Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

_____ _____

Ruoming Wu Date

Comparative Analysis of Four Intraclass Correlation Coefficient Estimation Methods on Continuous Outcomes in a Survey Study

By

Ruoming Wu

Master of Science in Public Health
Biostatistics

_____

Paul Weiss, MS.
Thesis Advisor

_____

Howard H. Chang, Ph.D.
Thesis Reader

_____

Date

Comparative Analysis of Four Intraclass Correlation Coefficient Estimation Methods on Continuous Outcomes in a Survey Study

By

Ruoming Wu
B.S.
Emory College of Arts and Sciences, Emory University
2015

Paul Weiss, MS.
Advisor

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Science in Public Health
Biostatistics
2016

Abstract

Comparative Analysis of Four Intraclass Correlation Coefficient Estimation Methods on Continuous Outcomes in a Survey Study

By Ruoming Wu

This study aimed to compare methods of estimating the intraclass correlation coefficient (ICC) on continuous outcomes in survey research utilizing cluster sampling. Analysis was performed on the dataset obtained from the Power Up for 30 program, which measured all elementary schools from different school districts in Georgia. Grade level teachers completed surveys regarding students' physical activities. ICCs were calculated for responses from each survey question at the district level and the school level using four methods: the ANOVA approach, the random intercept approach, the GEE approach and the design effect approach. The result indicated that different ICC estimation methods led to various results. ICC estimates obtained using the design effect approach greatly diverged from ICC estimates obtained using the other approaches. For all survey questions, ICC estimates at the school level were greater than that at the district level, suggesting that responses toward all questions exhibited greater consistency among different grade level teachers at the school level, compared to at the district level. Future studies should consider ICC estimation methods that take into account the non-normality of the dataset. Also, average number of respondents should be consistent at the district and the school level.

Comparative Analysis of Four Intraclass Correlation Coefficient Estimation Methods on Continuous Outcomes in a Survey Study

By

Ruoming Wu
B.S.
Emory College of Arts and Sciences, Emory University
2015

Paul Weiss, MS.
Advisor

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements of the degree of
Master of Science in Public Health
Biostatistics
2016

Acknowledgements

I am extremely grateful for my thesis advisor, Paul Weiss, whose expertise, generous guidance and support made it possible for me to work on a topic that was of great interest to me.

I am also thankful for Dr. Howard Chang for taking time to read my thesis.

Contents

1 Introduction

Survey research has been massively utilized in public health studies. The primary objective of adopting the survey research approach is to select a sample that would best represent the nature of the population [1]. Among all the survey sampling methods, cluster sampling has been most frequently used on account of its cost-effectiveness [2]. Such a sampling method is composed of multiple groups or clusters of smaller or individual units [3]. Geographical regions and school districts are both examples of clusters. An important assumption in cluster sampling states that observations from the same cluster are likely to be correlated to each other, rather than independent [4]. Such within cluster correlation is quantified by the intraclass correlation coefficient (ICC), denoted by ρ. Specifically, the ICC measures the relative homogeneity of the observations within the clusters, compared to the total variation among all the observations from the dataset [3]. It is algebraically defined as the proportion of the total variance that exists between clusters:

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_w^2} \tag{1}$$

where σ_B^2 and σ_w^2 represent the between- and within-cluster variance, respectively [5]. The ICC also serves as an indicator of whether the assumption of independent error terms often used for statistical tests is violated [6]. Previous studies have demonstrated that ignoring ICCs in statistical analyses results in a severe inflation of the Type I error [7]. Moreover, studies have also shown that cluster sampling generally exhibits lower statistical power than the simple random sampling method. As a result, the sample size for cluster sampling studies has to be increased, in order to achieve sufficient statistical power, compared to that of simple random sampling [8]. The magnitude of such an increase in sample size is usually characterized by the design effect, which is defined as the ratio of the variance of an estimator under cluster sampling

to that of an estimator under simple random sampling [4]. The design effect (*Deff*) is algebraically given by

$$Deff = 1 + (n_A - 1)\times\rho \quad (2)$$

where n_A is the average cluster size [9]. Since the effective sample size and statistical power can be significantly affected by the value of ICC, obtaining an accurate estimation of ICC is crucial to a study.

