REVIEW ARTICLE

Statistical Procedures Used in Pretest-Posttest Control Group Design: A Review of Papers in Five Iranian Journals

Nahid Dehghan Nayeri[1,2], Farshid Alazmani Noodeh[3], Hamid Sharif-Nia[4,5], Ameneh Yaghoobzadeh[6], Amir Hossein Goudarzian[7]

[1] *Department of Critical Care Nursing and Nursing Management, School of Nursing and Midwifery, Tehran University of Medical Sciences, Tehran, Iran*

[2] *Nursing and Midwifery Care Research Center, Tehran University of Medical Sciences, Tehran, Iran*

[3] *Department of Critical Care Nursing, Faculty of Nursing, AJA University of Medical Sciences, Tehran, Iran*

[4] *Department of Nursing, Amol Faculty of Nursing and Midwifery, Mazandaran University of Medical Sciences, Sari, Iran*

[5] *Psychosomatic Research Center, Mazandaran University of Medical Sciences, Sari, Iran*

[6] *Department of Nursing, School of Nursing, Arak University of Medical Sciences, Arak, Iran*

[7] *Department of Psychiatric Nursing, Tehran University of Medical Sciences, Tehran, Iran*

Received: 01 Oct. 2022; Accepted: 18 Aug. 2023

Abstract- The pretest-posttest control group design is one of the most widely used quantitative experimental design models for evaluating the efficacy of programs, treatments, and interventions. Despite the prevalence and utility of this research design, best practices for data analytical procedures are not clearly defined. Invalid results decrease the chance of generalization. Given that Iranian Journals are interested in publishing pretest-posttest control group design studies, it is important to denote the accuracy of them. The aim of the current study is to explore the correct procedure for using ANCOVA in pretest-posttest control group designs to mitigate the potential limitations of this approach. This study explores the use of ANCOVA in pretest-posttest control group design. It has been done by analyzing data from experimental studies published in five Iranian journals indexed in PubMed or Scopus between 2011 and 2018. The results indicate that among the 280 published experimental studies in these journals, 53 papers (18.9 percent) used ANCOVA as the statistical test in pretest-posttest studies. The power of the test represents the probability of detecting differences between the groups being compared when such differences exist. Our analysis concludes that ANCOVA, which runs a multiple linear regression, is a suitable method for comparing and examining pretest-posttest study designs. Implications of this study have potential utility for researchers employing the use of pretest-posttest control group designs in various fields in and outside of Iran.

Acta Med Iran 2023;61(10):584-591.

Keywords: Pretest-posttest study; Analysis of covariance; Nursing

Introduction

Pretest-posttest control group designs are a classical experimental design model that is widely used to examine changes or outcomes in an intervention or treatment group (case) by comparing data from before and after the intervention or treatment (1). This type of research design is particularly common in educational, medical, and psychological research (2).

The purpose of most experimental research is to determine the effect an intervention has on a treatment group (3,4). The first step involves randomly assigning individuals to the intervention and control group so that they can be grouped without bias, although this does not mean the groups are identical in terms of basic characteristics despite most researchers seeking to maintain a sufficient degree of homogeneity prior to randomization (5).

Corresponding Author: A. Yaghoobzadeh
Department of Nursing, School of Nursing, Arak University of Medical Sciences, Arak, Iran
Tel: +98 9375651975, E-mail address: a.yaghoobzadeh@yahoo.com

Statistical methods used to resolve the heterogeneity between groups in terms of basic data scores in pre-test and post-test studies are useful (6). Given that the statistical power is likely to detect a significant difference when there is a real difference between groups (7), (even if scores between groups are homogeneous and comparable), it is possible that the posttest scores would be unable to provide powerful results without considering the pretest scores (8). With this in mind, the minimum level of statistical power needed to determine the actual difference between treatments is estimated to be between 80% and 90% (9). There are four elements for determining the statistical power: 1. Statistical testing; 2. Determining the statistical significance level (α); 3. The sample size; and 4. The effect size (10). Although the first three elements are self-explanatory of their features, Cohen defines the effect size as the difference between the groups mean to one unit standard deviation instead of the standard error (11); hence, the effect size will increase by decreasing the standard deviation (12). In fact, the effect size represents the actual difference between the scores, regardless of how it relates to a general population. It also calculated based on the mean and standard deviation of the sample without considering the standard error (13).

