Opinion Article

An opinion regarding equivalence testing for evaluating measurement agreement

---

ABSTRACT

The novel statistical approach 'equivalence testing' has been proposed in order to statistically examine agreement between different physical activity measures. By using this method, researchers argued that it is possible to determine whether a method is significantly equivalent to another method. Recently, equivalence testing was supported with the use of 90% confidence interval, obtained from a mixed ANOVA, which I believe is a more robust approach. This paper further discusses the use of this method in comparison to a more well-established statistical analysis (i.e. mixed design ANOVA), as well as various limitations and arbitrary assumptions in order to perform this analysis. The paper concludes with some remarks and considerations for future use in similar approaches.

Keywords: Mixed design ANOVA; p-value; confidence interval; methods' comparison

1. INTRODUCTION

Equivalence tests have gained some attention during the past two decades, mainly starting with applications in the pharmaceutical industry and biology. I recently came across the paper of Dixon and colleagues (2018) regarding the novel statistical approach 'equivalence testing' in order to statistically examine agreement between different physical activity (PA) measures and to evaluate the validity of a new method. The researchers suggest that the use of standard statistical tests of mean differences (e.g. ANOVA, t-test) is employed in similar research approaches, which generally focus in significant differences, rather than equivalences. I have to mention that this test is proposed only for group-level measurement agreement, because for individual-level agreement other tests are widely accepted as more adequate and valid (i.e. Bland-Altman plots, Mean Absolute Percentage Error and Root Mean Square Error of Approximation).

2. RATIONALE

Initially this statistical technique was introduced by Lee, Kim and Welk's [1] study, with the exception that there were not mentioned any p-values. They stated that 'in traditional hypothesis testing, the focus is on testing for a significant difference', however by 'using an equivalence test, it is possible to determine whether a method is significantly equivalent to another method' [1, p. 1843]. Since then, a number of studies have used this method in order to evaluate the agreement between methods in sport science [2-4].

In that initial approach [1], as well as some of following studies [4], equivalence was supported with the use of 90% confidence interval (CI), obtained from a mixed ANOVA. I believe that was a more robust approach, taking into consideration the misuse of p-values in order to support statistical hypotheses [5]. In fact, the American Statistical Association released specific guidelines on the use of p-values stating, among else, that p-values do not measure the probability that the studied hypothesis is true, do not measure the size of an effect or the importance of a result and not provide a good measure of evidence regarding a model or hypothesis. For this reason, the use of methods that emphasize estimation over testing was suggested, such as confidence, credibility, or prediction intervals and Bayesian methods [6]. In order to better understand the context and significance of this statement, Yaddanapudi's [7] editorial paper explains its salient features. To make it more concrete, the point in the American Statistical Association statement is not that p-values give the wrong answer; the point is that p-values usually commit what is called 'errors of the third kind: solving the wrong problem' and cannot be a good guide for probability testing [8].

The basic assumption made in order to justify equivalence testing was that standard statistical tests of mean differences are designed to detect differences, not equivalence and failure to reject the null hypothesis of no difference does not necessarily provide evidence of equivalence [9]. I am not convinced that this statement is correct. It is widely accepted that hypothesis testing is an important activity of empirical research. The initial null hypothesis (H_O) assumes that population means are equivalent and only if there is strong evidence to the contrary (alternative hypothesis; H_A), it can be assumed that there are differences among group means [10].

Furthermore, multivariate inferential procedures (i.e. repeated measures ANOVA) include hypothesis tests that allow several variables to be studied by preserving the significance level without inflating type I error rate [11]. The sample size is an issue, however with the correct use of appropriate tests, such as Pillai's trace, small or unequal sample sizes are not considered problematic, because the greatest protection against type I errors is offered [12]. Additionally, mixed-model designs are recommended in most cases because they can control for the repeated nature of the data (i.e. collection of data from PA monitors for multiple activities) [13]. This is not possible in equivalence testing, even though this approach might have limited value, because a single regression model is fitted to the average of the estimates throughout the range of all activities and not each activity separately [9].

Lastly, the confidence intervals for equivalence suggested by the authors (i.e. 10% and 2%) are somewhat arbitrary, an issue also highlighted by Dixon and colleagues [9]. This might be acceptable, since equivalence bounds in sport science are not set by regulations, as it happens for drug development (i.e., differences up to 20% are not considered to be clinically relevant). Such general regulations about what constitutes a meaningful effect seem unlikely to emerge, even though these could be extremely helpful and of increased value, especially in sport and exercise medicine. However, these intervals remain arbitrary and no statistically-based justification has been proposed in order to justify them.

Choice of equivalence bounds should be given careful thought, because the selected value will have enormous impact on sample size and interpretation of the observed results. An equivalence bound should be considerably smaller than the "clinically important difference" that would be used in a power analysis for assessing superiority between methods, and rationale for the chosen bound should be explained [14]. The value of an equivalence test is determined by the strength of the justification of the equivalence bounds. If the bounds chosen are based on the observed data, an equivalence test becomes meaningless [15].

