Testing the Efficiency of Add-On Effects with Crossover Design

Jiangtao Gou and So Young Lim

Abstract

This research is aimed to compare students' academic performance based on two different teaching methods, traditional pedagogy and a lecture with in-class activity called Paper Toss game. A randomized, modified crossover design was used to measure the outcomes based on the scores from pretest and posttest. In general, the treatment effects were mainly focused for analysis in crossover designs. But, we were interested in analyzing data with the period effects. Three different approaches, distribution-based methods, Generalized Estimating Equation (GEE) and Generalize Linear Mixed Model (GLMM) were investigated by using R programming. For analysis using the distribution-based method, T-test was performed. The R functions of geeglm in packages geepack for GEE and the function of glmer in packages lme4 for GLMM were utilized to analyze the data. Even though some useful finding for evaluation purposes were revealed through the research framework, there was no significant impact on students' performance with additional teaching method.

Introduction

Crossover designs have been broadly introduced in the field of statistics in recent years and are largely used in researches of innumerable disciplines such as medicine, pharmacy,

manufacturing, engineering, education and so on. The crossover study is defined as a longitudinal or repeated study such that experimental units receive one treatment to the other during the different course of the trials. In other words, the same number of treatment is given to the participants during the same number of study periods. So, all the participants can serve themselves as the best control of their own in a different sequence of treatments. In a crossover design, the order of treatment administered to the experimental units is called a sequence and the time each treatment is provided is called a period. According to Connelly, (2014), the treatment effect is defined as the extent of improvement due to the intervention being tested while the period or order effect is described as the possible influence of the sequence on the intervention's effectiveness. The main purpose of this study is to separate the treatment effect from the period effect (Welleck & Blettner, 2012).

Depending on the purpose of the research, the crossover design can be developed in a various range of complexity, from the simplest model such as *AB/BA* to a complex combination model like *ABCD/BDAC/CADB/DCAB*. The most prevalent crossover design is comprised of two sequences, two periods, and two treatments in a setting. Basically, participants are randomly assigned into one of two groups. The first group receives a treatment, *A,* in the first period, and then received another treatment, *B,* in the second period while the second group receives the treatments in the reverse order, *B* first and then *A*. Moreover, researchers can conduct a constant treatment administration, such as *A*, *A* and *B*, *B,* depending on the purpose of the study. With more than two treatments, many different combination of study design can be suggested.

With examples, a variety of models in crossover design could be understood easier. Wilson at el. (2016) conducted a research using a randomized crossover design to evaluate the impact of

10-minute activity break outside the classroom on on-task behavior of young boys. For a subject in this research, 58 boys aged 11.2 ± 0.6 years were recruited from elementary school in Adelaide, South Australia. A half of the group completed four weeks of the 10-minutes active lesson break followed by four weeks of the passive lesson break (*A/B*). And, the other half of group did the reverse(*B/A*). Wilson at el. (2016) used the simplest crossover design, but, more complexity is possible. For instance, Taib at el. (2012) hypothesized that consuming lactose–isomaltulose-containing milk during morning will boost children's attention and memory resulting in improvement of cognitive performance. And, they conducted a research using 4 × 4 Latin Square design, which is an unique type of crossover design such that each treatment occurs only once within each sequence and once within each period (Vonesh & Chinchilli, 1997). There were four different test products, a glucose-malto- dextrin drink (*A*), and three different milk products: standard growing up milk (*B*), reformulated growing up milk (*C*), and standard growing up milk with isomaltulose (*D*) were provided to a total of 30 children. Each product was given for four different days, and the order of the four treatment was counterbalance between subjects. That is, an order of treatment, *ABCD,* for the first sequence, *BCDA* for the second, *CDAB* for the next and *DABC* for the last were administered.

One of the chief advantages of crossover design is to reduce a risk of having confounding variables. Participant variation can be unconcerned in crossover design. As mentioned above, each subject performs as their own controls for evaluation in a randomized setting. Because the participants receive both treatments, researchers also can evaluate them and express preferences for and against particular treatments (Mills et al., 2009). Besides, requiring fewer participants can be vital when study populations are extremely limited. In general, statistical analysis requires a

large number of randomized sampling in order to ensure the reliability of a research. This design is statistically very efficient because the feasible sample sizes for a study can be much smaller compared to non-crossover designs. Stated differently, less numbers are needed to meet the same criteria for reducing Type I and Type II errors (Connelly, 2014) and expected to have a compatible level of statistical power. For instance, an interesting approach to an altered crossover design was suggested by Reich & Milstone (2013) that randomizing by group or cluster in crossover design could be more advantageous compared to randomizing by individuals in certain setting. Additionally, the authors made comparison between crossover trial and non-crossover trial in a various range of clusters and found that a study conducted with a non-crossover design is likely to require approximately 40 more clusters in order to achieve the same statistical power as a study with crossover design. They also proved that the standard cluster-randomized design retains 50% less power than the cluster-randomized crossover design. Thus, statistical power gained from comparatively small number of participants in crossover design can be beneficial for studies with extreme case of recruiting experimental units.

