the popular p value practice a problem? An example may help to explain. When I teach my graduate research methods class, I usually conduct a survey about students' background on my first day's class so that I can prepare my teaching according to the students' background and needs. Two of the questions in the survey are about the students' undergraduate Grade Point Average (GPA) and the Graduate Record Examinations (GRE) scores. Table 1 illustrates 14 students' responses in 1 year's survey.

Say if I am interested in knowing the impact of undergraduate training on students' GRE test performance, I can run a correlation between GPA and GRE using the data in Table 1. The correlation coefficient (r) is 0.178, with a p value of 0.544. Since the p value is larger than 0.05, we can then conclude that there is no relationship between GPA and GRE. But let's go further and do a small experiment: We simply copy the sample data and paste them into the existing data set to increase the n in the statistical software we are using, and re-compute r and p value each time (Note: This experiment is only trying to make my point and SHOULD not be done in a real study!). We repeated this process eight times and summarized our computational results in Table 2.

As expected, r never changed since the original pair data remain the same even if the data were duplicated. To further confirm this observation, we also plotted the data when $n = 14$ and when $n = 126$. As illustrated in Fig. 1, the relationship was kept exactly the same, except that a "+" in the right illustration represents nine pairs of data, rather than one. However, p value reduced as the sample size increased and became less than 0.05 or "significant" when n reached 126 or at our 8th addition of data duplication. If we now draw a conclusion about the relationship between GPA and GRE based on the p value, we will arrive at a completely different one: GRE is significantly related to GPA. This clearly demonstrates the problem in drawing conclusions merely based on a p value since it is BIASED by the sample size! When a sample size is large enough, almost all statistical findings could get a p value less than 0.05 and become "significant"; in contrast, even if there is a high correlation, or a meaningful treatment effect, the p value could be larger than 0.05 if the sample size is small.

The problem of making a research conclusion based on merely p value has been criticized for a long time. It is nearly a century (over actually if we count Karl Pearson's work in 1901) since Ronald Fisher advocated the concept and procedure of hypothesis testing in 1925. Known today as "significance" testing, the hypothesis testing is the most widely used decision-making procedure in scientific research. Meanwhile, hypothesis testing has been criticized from the very beginning, mainly for three aspects[1–5]: (a) hypothesis testing (deductive) and scientific inferences (inductive) address different questions; (b) hypothesis testing is a trivial exercise, to which Tukey[6] drove home this point when he commented "the effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'Are the effects different?' is foolish"; and (c) hypothesis testing adopts a fixed level of significance (i.e., $p < 0.05$ or 0.01), which forces researchers to turn a continuum of uncertainty into a dichotomous "reject or do-not-reject" decision. Furthermore, as illustrated above, since a large sample size can lead to almost every comparison being "significant", this makes the word "significant" itself meaningless.

E-mail address: weimozhu@illinois.edu

Peer review under responsibility of *Shanghai University of Sport*

Production and hosting by Elsevier

doi:10.1016/j.jshs.2012.02.002

Table 1
Students' GPA and GRE scores.

Student ID	GPA	GRE
1	3.4	1025
2	3.8	1185
3	3.8	1440
4	3.7	910
5	3.6	1030
6	3.9	1310
7	3.5	1270
8	3.5	1100
9	3.7	790
10	3.2	1170
11	3.8	1400
12	3.1	1120
13	3.3	1200
14	3.5	1160

Abbreviations: GPA = Grade Point Average; GRE = Graduate Record Examinations.

Table 2
Impact of increasing sample size on r and p value.

Number of students	14	28	42	56	70	84	98	112	126
r	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178	0.178
p value	0.544	0.366	0.261	0.190	0.141	0.106	0.080	0.061	0.047

In 1970, a group of sociologists criticized extensively the p value practice in their book *The Significance Test Controversy*[7] (see also more recent similar publications *What if There Were No Significance Tests*? edited by Harlow et al.[8] and *The Cult of Statistical Significance* by Ziliak and McCloskey[9]). Almost 20 years ago, Cohen[3] published his well-known article *The earth is round ($p < 0.05$)*, in which he concluded that "After four decades of severe criticism, the ritual of null hypothesis significance testing (mechanical dichotomous decisions around a sacred 0.05 criterion) still persists." If we look at today's widely spread, much worse p value driven practice, we have to conclude *Sadly, the earth is still round ($p < 0.05$)*!