It should be noted that only significance should not be considered a criterion, because the smallest variations may be significant in the large sample size (14). Although the effect size is estimated by several formulas, it is generally obtained in two ways: 1. the standardized difference between the two meanings (Cohen's d) and correlation (Pearson's r) between the independent variable and the scores in the dependent variable (15). The numerical values (Table 1) using Pearson's r is between zero and one, while the values of Cohen's d can be more than one (16). Another key concept in this field is statistical power, which measures the accuracy of the probability of rejecting the null hypothesis (14). It is influenced by four factors: 1. effect size; 2. significance level (α); 3. The test's power; and 4. number of participants in the study (13).

It is also necessary to consider the internal validity, which is to rely on the accuracy of the results and the causal relationship between the pretest and posttest scoring changes in the intervention group rather than the control group, in this type of study. Processes used to determine the external validity assist with strengthening the degree of generalizability of the treatment effects to the study population, treatment, and measuring instruments (17).

One of the threats to internal and external validity is the pretest score and the interactive effects between the intervention and the control group (18). The pretest score can have an effect on the posttest score, which is known as the Carryover Effect (19). This occurs as the initial treatment affects the response of the participants in the secondary treatment (20). In other words, the interactive effect occurs when the participants stimulated toward the intervention and treatment by pretest. Moreover, their responses (posttest scores) are different from those that did not meet the pretest scores (21). Participants with lower scores in pretest may have lower scores in the posttest compared to their counterparts with higher scores, even if the treatment has a significant impact on the posttest scores, while, it will be doubtful to make more changes in the results of the participants who had higher scores in the pretest (22). Although maturation and history are two major factors that threaten internal validity in this type of study, the interaction between the pretest score and the intervention is considered as a major and serious threat factor for the external validity (21).

Table 1. The effect size

Size	Pearson's r	Cohen's d
Small	.1-.3	<.25
Average	.3-.5	.25-.4
Large	.5-1	.4-∞

In general, there are broad opinions about best practices for data analysis for Pretest-posttest control group designs. Five methods are commonly used to analyze data based on the literature: 1. analysis of variance on the posttest score alone; 2. analysis of variance on different scores; 3. analysis of variance on the percentage change scores; 4. Analysis of covariance; and 5. Blocking of the primary scores (23). However, Repeated Measures ANOVA is also used when the scores measured more than twice after the intervention (21).

Analysis of covariance (ANCOVA) is one way to control for the effects of a pre-test score, so that conditions are provided to examine the effects of treatment (intervention) apart from the potential effect of the pretest score (24-26). It assists to determine whether the adjusted group mean is significantly different or not (27). ANCOVA, as the statistical method, control the threat of internal and external validity by controlling the effect of pretest scores in pretest-posttest studies. Therefore, reliable results will be provided (23). ANCOVA increases and strengthens the statistical power of the results By decreasing the error of the intergroup variance (27). To understand this, the F-statistic has to be recognized first which is considered as an assessment of

the difference between groups; The F-statistic is calculated by dividing the explained variance of the inter-group (for example, between the intervention and control groups) with an unexplained variance of intra-group. Therefore, it will be greater than critical value, it can be concluded that the difference between the two groups is significant (28). Also, the unexplained variance includes the error variance (e.g., the difference between individuals) with the effect of other factors (29). Therefore, the effect of the confounders, such as the pre-test score (covariate), affects the denominator. So, when control the effect of covariate score on the outcome score (for example, the post-test score) is controlled, in fact, the F-statistic will increase by removing it from denominator. As a result, the power the statistic will increase to find meaningful effects (if it exists) (30). Another use of ANCOVA is to modify the differences in the unequal groups, which aims to correct the difference between the groups that affects the posttest score (31). In addition, the covariate may be strongly related to the independent variable, which results in a significant difference in the outcome variable by removing the variance of the dependent variable (posttest) associated with the covariate (pretest). This also make the results meaningless (32).

As with other statistical tests, using ANCOVA needs to meet some presupposition as following: 1. Linearity of the regression (linear correlation between the dependent variable and the covariate); 2. Homogeneity of the error variance (error of random variable with a zero mean and an equal variance for different intervention and control groups); 3. Errors independency (the lack of correlation errors); 4. Normal distribution of errors; 5. Evaluation of variances homogeneity using Levene's test or Box test; and 6. Homogeneity of regression slope (the equality of regression slope of different lines among groups). In other words, there shouldn't be significant differences between groups in terms of the interaction of covariate scores and independent variables (33). Additionally, the slope of the regression lines needed to be the same among the groups for the covariates (in relation to the dependent variable). This presupposition known as homogeneity of the regression slope, which can be evaluated by a F test on the interaction of independent variables with covariates. If the F test was significant, this presupposition will be failed (34). The assumption of the linear relationship between the pre-test and post-test scores and the homogeneity of the slope of regressions are very important among the mentioned presuppositions as the researchers must first assess these two presuppositions (35).