The pivotal limitation of the crossover design is the potential occurrence of carryover effect, in which the effect of the first treatment remains until the following course of the treatment is applied leading to a contamination of the next treatment effect. Accordingly, an effect from the treatment accomplished for the very last possibly show better effectiveness if the carryover effect from the previous treatment plays a role as a booster or it could show less effectiveness if the carryover effect from the previous treatment plays a role as a retractor. Often, this carryover effect results in a cause of statistical bias. In order to thwart this problem, researchers must administrate a washout period between treatments such that adequate time is given to participants between

treatment to eliminate any possible effect transferred. Thus, Welleck & Blettner (2012) strongly suggested that a researcher must know the length of time sufficient for the treatment to be effectively washed out. In the example of the research conducted by Wilson at el. (2016) stated above, the authors administered a two-week washout period in between two treatments, four weeks of the 10-minutes active lesson break (*A*) and four weeks of the passive lesson break (*B*). And, in the other example of study, Vonesh & Chinchilli (1997) performed their research with one week of a washout period between each test day. Since concerns with carry-over effects or residual effects cannot be dismissed, washout periods between the experimental treatments must be amply implanted, in that, theoretically, experimental units are supposed to restore their baseline status.

In addition, the use of this designs, especially for the clinical trial, are very restricted to patients who from chronic, stable conditions such as diabetes. For example, Helge et al., (2012) hypothesized that the increased fat-rich dietary intake would lead to a higher content of muscle ceramide as well as lower insulin sensitivity compared to carbohydrate-rich dietary consumption and conducted a study using a crossover design. Eleven patients with type 2 diabetes who categorized based on the WHO classification participated in this research. The participants were randomly assigned to undergo the two different types of diet. Fat-rich diet was composed of food containing protein, fat and carbohydrate in a 1:3:1 ratio while carbohydrate-rich diet was made of in ratio of 1:1:3 respectively. Each diet was consumed for three weeks, and about five weeks of washout period was administered under no restrictions of daily diet and physical activity. A key finding in this study was that there was no difference in muscle ceramide content and insulin sensitivity between fat-rich diet and carbohydrate-rich diet in type 2 diabetes patients.

The moderate medical conditions will not show dramatic changes in the duration of the

study, so the effect of the treatment is expected only to alleviate symptoms but not to cure the disease completely. This is one of the reasons why crossover design has been recognized as one of the most effective methods, predominantly for medical and pharmaceutical studies. Here is another example. Lin et al. (2011) used an experimental crossover design to investigate whether Montessori method could improve the eating ability of twenty-nine residents with dementia in long-term care facilities. Subjects were randomly assigned into two different interventions, either the Montessori(A) or control(B). A two-period crossover design was used, with fifteen residents assigned to the sequence of Montessori intervention-routine activities **(**A/B) and fourteen residents assigned to the sequence of routine activities-Montessori intervention (B/A). There was a two-week washout between each intervention period. Because all residents served as their own control, residents were unaware of which intervention sequence of the study they were involved in. As a result, the authors made a conclusion that Montessori-based activities could help the participant promote their self-eating time while reducing the time being fed by their caregivers. Thus, applying a crossover design in this research was very useful because no complete cure for the patients with dementia was made in the experiment.

Taking advantages of using crossover designs in educational research has been enlightened in comparatively recent years. Like all other disciplines, education has always undergone dramatic shift as an advancement in technology leads to an educational innovation of teaching methods. Since technology-enhanced teaching methods are ceaselessly introduced to educational institutions, researches in teaching evaluation has drawn attention to improve effectiveness of new teaching strategies. However, researches in school setting tend to be considered more complex compared to other studies. According to Berliner (2002), in education, broad theories and

ecological generalizations often fail because they cannot incorporate the enormous number or determine the power of the contexts within which human beings find themselves. In other words, unlike a scientific laboratory, a school is an organic, dynamic, complex institution with hundreds of variables intervening to complicate the "experiment" (Gordon, 2007). Both Berliner (2002) and Gordon (2007) could not emphasize enough the fact that educational researches required sophisticated research designs because a large number of contexts such as individual differences, school atmosphere, teaching methods, socioeconomic status, interaction and so forth can affect not only spontaneously but also impulsively, so the consistencies across the nature of numerous factors cannot be easily found for scientific analysis in education.