Cluster sampling design has received a significant amount of attention and has been massively implemented in public health studies. Accordingly, calculations of the ICC are frequently applied in these studies, as well as the illustration of the design effect. For example, the cholesterol education and research trial (CEART) study randomly assigned primary care physician practices to patients. The ICC, along with the design effect, for primary care practices were estimated based on patient health outcome such as diastolic blood pressure, body mass index and triglycerides using the analysis of variance (ANOVA) approach [8]. Another example is the study of adolescent cigarette use measurements. School-level ICCs for smoking-related variables were estimated through the mixed model regression approach [6]. On the other hand, the study of acquiring an accurate estimate of the ICC has received much less attention. Factors affecting the value of ICC include scope of the cluster and subject variability. Studies have suggested that the ICC within a family or a household is expected to be higher than that within a geographical area because family members are considered to be more correlated than observations taken from a large geographical area [10]. In addition, studies have also indicated that the ICC is sensitive to subject variability, the degree to which variation of observations in the sample vary from the

"true" variation in the target population. It appears that when measurement errors are fixed, ICC values can widely fluctuate due to subject variability [11].

The main purpose of this study is to compare methods of estimating the ICC for a clustered continuous outcome. Similar studies in the literature include comparison of methods for estimating the ICC for binary responses, where five methods were performed and compared on cluster randomized trials of cancer screening datasets [10]. Another study on the association between average work hours, a continuous outcome, and nurse burnout utilized and compared two distinct multilevel regression model methods to estimate the ICC [12]. In this study, responses regarding the frequencies of students' physical activities from various grade teachers are analyzed at the district and the school level. Specifically, four methods are applied to estimate the ICC, namely, the ANOVA estimator, the random intercept model approach, the generalized estimating equation approach and the design effect approach. This study aims to compare the performance of these methods of estimating the ICC for the district and school level responses and generalize the findings to the design of future survey research using cluster sampling.

2 Methods

2.1 Dataset

The dataset is collected from the Power Up for 30 program, conducted by the Georgia Department of Public Health. This program aims to improve the physical well-being of elementary school students by encouraging every elementary school in Georgia to incorporate 30 minutes of physical activity each weekday. The survey-based study measured all elementary schools from different school districts. Within each school, teachers from each grade level (Grade 1-5) answered the students' physical activities related questionnaire. Their responses to several questions are modeled on their positions at the school. This study restricts the number of respondents' in each cluster to be greater than 20, since the ICC may not be well estimated for small clusters.

2.2 Statistical methods

Four methods of estimating the ICC are applied to the datasets: the ANOVA estimator, the random intercept model approach, the generalized estimating equation approach and the design effect approach. In general, suppose there are a total of k clusters. Within each cluster, denoted by i, there are n_i subjects. The outcome variable, denoted by y_{ij}, represents the response of the j^{th} subject in the i^{th} cluster.

2.2.1 The ANOVA estimator

The one-way random effects model is frequently applied to estimate the ICC. The model is given by:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{3}$$

where μ represents the overall average response of the target population, α_i represents the specific effect of the i^{th} cluster and ε_{ij} represents the error term, for i from *1* to k and j from *1* to n_i. The model assumes that the α_i's are independently and identically distributed and follow a normal distribution with mean 0 and variance σ_α^2 and the ε_{ij}'s are also independently and identically distributed but follow a normal distribution with mean *0* and variance σ_w^2. The α_i's and the ε_{ij}'s are independent. Under such a model, the ANOVA estimator [13] for the ICC is given by:

$$\widehat{\rho_A} = \frac{MSB - MSW}{MSB + (n_A - 1)MSW} \tag{4}$$

where n_A represents the average cluster size, and is given by

$$n_A = \frac{1}{k-1}\left[N - \frac{\sum_{i=1}^{k} n_i^2}{N}\right], and\ N = \sum_{i=1}^{k} n_i . \tag{5}$$

The MSB and MSW are between- and within-cluster mean squares, which can be obtained from the one-way analysis of variance.