A commonly reported disadvantage to pretest-posttest control group designs is due to carryover effects. Carryover effects initially arose as an issue in repeated measures clinical experiments when it was found that certain factors could ‘carry over’ from one treatment to another. One way of responding to this issue is to use randomization and statistical testing to control for such effects. Although chi-square test, t-test, and analysis of variance are considered important tests for controlling carryover effect, it is the analysis of covariance (ANCOVA) that has the greatest control power.

Incorrect data analysis practices can lead to confounding results in experimental research designs. While incorrect data analysis procedures can lead to confounding results in any experimental research designs, generalization is one of the most pressing concerns for research employing pretest-posttest control group designs. Invalid results decrease the chance of generalization (36). Given that Iranian Journals are interested in publishing pretest-posttest control group studies in a variety of fields including medicine, education and psychology, it is important to denote the accuracy of the correct procedures for data analysis. The aim of the current study is to explore effective practices for ANCOVA use in pretest-posttest control group design.

Methods

A systematic review was conducted in order to identify experimental and quasi-experimental studies with control group designs published in five Iranian journals indexed in PubMed or Scopus (i.e. Iranian Journal of Nursing and Midwifery Research, International Journal of Community Based Nursing and Midwifery, Hayat, Journal of Caring Sciences, Nursing and Midwifery Studies). For the first step of the review, one of the authors evaluated in detail the articles published during the period 2011-2018 (except one journal which was assessed from 2013). In the primary survey, articles were selected based on four inclusion criteria: 1) That the article was written in either English or Persian (abstract was English); 2) That the articles presented the results of pretest-posttest studies; 3) That the studies used a control group; and 4) a continuous dependent variable/continuous dependent variables. Articles were excluded if they met any of the following criteria: 1) Lack of a control group; 2) Use of more than two groups; 3) Only posttest design; 4) Use of a factorial design; and 5) Use of a cross-over design. Once these criteria were considered, 280 studies remained for final

assessment. Two researchers assessed the titles, material, methods, statistical procedures, and results section of articles separately in each phase. This was done in order to increase reliability of chosen papers.

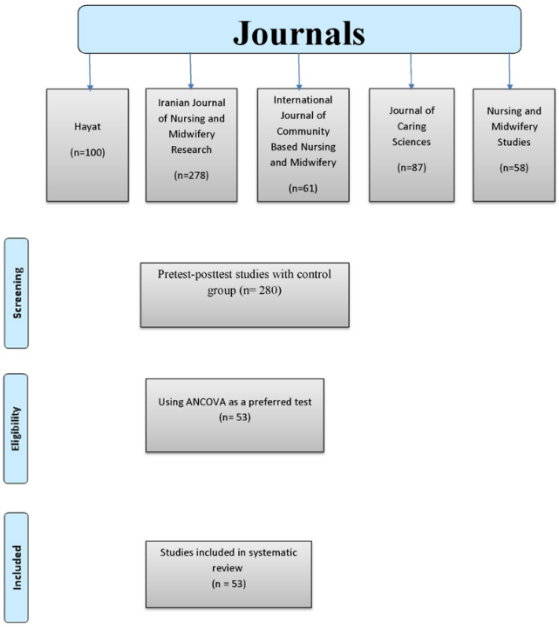

Figure 1. Number of entered studies

Results

The initial search identified 584 experimental studies published in five Iranian journals indexed in PubMed or Scopus were evaluated during 2011-2018. After examination of the abstract and full text of each paper, 280 papers were found to be eligible based on the inclusion and exclusion criteria. 70 papers (25 percent) controlled the base-line score while 53 papers (18.9 percent) did not use ANCOVA as the preferred test. One the remarkable result of the current study was that none of the evaluated papers report effect size.

"In order to clarify the differences between the results of the independent t-test and ANCOVA, the authors of the present study analyzed the unreal data of the study using these two different approaches, then compared the effect size of them."