In particular, the question of how effectively new teaching method can be evaluated has been always a great challenge faced by the researchers. In order to explore this question, researchers need to choose proper mythologies among a variety of statistical techniques including qualitative and quantitative research methods. In education, qualitative research methods are preferred over quantitative since data is commonly collected through observation. Thus, a case study method has been selected as one of the most popular implementations in educational researches. For example, Nedungadi et al. (2010) conducted a study to create a realistic mathematics and science laboratory environment by using adoptive simulations called Amrita learning. As a formative evaluation, the authors adopted a case study method in which participants were randomly assigned into two groups, either the traditional lab or simulations. The average scores from pretest and posttest from both groups were collected to compare the improvement of student learning. Even though both groups showed similar learning enhancement overall, the average score from the group using the adaptive simulation was about 5.4% higher compared to

the scored from the group using traditional laboratory. Along with a case study, survey or questionnaires are also commonly used in educational research because these are convenient to get a direct feedback from participants. The research conducted by Nedungadi et al. (2010) also used questionnaires to compare feedback from participants using the traditional laboratory and the simulation. Authors allowed the participants in the traditional lab group work on the simulation and the participants in the simulation try out the physical experiment after the posttest were made, but before the feedback was given. They found that the majority responded that the simulation was very useful to learn the topics because it provided additional learning options that the physical experiment cannot provide.

However, these methods have an inevitable limitation in which they cannot provide comparison of the efficacy between traditional standard lectures and desirable pedagogies without concerning between group variability. The participants in the study by Nedungadi et al. (2010) learned through simulation or traditional style three times a week for a six week-period. And then, they tried out the other teaching method after posttest were made. But, the period they tried the other teaching method did not be specified in this paper. Since participants already had some idea about what they were supposed to learn, it is hard to measure how they exactly feel about the other teaching method they did not involve. That is, if researchers make a conclusion only based on what the participants respond in a survey, it could be deemed as very objective, not subjective because how they feel about their achievement during experiment they involved could be drastically different from the analysis based on scientific method of actual achievement they made.

Subsequently, researchers have kept trying to find a different approach when it comes to evaluating effectiveness of teaching method without testing at the level of the teaching activity.

An innovative approach to educational evaluation was proposed by Smith at el. (2012) by using a student engagement framework. Their approach was derived from an idea that since student engagement can be seen as a measurement of how actively they absorb learning materials, determination of the level of students'' engagement can be a great means to evaluate learning outcomes. So, the authors modified two well-known surveys from the United States and Australia and used as a tool to evaluate a specific online learning. The sample population were four hundreds twenty-one student enrolling in a public health course. Participants evaluated the online role-play two weeks after they completed and three days after they submitted an assessment. And, online survey comprised thirty-seven Likert questions rated on a five-point Likert scale was delivered by e-mail and one hundred forty-three students responded. Based on report made by student, the conclusion was made that the student engagement framework could be useful in some way for the evaluation purposes. Again, even though collecting data using survey or questionnaire could be very practical and easy to access, there is a doubt on its validity how truthful a respondent is being. It is not deniable that making a scientific conclusion using these methods is hard to control possible cause of variation from the response variables.

In order to avoid having this bias, crossover designs are being applied in educational studies to compare learning outcomes based on the different teaching methods. By using crossover methods, between experimental unit variation can be disregarded because each unit also plays a role as control. In fact, the statistics validity with using crossover design in educational research has been proved with the previous literatures. For example, Barry and colleagues (2016) conducted a study using a randomized crossover design in order to find how team-based learning (TBL) and traditional lecture could differently affect student learning outcomes and their confidence. A total

of 30 students were participated and divided into two groups in crossover design setting. Participants learned therapeutic topics either for first three weeks in a way of TBL and another three weeks in traditional pedagogy or vice versa. After each teaching method is completed, assessment examinations were given to test application, recall and combined examination. And, in the end of semester, a confidential survey with Likert scale questions was made to measure their confidence toward materials they learned. Researchers analyzed the data collected with paired t-test and direct comparison using Cohen's d and found that overall student assessment scores with TBL was significantly higher compared to scores with traditional teaching method. Moreover, student confidence ratings were significantly higher when they studied in a TBL strategy. Overall, active learning and engagement motivated by TBL promote self-study resulting in better learning consequences along with confidence of mastering materials.

Many of previous researches using crossover study mainly focus on assessment of the treatment effect. However, current literatures provide no scientific evidence in the efficacy of teaching method by interpreting the period effect. Thus, we sought to evaluate the period effect applying crossover design in a setting of teaching evaluation.

Distribution-based methods

Milliken & Johnson (1984) suggest that the objective of crossover designs is to eliminate variation in comparing treatments by observing all treatments on the same experimental unit. That is, researchers randomly assign N experimental units to the pth sequence where S sequences of the T treatment are constructed to compare treatment effects.

In general crossover design, treatment effects, period effects and carry-over effects must be obtained. Even though possible presence of carry-over effect is anticipated, the study objective

is mainly focused on how effectively treatments are practiced, not evaluating how carry-over effects are present. With having sufficient wash-out period, carry-over effects tend to be neglectable for researchers.