3. CONCLUDING REMARKS

In future similar researches, I believe it would be more appropriate to compare the results derived from different statistical methods (i.e. equivalence testing *vs* mixed design ANOVA) and not only present the results from a single method. This approach could provide evidence of similarities and differences between the methods, so that the readers can understand more adequately what extra the new method has to offer. Lakens, Scheel and Isager [15] also recommend that researchers should perform both a null-hypothesis significance test and an equivalence test on their data, in order to improve the falsifiability of predictions in science.

However, in order to correctly address these results, CI and effect sizes, a set of statistics that indicates the relative magnitude of the differences between means [10] should also be calculated for all methods and not simply rely on p-values, as it happens nowadays with equivalence testing. Especially for effect sizes, the biggest challenge for researchers will be to specify the smallest effect size of interest, because not specifying a smallest effect size of interest for research questions at all will severely hinder theoretical progress [15].

Lastly, in order this attempt to introduce equivalence testing in sport and exercise science to be successful, the following considerations should be taken into account: a) Develop easy-to-use and accessible software; b) Express equivalence bounds in standardized effect sizes rather than raw scores; c) Related articles should discuss both power analyses and statistical tests for dependent t-tests, repeated measures or mixed design ANOVA and meta-analyses; d) Guidance should be provided on how to set a priori equivalence boundaries, given that there are often no specific theoretical limitations on how small effects are predicted to be nor cost-benefit boundaries of when effects are too small to be practically meaningful [16].

The interesting article of Dixon and colleagues [9] adds further to our understanding regarding the adequate use of equivalence testing for evaluating measurement agreement in sport science. While it is exciting to see increased attention to the development and dissemination of new statistical approaches and equivalence testing can provide another tool in the toolbox for scientists, they should be cautious about making and adopting statistical recommendations, because these could be considered as another ‘trend’.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Lee J-M, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc. 2014;46(9):1840-48. doi:10.1249/MSS.0000000000000287
2. Florez-Pregonero A, Meckes N, Buman M, Ainsworth. Wearable monitors criterion validity for energy expenditure in sedentary and light activities. J Sport Health Sci. 2017;6(1):103-10. doi:10.1016/j.jshs.2016.10.005

3. Kim Y, Welk GJ. Criterion validity of competing accelerometry-based activity monitoring devices. Med Sci Sports Exerc. 2015;47(11):2456-63. doi:10.1249/MSS.0000000000000691
4. Morris CE, Wessel PA, Tinius RA, Schafer MA, Maples JM. Validity of activity trackers in estimating energy expenditure during high-intensity functional training. Res Q Exercise Sport. 2019 [Epub ahead of print]. doi:10.1080/02701367.2019.1603989
5. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016;531(7593):151. doi:10.1038/nature.2016.19503
6. Wasserstein RL, Lazar NA. The ASA's statement on p-values: Context, process, and purpose. Am Stat. 2016;70(2):129-33. doi:10.1080/00031305.2016.1154108
7. Yaddanapudi LN. The American Statistical Association statement on p-values explained. J Anaesthesiol Clin Pharmacol. 2016;32(4):421-423. doi:10.4103/0970-9185.194772
8. Startz R. Not p-Values, Said a Little Bit Differently. Econometrics. 2019;7(1):11. doi: 10.3390/econometrics7010011
9. Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, Welk GJ. A primer on the use of equivalence testing for evaluating measurement agreement," Med Sci Sports Exerc. 2018;50(4):837-45. doi:10.1249/MSS.0000000000001481
10. Tabachnick BG, Fidell LS. Using multivariate statistics. 5th ed. Boston, MA: Allyn & Bacon; 2007.
11. Rencher AC. Methods of Multivariate Analysis. 2nd ed. New York, NY: John Wiley & Sons, Inc; 2002.
12. Park E, Cho M, K, CS. Correct use of repeated measures analysis of variance. Korean J Lab Med. 2009;29:1-9. doi:10.3343/kjlm.2009.29.1.1
13. Welk GJ, McClain J, Ainsworth BE. Protocols for evaluating equivalency of accelerometry-based activity monitors. Med Sci Sports Exerc. 2012;44(Suppl1):S39-49. doi:10.1249/MSS.0b013e3182399d8f
14. Mascha EJ, Sessler DI. Equivalence and noninferiority testing in regression models and repeated-measures designs. Anesth Analg. 2011;112(3):678-687. doi:10.1213/ANE.0b013e318206f872
15. Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. AMPPS. 2018;1(2):259-269. doi:10.1177/2515245918770963
16. Lakens D. Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. Soc Psychol Pers Sci. 2017;8(4):355-62. doi:10.1177/1948550617697177