Applicable example

a) A researcher designed the study with the aim of assessing the effect that face-to-face education had on the quality of life (QOL) of older adults with diabetes. This was the experimental study with pretest-posttest group. 60 older adults were referred to a clinic and then randomly allocated to the intervention and control groups based on the inclusion and exclusion criteria. The Diabetes quality of life (DQoL) instrument was then completed by the two groups as the pretest. The intervention group underwent 12 of sessions over a period of one month while the control group received no intervention. The control group had no contact with the intervention group during the study. At the end of the 12^{th} session, each group was given a questionnaire to answer. Data gathered from both groups before and after training then analyzed using SPSS version 25. The total scores calculated the pretest and posttest results of the two groups separately (Table 2). The normality of the data was tested using Shapiro-Wilk. Data were then analyzed using independent t-test (posttest in intervention and control groups) and ANCOVA (with pretest as covariate). The effect size of the two approaches was measured using G * Power 3.1 software.

Table 2. The scores obtained from Diabetes quality of life (DQoL) instrument

Group	Pretest	1st posttest	2nd posttest	3rd posttest
1	25	32	33	39
1	23	27	28	38
1	21	26	27	37
1	24	29	30	38
1	26	30	31	39
1	25	31	32	39
1	26	33	34	42
1	21	26	27	37
1	20	24	25	33
1	20	25	26	36
1	26	33	34	40
1	24	28	29	39
1	22	27	28	32
1	25	30	31	35
1	27	31	32	36
1	26	32	33	37
1	29	35	39	43
1	22	27	29	33
1	21	25	29	33
1	21	26	29	33
1	27	34	38	42
1	25	29	33	37
1	23	28	32	36
1	26	31	35	39
1	28	32	38	42
1	27	33	37	41
1	30	35	39	45
1	23	28	32	38
1	22	26	29	34
1	22	27	28	32
2	18	20	21	18
2	25	25	29	26
2	24	25	28	23
2	23	23	27	23
2	21	22	29	21
2	25	24	28	22
2	23	23	26	22
2	21	22	26	22
2	26	27	29	25
2	25	27	29	25
2	17	18	17	16
2	26	26	26	25
2	25	26	27	25
2	24	24	24	25
2	22	23	23	23
2	26	25	26	27
2	24	36	24	23
2	22	23	23	22
2	27	28	26	26
2	26	28	27	27
2	18	19	18	19
2	27	27	26	28
2	26	27	22	26
2	25	25	23	26
2	23	24	21	24
2	27	26	25	27
2	25	25	24	25
2	23	24	23	24
2	26	25	25	26
2	26	26	25	26

A) First approach: paired t-test

In the intervention group, the mean and standard deviation of QOL was 24.23±2.71 and 29.33±3.18 (P<.001) before and after the intervention, respectively.

The effect size was estimated to be 1.715.

Second approach: independent t-test

The mean and standard deviation of posttest QOL after the intervention was 25.00±3.41 and 29.33±3.18 (*P*<.001) in the intervention and control groups, respectively. The effect size was estimated to be 1.313.

Third approach: ANCOVA

Initially, the important presupposition of regression slope homogeneity was evaluated. The insignificant interaction effect of covariate scores and groups established this important presupposition (*P*=.227). Moreover, Levene's test confirmed the homogeneity of variance (*P*=.469). However, the mean and standard deviation of posttest score for QOL were 29.33±3.18 and 25.00±3.4 in the intervention and control group respectively. After the intervention and without modifying the pretest scores, the means of the intervention and control groups were 29.19 (CI95%:28.55-29.83) and 25.13 (CI95%:24.49-25.77) respectively after modifying the pretest scores. Finally, ANCOVA results indicated that this difference was statistically significant (*P*<.001). Eta was estimated to be .890. The estimated effect size was estimated at 2.850 using this approach.

B)

Imagine that three measurements are made on a weekly basis after the intervention and each week the training is given to the intervention group. The researcher is seeking to answer the question as to whether there is a significant difference between the intervention and control group after three weeks of education in three times measurements. In other words, is there any significant difference between the intervention and control groups over the three weeks education and repeated evaluation of the quality of life of diabetes patients? In order to analyze the data, the normality of the data in each group was tested using Shapiro-Wilk. Data were then analyzed using three different approaches: repeated measurement of analysis of variance (comparison of mean and standard deviation of four times of measurements in two groups); repeated measurement of analysis of variance by controlling the effect of pretest (comparison of mean and standard deviation of three times measurements in two rounds by considering pretest score as covariate); and, finally, ANCOVA test (considering pretest score, first posttest and second posttest scores as covariates). The effect size of the three approaches was measured using G * Power 3.1 software.