For the N_i experimental unit, the logit probabilities of the responses in periods 1,2 are α and $\alpha + \pi + \tau + \lambda_A$ respectively in the first group receiving treatment *A/B* and $\alpha + \pi$ and $\alpha + \tau + \lambda_B$ respectively in the second group, receiving treatment *B/A* whereas π and τ are treatment and period effects respectively and λ_A and λ_B are carryover effects. The model for 2×2 crossover design is shown in Table 1.

Table1: The model for 2×2 crossover design

Sequence	Period 1	Period 2
AB	α	$\alpha + \pi + \tau + \lambda_A$
BA	$\alpha + \pi$	$\alpha + \tau + \lambda_B$

Noted that, in order to test the difference between two quiz questions, the hypothesis is tested with H_0: π=0. And, the hypothesis to test for the effectiveness of the additional teaching technique is H_0: τ=0. For Testing carryover effect, the hypothesis is suggested to be H_0: $\lambda_A = \lambda_B$.

By referring to a model introduced by Milliken & Johnson (1984), the response of an observation from the *m*th experimental unit who receives the *i*th sequence and *k*th treatment in time period *j* is shown as $y_{ijkm} = \mu_{ijk} + \varepsilon_{im} + e_{ijm}$ where $i = 1, 2, \ldots, s$, $j = 1, 2, \ldots, t$, $m = 1, 2, \ldots, t$, $k = 1, 2, \ldots, n_i$. The μ_{ijk} is the mean response with treatment *k* applied in sequence *i* at time period *j*, $\varepsilon_{im}$is the random error associated with the *m*th experimental unit in sequence *i*, and e_{ijm} is the random error associated with time interval *j* of experimental unit *m* in sequence *i*. Then, model can be rewritten as $y_{ijkm} = \mu + \tau_k + \Pi_j + \sum_{r=0}^{j-1} \lambda_{jk} + \varepsilon_{im} + e_{ijm}$ where τ_k is the treatment

effect, Π_j is the time interval effect, and λ_{jk}, denoted the carryover effect of the k_rth treatment, which is administered in time period r = 0, 1, …, j-1 on time period j where $\lambda_{jk_r} = 0$ for all r where r < j-1, j = 2, 3, …, t. Consequently, the model can be rephrased as $y_{ijkm} = \mu + \tau_k + \Pi_j + \lambda_{k_{j-1}} + \varepsilon_{im} + e_{ijk}$. This is the naivest way to interpret a crossover design and can be analyze by using T-test.

Generalized Estimating Equation (GEE)

The generalized estimating equation (GEE) was originally introduced by Liang and Zeger (1986), which is an extension of generalized linear model. GEE methods are useful to analyze longitudinal data especially when response variables are numerical, either discrete or continuous. GEE approaches are often preferable in the situation where application of likelihood based methods is not valid. That is, GEE can be used in a various range of analysis including data with Poisson, binomial and gamma distributions. Valois (1997) stresses that with GEE, different subjects can have different numbers of repeated measurements and these measurements do not need to be taken at the same time intervals for all subjects. In longitudinal studies, the general structure of repeated measurements layouts each observation containing a subject identifier, an observation number, a time identifier, a response and some covariates.

Based on the GEE layout for cross design introduced by Valois (1999), each experimental unit, N_i, where i = 1, 2, … ,n has a vector of p response $Y_i' = (y_{j1}, y_{j2}, \dots, y_{jp})$. A matrix X_i of m = [1 + π + τ + λ] whereas the intercept, π , τ and λ described as treatment effects, period effects, carryover effects covariates as $X_i = \begin{pmatrix} 1 & x_{j11} & \cdots & x_{j1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{jp1} & \cdots & x_{jpm} \end{pmatrix}$. Therefore, the vector (y_{jk} , x_{jk1} , x_{jk1},

$\dots, x_{jkm}$) is observed at time t_{jk}, for experimental unit, *i* at period *k*. All the covariates will take the value of 0 or 1. In this agenda, the covariance structure does not necessarily be specified in order to earn reasonable estimates of regression coefficients and standard errors.

In general, with the GEE, link function can be any $g(\mu_i) = x_i^T \beta$ such as identity, log or logit. The random component can be any distribution of the response like binomial, multinomial, normal, etc. that The advantages gained by using the GEE is that it is computationally easier than when categorical data is associated compared to using maximum likelihood estimation. Moreover, The GEE does not require multivariate distribution to compute and it can estimate misperceived correlation structure correctly. However, there are also downside of using the GEE. Since there is no likelihood function involved, the GEE are not able to specify the joint distribution. Besides, the standard errors with the GEE tend to underestimate the true measure unless sample size are large enough.

Generalized Linear Mixed Model (GLMM)

McCullagh and Nelder (1989) firstly proposed an extension of linear models, called generalized linear models (GLMs). Unlike other linear models, the GLMs are applicable to situations where the observations are not only continuous, but also discrete, or categorical. Since the observations do not require to be continuous, a variety of models that includes normal, binomial and Poisson is also applicable with GLMs. Compare to a classical linear model, the mean of the observation in GLMs is associated with a linear function of some covariates through a link function the variance of the observation is a function of the mean (McCullagh and Nelder, 1989). The GLMM is extended from GLMs by including random effects in the predictor. The GLMM can also be seen as an extension of Gaussian linear mixed model based on two key elements they share.