2.2.2 The random intercept model approach

For the random intercept model, a type of regression model, the baseline value for each cluster is different, but the slope is the same for all cluster. The model is given by:

$$y_{ij} = \mu + \theta_i + x\beta + \varepsilon_{ij} \tag{6}$$

where μ represents the overall average intercept, θ_i is the difference between the mean of cluster i and the average intercept, β is the vector of coefficients that do not vary across clusters and ε_{ij} represents the error term. It is assumed that θ_i follows a normal distribution with mean *0* and variance τ^2 and ε_{ij} follows a normal distribution with mean *0* and variance σ^2. In particular, τ^2

measures the between-cluster variability in the intercepts and σ^2 measures the within-cluster variability in the residuals. In this case, the ICC is computed as:

$$\hat{\rho} = \frac{\tau^2}{\tau^2 + \sigma^2}. \tag{7}$$

2.2.3 The generalized estimating equations approach

Another regression model implemented in this study is the generalized estimating equations (GEE). As a very general statistical estimating approach, GEE allows user-specified working correlations. A key feature of GEE is that it is robust to misspecification of the working correlation. In addition, robust standard errors, another nature of GEE, ensure the validity of regression parameters even if the working correlation is misspecified [14]. For the continuous outcome in this study, a Gaussian distribution is selected with an identity link. The GEE model is given by:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_i \tag{8}$$

where the error term ε_i follows a normal distribution with mean 0 and variance matrix V_i. Such a variance matrix is given by:

$$V_i = \sigma^2 R_i(\alpha) \tag{9}$$

where $R_i(\alpha)$ is an $n_i \times n_i$ correlation matrix and α is the parameter that indicates the functional form of the correlation. In cluster sampling, it is normally assumed that observations from various clusters are independent but that from the same cluster are usually dependent. Thus, an exchangeable correlation structure, which is given by

$$\begin{bmatrix} 1 & \alpha & \cdots & \alpha & \alpha \\ \alpha & 1 & & \alpha & \alpha \\ \vdots & & \ddots & \vdots & \\ \alpha & \alpha & & 1 & \alpha \\ \alpha & \alpha & \cdots & \alpha & 1 \end{bmatrix} \tag{10}$$

is assumed for this study. The working correlation structure coefficient represents the ICC since it accounts for the within-cluster correlation.

2.2.4 The design effect approach

As mentioned earlier, the design effect is utilized in cluster sampling to measure the efficiency of the sample. The design effect is algebraically given in (2) where n_A is the average cluster size. The formula for n_A is given in the ANOVA estimator section. Consequently, the ICC can be derived from (2) as follows:

$$\hat{\rho} = \frac{Deff - 1}{n_A - 1}. \qquad (11)$$

2.3 Computation

All analyses were performed in SAS software (version 9.4; SAS Institute Inc, Cary, NC). For the ANOVA estimator, SAS functions were written to compute the ICC based on equation (5). For the random intercept approach, we used SAS PROC MIXED to model the random effect. Estimates of the variance attributable to district- or school- level and to residual errors from PROC MIXED output were utilized to compute the ICC. For the generalized estimating equations approach, we used SAS PROC GENMOD. The design effects at district and school-level were calculated in SAS callable SUDAAN (version 11; RTI International, Research Triangle Park, NC). ICCs were derived from the design effects according to equation (11).

3 Results

Cluster characteristics and the results of ICCs using the four aforementioned estimation methods for questions with continuous outcome at the district- and the school-level are shown in Tables 1-7. The average number of respondents, which represents the average cluster size, for all questions estimated at the district- and the school-level were approximately 60.20 and 4.75, respectively.

Table 1
Cluster characteristics and results of ICC from different estimation methods for the survey question: "How many days per week is recess scheduled at your school?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.463	31.84	0.286	0.312	0.212	0.519
School Level	4.783	3.080	0.504	0.518	0.499	0.550

Table 2
Cluster characteristics and results of ICC from different estimation methods for the survey question: "On average, how many minutes is a scheduled recess period at your school?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.318	31.860	0.220	0.229	0.210	0.520
School Level	4.772	2.610	0.376	0.393	0.397	0.427

Table 3
Cluster characteristics and results of ICC from different estimation methods for the survey question: "In general, how often is physical activity integrated into the classroom?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.181	6.510	0.0328	0.0354	0.0600	0.0870
School Level	4.761	1.360	0.0769	0.115	0.118	0.0625