First approach: repeated measurement of analysis of variance

Changes in quality of life were measured four times using the correction of Greenhouse-Geisser and with the help of repeated measurement of analysis of variance. Changes in the scores for quality of life are different and significant (eta_2: .726, *P*<.001, F (2.40,139.50): 153.35). The effect size was estimated to be 1.627.

Second approach: repeated measurement of analysis of variance by controlling the effect of pretest

Changes in quality of life were measured three times using the correction of Greenhouse-Geisser and with the help of repeated measurement of analysis of variance. Changes in the scores for quality of life were different and significant**.** (eta_2: .571, *P*<.001, F (1.81,103.54): 75.86). The effect size was estimated to be 1.153.

Third approach: ANCOVA

Initially, the important presupposition of regression slope homogeneity was evaluated. The insignificant interaction effect of covariate scores and groups established this important presupposition (*P*=.404). Moreover, Levene's test confirmed the homogeneity of variance (*P*=.128). The mean and standard deviation of posttest score for quality of life in third time measurement were 37.50±3.45 and 23.90±2.79 respectively in the intervention and control group after three times intervention and without modifying the pretest scores, first posttest, and second posttest score. However, the means of the intervention and control group were 37.12±3.65 (CI95%:36.18-38.073) and 24.27±3.65 (CI95%:23.32-25.21) respectively after modifying the pretest scores. Finally, ANCOVA results indicated that this difference was statistically significant (*P*<.001). Eta was estimated to be .807. The effect size was estimated at 2.050 using this approach."

Discussion

In designing pretest-posttest with control group studies, some of the researchers use independent t-test, paired t-test, or repeated measure ANOVA instead of using ANCOVA in order to assess their hypothesis. In all these methods, the use of pretest scores helps to reduce error variance, thus producing more powerful tests than designs with no pretest data (21). Generally speaking, the power of the test represents the probability of detecting differences between the groups being compared when such differences exist.

It seems that the ANCOVA test, which runs a multiple linear regression, is a suitable method for comparing and

examining pretest-posttest study designs. Generally, it has more power to discover the effects of treatment and intervention (37). ANCOVA controls the threats of internal and external validity by combining the regression and ANOVA as the statistical test. It also controls the effects of the pretest score in pretest-posttest studies. It thus provides more reliable results (23). In other words, increasing reliability can enhance the generalization of the study results (38). ANCOVA increases and strengthens the statistical power of the results by reducing the error of the intergroup variance (27). Hence, the following consideration is necessary; random allocation of samples is necessary for accuracy and estimation of effect sizes.

There are several different approaches for comparing and analyzing the scores obtained in experimental studies of two groups with pretest and posttest. The estimated effect size using ANCOVA was greater than that for other tests. It seems that ANCOVA, which runs a multiple linear regression, is an appropriate approach for comparing and analyzing these kinds of studies. While some researchers believe that, even if ANCOVA presuppositions are not established, researchers can still use this test and do not need to re-analyze the data, others hold that if the important presupposition of regression slope homogeneity is not recognized then other methods, such as Quade's nonparametric ANCOVA, Puri and Sen's nonparametric ANCOVA, or parametric ANCOVA, should be used for ratings. Finally, we insist on these two points: If groups are not randomly assigned, researchers should interpret the results of this type of study carefully and estimate and report the effect size with caution.

Acknowledgments

This article is the assignment of PhD course. I, Ameneh Yaghoobzadeh, would like to thank my professors, Professor Nahid Dehghan Nayeri and Dr. Hamid Sharif Nia who made this opportunity for me to learn more.