Both the GLMM and Gaussian linear mixed model hold conditional independence given the random effects and a conditional distribution of the random effects.

In a crossover design setting, we followed mixed logistic model with crossed random effects suggested by McCullagh and Nelder (1989). For the experiment, let u_i and v_j be the random effects corresponding to the *i*th a treatment effect and *j*th a period effect. Then, on the logistic scale, the probability is modeled in term of fixed effects+ $u_i + v_j$. We can also assume that the random effects are independent and normally distributed with means 0 and variances σ_i^2 for the treatment effect and σ_j^2 for the period effect. With the binary responses, y_{ijk} are conditionally independent such that $\text{logit}\{P(y_{ijk} = 1|u, v)\} = x_{ij}\beta + u_i + v_j$. Here y_{ijk} represents the *k*th binary response corresponding to the same pair of *i*th a treatment effect and *j*th a period effect and x_{ij} is a vector of fixed covariates, and β is an unknown vector of regression coefficients. Therefore, the GLMM is often referred to as conditional models in contradiction of the marginal GEE models.

Methods

We proposed an altered crossover design which can be used to evaluate a new teaching method in a single class in term of period effect. The idea of Simpson's paradox (Simpson, 1951) is demonstrated as one of the important topics to discover in statistics class. It is an interesting topic, but, for students, it is also viewed as one of the most confusing statistical concepts to grasp. Simpson's paradox is an exceptional case of confounding in which combining data from several dissimilar categories into a single category conveys revered association of two variables as a consequence. With confounding variables, researches are often misdirected to the conclusion with spurious reasoning.

To help students comprehend the concept of Simpson's paradox, In-class activity called

Paper Toss game was designed to a subject of 69 student volunteers in an undergraduate level course STAT 213 Introduction to Applied Statistics class at Hunter College of CUNY in fall 2015. This activity is simply a paper toss competition between a student and an instructor. Participants throw several crumpled pieces of paper balls into a bucket and count how many times they can succeed. There were two different levels, easy and difficult, based on how far standing line is set up from the bucket. For the easy level, participants stand right next to the bucket, so they are unlikely to miss any shot unless they fail it on purpose. For the difficult level, the distance between participants and the bucket is far enough, about 15-20ft depending on the bucket size, to be more likely to miss a shot. Before the competition begin, the instructor states that he will let the student win for both hard and easy tosses, but make the student lose the game indeed.

A student volunteer makes an attempt for hard tosses first. Once the student make a success shot into the bucket, the student's turn is over. For easy tosses, the student throws a paper ball into the bucket five times. After the student finished attempts for both levels, the instructor modifies a game strategy based on how well the student performs such that the instructor wins the game overall even though the score from both hard and easy tosses are lower than the student.

All participants in class were randomly assigned into two groups, Group1 and Group 2, in crossover design setting. Two versions of quiz, Quiz A and Quiz B based on Simonoff (2003) and Moore (2010) were devised. Throughout the traditional lecture, the idea of the contingency table and Simpson's paradox with the example of sex bias in graduate admission at Berkeley (Bickel et al., 1975) and the example of helicopter evacuation (Moore, 2010) is delivered to Students in Group 1. After the lecture is completed, they take Quiz A. Next, Simpson's paradox is explained by engaging in paper toss game and class discussion is made after that. And then, they take Quiz

B. On the other hands, students in Group 2 take Quiz B after the lecture and take Quiz A after the paper toss activity.

Explanatory variables are shown in Table 1. Note that carryover effect for Quiz A an Quiz B are completely determined by treatment effect and period effect, so we cannot include all four effects in the model. Otherwise, the model matrix is rank deficient. We include treatment effect and period effect. The explanatory variables are shown in Table 2.

Table2: Explanatory variables

Group	Treatment Effect (Quiz Version)	Period Effect	Carryover Effect (Quiz A)	Carryover Effect (Quiz B)
1	0	0	0	0
	1	1	1	0
2	0	1	0	1
	1	0	0	0

In order to evaluate the activity's effectiveness, the scores from the two quizzes were compare in two ways, Generalized Estimating Equation (GEE) and Generalized Linear Mixed Model (GLMM), by using R programming. The R function of 'geeglm' in packages geepack for GEE and the function of 'glmer' in packages lme4 for GLMM are used to analyze the data. We focus on the two treatments and two periods design with binary response and assigned a correct score as 1 and incorrect score as 0 to test the period effect to evaluate the efficacy of the add-on teaching method.