Table 4
Cluster characteristics and results of ICC from different estimation methods for the survey question: "On average, how many minutes per day do your teachers integrate physical activity into the classroom?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.146	4.490	0.0339	0.0439	0.0440	0.0590
School Level	4.759	1.230	0.0646	0.0742	0.0708	0.0612

Table 5
Cluster characteristics and results of ICC from different estimation methods for the survey question: "How long is each bout of the integrated physical activity?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.073	10.930	0.0694	0.0715	0.0964	0.168
School Level	4.753	2.300	0.308	0.308	0.309	0.346

Table 6
Cluster characteristics and results of ICC from different estimation methods for the survey question: "Approximately how many school staff members participate in these physical activity opportunities offered by your school?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	60.073	10.930	0.0694	0.0715	0.0964	0.168
School Level	4.753	2.300	0.308	0.308	0.309	0.346

Table 7
Cluster characteristics and results of ICC from different estimation methods for the survey question: "At your school, how many parents are involved in promoting physical education/physical activity before, during, and after school?"

	Cluster Characteristics		ICC Estimation Methods			
	Average number of respondents	Design Effect	ANOVA	Random Intercept	GEE	Derivation from Design Effect
District Level	50.876	3.270	0.0307	0.0366	0.0175	0.0386
School Level	4.738	1.810	0.173	0.177	0.191	0.217

3.1 District Level

At the district level, the ICC values from the ANOVA estimator were similar from those obtained using the random intercept model approach for all questions. The ICC estimates from the design effect were noticeably greater than those obtained from other methods. In particular, the greater differences between ICCs are associated with the greater the design. For example, the question shown in Table 1 exhibited a design effect of 31.84, and the ICC from the design effect approach was 67% higher than the mean of ICCs obtained using other methods. Additionally, the question shown in Table 7 indicated a small design effect with the value 3.27. We consequently observed a small difference between the ICC from the design effect approach and those obtained using the ANOVA and random intercept methods. The ICCs from the GEE approach fluctuated and were different from that obtained using the other three methods, except for those shown in Table 4.

ANOVA values of ICC estimates were 0.29 and 0.22 for questions in Table 1 and 2, respectively, and these two values were significantly higher than those from the other questions. This observation indicated that greater consistency were found in the responses from different grade level teachers for these two questions. On the other hand, for questions shown in Table 3-7, the ICCs ranged from 0.03 to 0.07. This observation suggested that greater discrepancy existed in the responses from different grade level teachers for these five questions.

3.2 School Level

At the school level, ICCs obtained from the ANOVA estimator, random intercept model approach and the GEE approach were close to each other for all questions. However, we observe

differences in ICCs using the design effect approach compared to the other approaches. The design effects at the school level, ranging from 1.23 to 3.08, were smaller than those at the district level, ranging from 3.27 to 31.86. For questions with design effects greater than 1.50, ICCs obtained from the design effect approach were greater than those obtained from the other methods (Table 1, 2, 6 and 7). For questions with design effects lower than 1.50, ICCs obtained from the design effect approach were smaller than those obtained from the other methods (Table 3-5). The difference between ICCs from the design effect approach and the mean of ICCs from other approaches at the school level was lower than the respective differences at the district level. Specifically, estimated ICCs from the design effect approach differed from the mean of the ICCs obtained from the other methods by 15%.

Values of ICC estimates were 0.50, 0.39, 0.31 and 0.18 using the random intercept approach for questions shown in Table 1, 2, 6 and 7, respectively. These high values suggested that greater consistency were detected in the responses from different grade level teachers for the aforementioned questions. For questions shown in Table 3-5, the ICC estimates ranged from 0.06 to 0.10. This observation suggested that greater discrepancy in the responses from different grade level teachers were detected for these three questions.