Knowing the p value based practice is wrong and seriously damaging to the scientific knowledge we are acquiring; it is time to take quick actions to stop the practice. Below is a quick summary what we should do concerning hypothesis testing:

1. NEVER draw a conclusion merely based on a p value.
2. In addition to p value(s), report both descriptive statistics (e.g., mean and SD) of variables being tested and the values of statistics themselves (e.g., t and F values).
3. Always include "statistically" as a prefix when using the word "significant" to describe a p value based finding.
4. Compute confidence intervals of any variables being tested and report them.
5. For correlations (rs), make a judgment on their meaning using either absolute criterion[10]:
   0−0.19 = No correlation
   0.2−0.39 = Low correlation
   0.4−0.59 = Moderate correlation
   0.6−0.79 = Moderately high correlation
   ≥0.8 = High correlation
   or compute the coefficient of determination by simply squaring the correlation coefficient. The coefficient of determination in this case describes the proportion of variability in the Y variable accounted for by the X variable.
6. For other statistical tests, report effect size,[11,12] which can be considered conceptually as the average group difference(s) after taking random factor out. In fact, many top scientific journals already have forbidden reporting p values only and require manuscripts to compute and report effect size.[13,14] I strongly recommend that the *Journal of Sport and Health Science* (*JSHS*) adopts this editorial policy.
7. Finally, further explain effect size under the context of "clinical/practical" significance[4,15,16] and link each unit of change in a dependent variable or outcome measure with its real life meaning, e.g., the true impact of a unit change in VO_2max or body composition on health.

In summary, Cohen[3] criticized the p value abuse as "the earth is round ($p < 0.05$)" almost 20 years ago. Yet, the words "significant/significance" are so attractive and researchers often jump to a "significant" conclusion even if the observed "$p < 0.05$" is merely the bias of a large sample size or a meaningless sampling variability. Sadly, while the misuse and abuse of "$p < 0.05$" have been well criticized in the literature and taken into account by many journals' publication

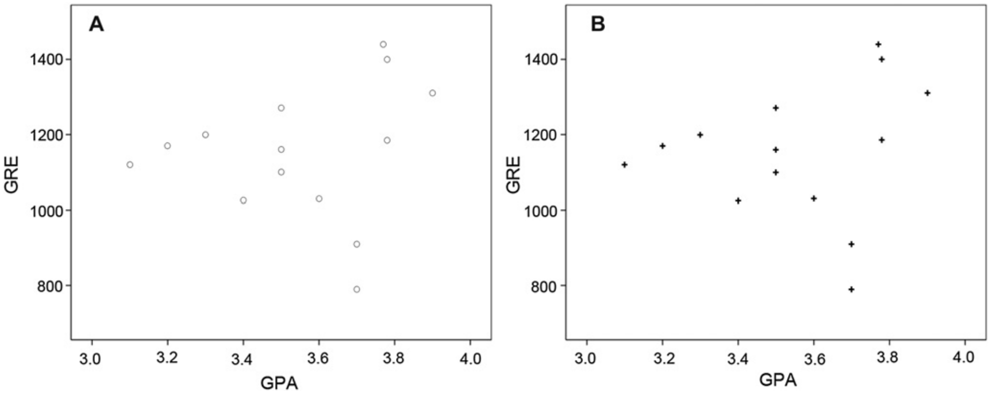

Fig. 1. Illustration of the correlation between Grade Point Average (GPA) and Graduate Record Examinations (GRE) (A: n = 14; B: n = 126).

guidelines, this inappropriate practice seems to be even more widespread now. To maintain scientific integrity, it is time to stop the p value practice and abuse. Suggestions on "should" and "should not" practice regarding statistical hypothesis testing are outlined. It is highly recommended that authors, reviewers, and editors of *JSHS* follow these suggestions.

References

1. Berkson J. Some difficulties of interpretation encountered in the application of the chi-square test. *J Am Stat Assoc* 1938;**33**:526−42.
2. Cohen J. Things I have learned (so far). *Am Psychol* 1990;**45**:1304−12.
3. Cohen J. The earth is round ($p < 0.05$). *Am Psychol* 1994;**49**: 997−1003.
4. Kirk RE. Practical significance: a concept whose time has come. *Educ Psychol Meas* 1996;**56**:746−59.
5. Lambdin C. Significance tests as sorcery: science is empirical-significance tests are not. *Theory & Psychology* 2012;**22**:67−90.
6. Tukey JW. The philosophy of multiple comparisons. *Stat Sci* 1991;**6**:100−16.
7. Morrison DE, Henkel RE, editors. *The significance test controversy.* Chicago: Aldine Publishing; 1970.
8. Harlow LL, Mulaik SA, Steiger JH, editors. *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates; 1997.
9. Ziliak ST, McCloskey DN. *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: The University of Michigan Press; 2008.
10. Safrit MJ, Wood TM. *Introduction to measurement in physical education and exercise science*. 3rd ed. St. Louis: Times Mirrow/Mosby; 1995. p. 71.
11. Cohen J. *Statistical power analysis for the behavioral sciences*. New York: Academic Press; 1969.
12. Huberty CJ. A history of effect size indices. *Educ Psychol Meas* 2002;**62**:227−40.
13. Altman M. Statistical significance, path dependency, and the culture of journal publication. *J Socio Econ* 2004;**33**:651−63.
14. Thompson B. AERA editorial policies regarding statistical significance testing: three suggested reforms. *Educ Res* 1996;**25**:26−30.
15. Meehl PE. *Clinical versus statistical prediction*. Minneapolis, MN: University of Minnesota Press; 1954.
16. Ogles BM, Lunnen KM, Bonesteel K. Clinical significance: history, application, and current practice. *Clin Psychol Rev* 2001;**21**:421−46.