Results

Each student j has two responses $Y_j = (y_{1j}, y_{2j})'$ and a model matrix X_j with two covariates

$$X_j = \begin{pmatrix} 1 & x_{11j} & x_{21j} \\ 1 & x_{12j} & x_{22j} \end{pmatrix}$$

where x_{11j} is the treatment effect, π, and x_{21j} is the period effect, τ. We are interested in testing for the effectiveness on the additional teaching technique, which is the hypothesis, $\tau=0$.

To test H_0: $\tau=0$ with the GEE, the data was interpreted by geeglm function with logit as link-function for a binomial model in R programming. The response variable is Score and the fixed effect was a random intercept with Question and Period. We confirmed the effects of Question and Period as $\beta_{Question} = -1.453$, SE = 0.358, Wald = 16.52, Pr = 4.8e-05 and $\beta_{Period} = 0.563$, SE = 0.451, Wald = 1.72 Pr = 0.19 respectively. The intercept of the model was somewhat significant as $\beta_{Intercept}$ =1.453, SE = 0.358, Wald =16.52, Pr = 4.8e-05. We found Wald statistics for Question is 17.19 with p-value of 3.4e-05 and its Chi square is 16.72 with p-value of 5.5e-05 and degree of freedom of 1, which was taken by ANOVA analysis and Wald statistics for Period is 1.72 with p-value of 0.19 and degree of freedom of 1 which was the same as Chi score earned by ANOVA function. Also, the estimated scale parameters for intercept was indicated as 1.01 with the standard error of 0.247. And, the estimated correlation parameters for alpha value was found as –0.127 with the standard error of 0.126.

To test H_0: $\tau=0$ with the GLMM, we used a binomial linear mixed effect model using the glmer with logit as link-function. We estimated fixed-effects parameters and random effects in a linear predictor using maximum likelihood. The random effect was Score whereas the fixed effects were a random intercept with Question and Period which were indicated in binomial response. We confirmed that the random effects for Question and Period showed zero variance and zero standard deviation which is a measure of validity for random effect. The fixed effects of Question and Period with assuming an interpret that is different for each ID was uncovered as $\beta_{Question} = -1.865$, SE = 0.423, z = –4.41, Pr = 1.1e-05 and $\beta_{Period} = 0.553$, SE = 0.414, z = 1.37 Pr = 0.17 respectively.

The intercept of the fixed effect was not significant as we found $\beta_{Intercept} = -0.0756$, SE = 0.0779, z = –0.971, Pr = 0.33. Also, the correlation of fixed effect was found as –0.631 for the intercept of Question, –0.375 for the intercept of Period and –0.247 for the intercept of Question and Period. With the ANOVA test, F-score was found as 17.62 which was the same as the value for sum of squares and mean squares with degree of freedom of 1 for Question. For Period, F-score was found as 1.87 which was also the value of sum of square and mean squares with degree of freedom of 1.

The result gained with the same fixed effect but with a random slop of period within group with correlated intercept showed the random effect for Question of 0.408 variance with 0.639 standard deviation and for Period of 2.323 variance with 1.524 standard deviation. The fixed effects of Question and Period as we revealed as $\beta_{Question} = -2.075$, SE = 0.561, z = –3.70, Pr = 2.1e-04 and $\beta_{Period} = 0.658$, SE = 0.514, z = 1.28 Pr = 0.200 respectively. The intercept of the model was somewhat significant as we found $\beta_{Intercept} = 1.603$, SE = 0.442, z = 3.63, Pr = 2.8e-04. Moreover, the correlation of fixed effect was shown as –0.717 for the intercept of Question, –0.229 for the intercept of Period and –0.308 for the intercept of Question and Period. With the ANOVA test, F-score was found as 17.01 which was the same as the value for sum of squares and mean squares with degree of freedom of 1 for Question. For Period, F-score was found as 2.08 which was also the value of sum of square and mean squares with degree of freedom of 1.

Discussion

Researches using longitudinal analysis have become very popular in observational studies and clinical trials, especially for the subject involving variations over time change like a cause-effect experiment. The two most well-known models, the GEE and the GLMM are often used in a cross-sectional data analysis in a longitudinal study design. According to Zhang et al. (2011), the

GLMM explicitly models the within-subject correlation by using random effects whereas the GEE implicitly accounts for such correlations by using sandwich-type variance estimates. In R programming, the function geeglm from geepack is used for fitting GEE models while the function lmer from lme4 is used for fitting generalized linear and nonlinear mixed effects models.