References

1. Schmidt NA, Brown JM. Evidence-Based Practice for Nurses: Jones & Bartlett Learning, LLC; 2014.
2. Dugard P, Todman J. Analysis of pre-test-post-test control group designs in educational research. Educ Psychol 1995;15:181-98.
3. Johnson B, Christensen L. Educational Research: Quantitative, Qualitative, and Mixed Approaches: SAGE Publications; 2010.
4. Lee J, Lee Y, Kim MH. Effects of Empathy-based Learning in Elementary Social Studies. Asia Pac Educ Res 2018;27:509-21.
5. Zhao W, Berger V. Imbalance control in clinical trial subject randomization—from philosophy to strategy. J Clin Epidemiol 2018;101:116-8.
6. Roberts MC, Ilardi SS. Handbook of Research Methods in Clinical Psychology: Wiley; 2008.
7. Ellis PD. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results: Cambridge University Press; 2010.
8. Vogt WP, Gardner DC, Vogt ER, Haeffele LM. Selecting the Right Analyses for Your Data: Quantitative, Qualitative, and Mixed Methods: Guilford Publications; 2014.
9. Chase LJ, Chase RB. A statistical power analysis of applied psychological research. J App Psychol 1976;61:234-7.
10. Lipsey MW. Design Sensitivity: Statistical Power for Experimental Research: SAGE Publications; 1990.
11. Cohen J. Statistical Power Analysis for the Behavioral Sciences: Taylor & Francis; 2013.
12. Leong FTL, Austin JT. The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants: SAGE Publications; 2006.
13. Mayers A. Introduction to Statistics and SPSS in Psychology: Pearson Education Limited; 2013.
14. Haghdoost A. Do You Want to Gain a Profound Insight into Sample Size and Statistical Power. Iran J Epidemiol 2009;5:57-63.
15. Rosnow RL, Rosenthal R. Definition and interpretation of interaction effects. Psychol Bull 1989;105:143-6.
16. Sullivan GM, Feinn R. Using effect size—or why the P value is not enough. J Grad Med Educ 2012;4:279-82.
17. Adams KA, Lawrence EK. Research Methods, Statistics, and Applications: SAGE Publications; 2014.
18. Lavrakas PJ. Encyclopedia of Survey Research Methods: A-M: SAGE Publications; 2008.
19. Cohen L, Manion L, Morrison K. Research Methods in Education: Taylor & Francis; 2007.
20. Polit DF, Beck CT. Nursing research: Generating and assessing evidence for nursing practice: Lippincott Williams & Wilkins; 2017.
21. Dimitrov DM, Rumrill Jr PD. Pretest-posttest designs and measurement of change. Work 2003;20:159-65.
22. Grove SK, Burns N, Gray J. Understanding nursing research: Building an evidence-based practice: Elsevier Health Sciences; 2014.
23. Bonate PL. Analysis of pretest-posttest designs: CRC Press; 2000.
24. Verma J. Data analysis in management with SPSS software: Springer Science & Business Media; 2012.
25. Lin MH. Effects of classroom blogging on ESL student

writers: An Empirical reassessment. Asia Pac Educ Res 2014;23:577-90.
26. Reising DL, Carr DE, Tieman S, Feather R, Ozdogan Z. Psychometric testing of a simulation rubric for measuring interprofessional communication. Nurs Educ Perspect 2015;36:311-6.
27. Tabachnick BG, Fidell LS, Ullman JB. Using Multivariate Statistics: Pearson; 2018.
28. Pace L. The Excel Data and Statistics Cookbook, Third Edition: Lulu.com; 2013.
29. Jackson S. Research Methods: A Modular Approach: Cengage Learning; 2007.
30. Kirk RE. Experimental Design: Procedures for the Behavioral Sciences: SAGE Publications; 2012.
31. Sapp M. Basic Psychological Measurement, Research Designs, and Statistics Without Math: Charles C. Thomas Publisher; 2006.
32. Miller GA, Chapman JP. Misunderstanding analysis of covariance. J Abnorm Psychol 2001;110:40-8.
33. Field A. Discovering Statistics using IBM SPSS Statistics 4e + eBook + WebAssign Single Term + SPSS Version 23.0. 5th ed. London: SAGE Publications; 2018.
34. Leech NL, Barrett KC, Morgan GA. IBM SPSS for Intermediate Statistics: Use and Interpretation, Fifth Edition: Taylor & Francis; 2014.
35. Gliner JA, Morgan GA, Harmon RJ. Pretest-posttest comparison group designs: Analysis and interpretation. J Am Acad Child Adolesc Psychiatry 2003;42:500-3.
36. Lee JH. Experimental methodology in English teaching and learning: Method features, validity issues, and embedded experimental design. Engl Teach Pract Crit 2012;11:25-43.
37. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. Br Med J 2001;323:1123-4.
38. Bonita R, Beaglehole R, Kjellström T, santé Omdl, Organization WH. Basic Epidemiology: World Health Organization; 2006.