4 Discussion

As would be expected, different intraclass correlation coefficient estimation methods lead to various results. The percent difference in estimated ICCs for the same research question could range from 15% to 66%. In particular, ICC estimates obtained from the design effect approach generally differed greatly from that obtained from other methods. One possible explanation would be that the ANOVA estimator, the random intercept model approach and the GEE approach all assumed normality for the dataset, whereas the design effect approach did not have this underlying assumption. Thus, it is sensible to propose that for questions with large separation in ICC estimates obtained from the design effect approach, as compared to that from the other methods, such as the questions shown in Table 1 and 2, the underlying datasets may not follow normal distributions. Consequently, when performing analysis on ICCs, no single method can be fully relied on. To be specific, as shown in Table 1, the ICC estimate from the GEE approach suggested that at the district level, certain degrees of consistency on the responses from different grade level teachers were observed, whereas the ICC estimate from the design effect approach suggested that at the district level, a significant amount of degrees of consistency on the responses from different grade level teachers were observed. As a consequence, uncertainty in ICC estimations may lead to unreliable conclusions.

In this study, at the district cluster level, large ICC estimates for questions regarding general physical activity frequencies were observed (Table 1 and 2). This result indicated that responses toward broad questions were consistent among different grade level teachers. On the other hand, ICC estimates were much lower for detailed questions regarding physical activity, suggesting that inconsistency was observed for response toward specific questions among different grade

level teachers. At the school level, ICC estimates were generally higher than those at the district level, for all questions. This observation implied that responses toward all questions exhibited greater consistency among different grade level teachers at the school level, compared to at the district level.

Limitations of this study include that the underlying outcome of interest was greatly skewed to the right. Nevertheless, the ANOVA, the random intercept and the GEE methods all assumed normality of the underlying dataset. Thus, the ICC estimates were not accurate using these three methods. Future studies may consider statistical methods that targets at non-normal datasets. For example, the GEE approach with dispersion parameter incorporated into the working correlation structure may be utilized as another ICC estimation method. Another limitation is that the average number of respondents at the district level was much greater than that at the school level. Thus, it would be expected that more consistency would be observed among subjects at the school level. This assumption raised a concern such that the value of ICC estimates at the school level would be lower if the average number of respondents was the same as that at the district level. In future studies, sufficient number of subjects within each cluster at the school level has to be achieved, in order to obtain accurate ICC estimates. In this study, we did not obtain confidence intervals for each ICC point estimate. Future studies should consider to incorporate bootstrap confidence intervals into statistical analysis to elucidate the existence of, or lack thereof, overlaps between ICC confidence intervals from different methods. Future studies may also consider examine the ICC for similar research studies in other states, in order to examine the consistency of grade level teachers' responses across the country.

Reference

[1] Sedransk J. Analytical surveys with cluster sampling. Journal of the Royal Statistical Society 1965;27(2):264-78.

[2] Nelms CO, Otis DL, Linz GM, Bleier WJ. Cluster sampling to estimate breeding blackbird populations in North Dakota. Wildlife Society Bulletin 1999;27(4):931-37.

[3] Cochran W. Sampling techniques. New York: Wiley; 1977.

[4] Kish L. Survey Sampling. New York: Wiley; 1965.

[5] Ukoumunne OC, Gulliford MC, Chinn S. A note on the use of the variance inflation factor for determining sample size in cluster randomized trials. Journal of the Royal Statistical Society 2002;51(4):479-84.

[6] Murray DM, Alfano CM, Zbikowski SM, Padgett LS, Robinson LA, Klesges R. Intraclass correlation among measures related to cigarette use by adolescents estimates from an urban and largely African American cohort. Addictive Behaviors 2002;27:509-27.

[7] Murray DM, Hannan PJ, Baker WL. A monte carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cronfield's penalties? Evaluation Review 1996;20(3):313-37.

[8] Parker DR, Evangelou E, Eaton CB. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). Contemporary Clinical Trials 2005;26:260-67.

[9] Gambino JG. Design effect caveats. The American Statistician 2009;63(2):141-46

[10] Wu S, Crespi CM, W WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemporary Clinical Trials 2012;33:869-80.

[11] Lee KM, Lee J, Chung CY, Ahn S, Sung KH, Kim TW, et al. Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. Clinics in Orthopedic Surgery 2012;4:149-55.

[12] Park S, Lake ET. Multilevel modeling of a clustered continuous outcome: nurses' work hours and burnout. Nurs Res 2005;54(6):406-13.

[13] Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60:807-11.

[14] Homish GG, Edwards EP, Eiden RD, Leonard KE. Analyzing family data: A GEE approach for substance use researchers. Addict behavior 2010;35(6):558-63