In the data analysis with the GEE, we chose an exchangeable model for the correlation structure within groups in the model in which each observation pair in a group shows the equivalent correlation. Because these are repeated measures data, an exchangeable structure could be considered as a good choice. To fit a GEE model with an exchangeable correlation structure, the logit link function was used and the main effects for the two predictors, Question and Period, were included with binomial variables, 0 and 1 where the response variable is Score. The geepack package also provides the ANOVA function which runs a Wald test. The Wald statistics, also called the Wald Chi-Squared test, is a way to reveal whether explanatory variables in a model are significant. If a variable is shown as significant, it can be interpreted that the variable influences an effect on the model. The variables demonstrated by no effect can be eliminated resulting in no meaningful change in the model. This test can be often utilized for models with binary variables or continuous variables. The null hypothesis for the Wald test is to confirm the consequential influence hold by a parameter in the model. If the null hypothesis is rejected, it proves that the variable can be removed with no harm. That is, the Wald test show that parameters are zero, you can remove the variables from the model without any concern. However, If the test shows the non-zero parameters, those variables in the model must be reported. In our research, we found that the Wald statistics = 1.72 with p-value=0.19. Hence, the period effect of the in-class activity does not appear to have a significant impact on evaluation of new teaching method.

Unlike the marginal model obtained by using the GEE, the GLMM is an alternative approach to explicate a repeated measure to fit a mixed effects model with random intercepts. The outcome gained by the GLMM is a conditional model and it is useful to interpret a subject-specific research data rather than a population-averaged data frame of a model. Since logistic regression takes account of a nonlinear link function, the degrees of an effect estimated in a marginal model is different from the effect predicted by conditional models. The parameter estimated by a conditional model is more likely to show a greater magnitude compared to those measured by a marginal model. In R, the lme4 package provides the glmer function to fit the GLMM and was used to analyze out binary data with maximum likelihood. Two analysis for the estimated the period effect of the intercept did not deviate from zero, so we can conclude that the period effect does not significantly affect the learning improvement with new teaching method.

For the analysis of distribution-based methods, we need to use a data set with continuous random variable that follows the normal distribution. The data under the same experiment setting was collected and analyzed with 2x2 crossover design by T-test using R programming. The data collected are shown in Table 3 below.

Table3: The data collected with continuous response

ID	Gender	PreQ Score	PostQ Score	Pre Q	Post Q
1	F	3	6	Airline	Hospital
2	F	4	4	Hospital	Airline
3	F	2	6	Airline	Hospital
4	M	4	5	Hospital	Airline
5	F	4	4	Airline	Hospital
6	M	2	6	Hospital	Airline
7	F	4	4	Hospital	Airline

As a result, the period effect in this 2x2 crossover study did not significantly affect at the

0.05 significance level with 2 degree of freedom because the actual significance level was 0.076. Therefore, we concluded that the additional game activity to boost learning outcomes was not significantly effective based on all three data analysis.

Conclusion

Finding an ideal evaluating interventions for testing an intervention's effectiveness will be never easy to be accomplished in a real work setting. So, setting the most suitable experimental design will always remain as a prodigious challenge for researchers. Researchers cannot assure that there exists such an error-free experimental design for a study, but they could make a better or worse choice of metrologies to find the effectiveness of treatments. According to Cox and Snell (1999), crossover method is defined as design in which physically the same individual is used as an experimental unit on more than one occasion. In other words, each treatment is given to each experimental unit in a prearranged sequence mostly for evaluating the treatment effect. More interestingly, we tried to evaluate the effectiveness of new teaching method in a crossover design with measuring the period effect, which has never performed previously. Even though the result showed that there was no significant impact on the period treatment, the continuous researches focused on period effects will lead researchers to enter new phase of the finding.

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed., Wiley series in probability and mathematical statistics). Hoboken, NJ: Wiley-Interscience.

Ackroyd, S., & Hughes, J. A. (1981). *Data collection in context* (Aspects of modern sociology. Social research). London ; New York: Longman.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Berliner, D. (2002). Comment: Educational Research: The Hardest Science of All. *Educational Researcher, 31*(8), 18-20. Retrieved from http://www.jstor.org.proxy.wexler.hunter.cuny.edu/stable/3594389.

Bleske, B., Remington, T., Wells, T., Klein, K., Guthrie, S., Tingen, J., . . . Dorsch, M. (2016). A randomized crossover comparison of team-based learning and lecture format on learning outcomes. *American Journal of Pharmaceutical Education, 80*(7), American Journal of Pharmaceutical Education, 2016, Vol.80(7).

Blyth, C. (1972). On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association, 67*(338), 364-366. doi:10.2307/2284382

Chinchilli, V., Phillips, B., Mauger, D., & Szefler, S. (2005). A General Class of Correlation Coefficients for the 2 × 2 Crossover Design. *Biometrical Journal, 47*(5), 644-653.

Connelly, L. M. (2014). Understanding Crossover Design. *MEDSURG Nursing*, *23*(4), 267-268.

Cox, D., & Snell, E. (1999). *Analysis of binary data* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Fieberg, J., Rieger, R., Zicus, M., & Schildcrout, J. (2009). Regression modelling of correlated

data in ecology: Subject-specific and population averaged response patterns. *Journal of Applied Ecology, 46*(5), 1018-1025.

Gellman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Helge, J., Tobin, W., Drachmann, L., Hellgren, T., Dela, I., & Galbo, F. (2012). Muscle ceramide content is similar after 3 weeks' consumption of fat or carbohydrate diet in a crossover design in patients with type 2 diabetes. *European Journal of Applied Physiology*, 112(3), 911-918.

Huang, Y., & Ke, B. (2014). Influence analysis on crossover design experiment in bioequivalence studies. *Pharmaceutical Statistics, 13*(2), 110-118.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications.* New York: Springer Science Business Media, LLC.

Lin, L., Huang, Y., Watson, R., Wu, S., & Lee, Y. (2011). Using a Montessori method to increase eating ability for institutionalised residents with dementia: A crossover design. *Journal of Clinical Nursing,* 20(21-22), 3092-3101.

Lui, K., & Chang, K. (2012). Hypothesis testing and estimation in ordinal data under a simple crossover design. *Journal Of Biopharmaceutical Statistics*, *22*(6), 1137-1147. doi:10.1080/10543406.2011.574326.

McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 109–142.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman & Hall, New York.

McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*, Wiley, New York.

Milliken, G. A., & Johnson, D. E. (1984). *Analysis of messy data* (Vol. 1). New York: Van Nostrand Reinhold Company.

Mills, E.J., Chan, A., Wu, P., Vail, A., Guyatt, G.H., & Altman, D.G. (2009). Design, analysis and presentation of crossover trails. *Trials, 10*(27). Retrieved from http://www.trialsjournal.com/content/ 10/1/27

Moore, D. S. (2010), *The Basic Practice of Statistics*, New York: W. H. Freeman and Company, 5th ed.

Nedungadi, P., & Raman, R. (2010). Effectiveness of adaptive learning with interactive animations and simulations. *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on, 6*, V6-40-V6-44.

Piantadosi, Steven. (2005) Crossover Designs. In: Piantadosi Steven. *Clinical Trials: A Methodologic Perspective*. 2nd ed. Hobaken, NJ: John Wiley and Sons, Inc.

Popper, K. (1968). *The logic of scientific discovery* (2nd Harper torchbook ed., Harper torchbooks ; TB 576). New York: Harper & Row.

Prunuske, Amy, Henn, Lisa, Brearley, Ann, & Prunuske, Jacob. (2016). A Randomized Crossover Design to Assess Learning Impact and Student Preference for Active and Passive Online Learning Modules. *Medical Science Educator, 26*(1), 135-141.

Reich, N., & Milstone, A. (2013). Improving efficiency in cluster-randomized study design and implementation: Taking advantage of a crossover. *Open Access Journal of Clinical Trials*, 11.

Simon, & Chinchilli. (2007). A matched crossover design for clinical trials. *Contemporary Clinical Trials,* 28(5), 638-646.

Simonoff, J. S. (2003), *Analyzing Categorical Data*, Springer-Verlag.

Simpson, E. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 238-241. Retrieved from http://www.jstor.org/stable/2984065

Taib, M., Shariff, Z., Wesnes, K., Saad, H., & Sariman, S. (2012). The effect of high lactose-isomaltulose on cognitive performance of young children. A double blind cross-over design study. *Appetite, 58*(1), 81-7.

Ulrich Halekoh, Søren Højsgaard, & Jun Yan. (2005). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software, 15*(2), 1-11.

Valois, M. (1999). Evaluation of the performance of the generalized estimating equations method for the analysis of crossover designs. Ottawa: National Library of Canada = Bibliothèque nationale du Canada.

Vegas, S., Apa, C., & Juristo, N. (2016). Crossover Designs in Software Engineering Experiments: Benefits and Perils. *Software Engineering, IEEE Transactions on*, 42(2), 120-135.

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and nonlinear models for the analysis of repeated measurements.* New York: M. Dekker.

Welleck, S., & Blettner, M. (2012). On the proper use of crossover design in clinical trials. *Deutsches Ärztebatt International, 109*(15), 278- 281.

Wilson, A., Olds, T., Lushington, K., Petkov, J., & Dollman, J. (2016). The impact of 10-minute

activity breaks outside the classroom on male students' on-task behaviour and sustained attention: A randomised crossover design. *Acta Paediatrica, 105*(4), E181-E188.

Walt, J., & Potgieter, F. (2012). Research method in education: The frame by which the picture hangs. *International Journal of Multiple Research Approaches, 6*(3), 220-232.

Zhang, H., Xia, Y., Chen, R., Gunzler, D., Tang, W., & Tu, X. (2011). Modeling longitudinal binomial responses: Implications from two dueling paradigms. *Journal of Applied Statistics, 38*(11), 2373-2390.

Zhang, H., Yu, Q., Feng, C., Gunzler, D., Wu, P., & Tu, X. M. (2012). A new look at the difference between the GEE and the GLMM when modeling longitudinal count responses. *Journal Of Applied Statistics*, *39*(9), 2067-2079. doi:10.1080/02664763.2012.700452.