THEORY AND APPLICATIONS OF INTRACLASS CORRELATION COEFFICIENTS AT CLUSTER RANDOMIZED DESIGN FOR STATISTICAL PLANNING VIA HIERARCHICAL MIXED MODELS

By

Chun-Lung Lee

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods — Doctor of Philosophy

2019

ABSTRACT

THEORY AND APPLICATIONS OF INTRACLASS CORRELATION COEFFICIENTS AT CLUSTER RANDOMIZED DESIGN FOR STATISTICAL PLANNING VIA HIERARCHICAL MIXED MODELS

By

Chun-Lung Lee

Research investigators rely on information of intraclass correlation coefficients for planning and conducting designs and experiments for scientific inquiries in educational and social studies. Randomized controlled trials and cluster randomized studies are deemed as the gold standard for evidence-based interventions, and both approaches have been applied successfully in many situations for more effective decision-making in education and social research. The cluster randomized designs for community-based research, in particular, have been widely used in the modern era, since they are often operated at the group level, like a whole community or worksite, in order for researchers more easily to deal with random assignment of an entire intact group rather than that of each individual subject. Hence, such cluster-randomized trials or group-randomized experiments have become important and useful to provide evidence-guided practice models for scientific inquiry and research.

The aim of this dissertation is to develop the methods for the intraclass correlation coefficients for binary and continuous outcomes in cluster-based intervention designs using hierarchal mixed model based on the scenarios of unconditional and conditional multilevel structures with cluster sampling schemes. Simulation studies are used to assess the statistical properties of intraclass correlation estimation and inference via the real data set of RSA-911 for people with disabilities served in the Michigan Rehabilitation Services Programs.

The results show that the average (unadjusted) intraclass correlation is about 0.01 for

competitive employment and about 0.02 for weekly earnings (quality employment) in Michigan. These average (unadjusted) intraclass correlations from RSA-911 are relatively low in comparison to education interventions or academic programs for assessments in reading and mathematics across K-12 (Bloom et al., 1999, 2007; Hedges & Hedberg, 2007; Schochet, 2008); however, they seem comparable to some extent from those psychological and mental health data in school-based intervention designs (Murray & Short, 1995).

For future study, researchers may look into different types of integrated large-scale complex data sets such as RSA-911 data with a set of covariates from Census data for investigating how intraclass correlation performs in statistical estimation and inference across multiple platforms. In addition, it would be interesting to study how to deal with missing values in the estimation procedure of intraclass correlation, and what remedial procedure can be added to improve estimation process. For the proposed method, it would recommend the total sample size should be greater than 1,500 and within group sample size would be better to be larger than 100 (with the number of groups about 15).

In conclusion, this study provides a comprehensive methodology for intraclass correlation estimation and inference using the mixed "analysis of variance" approach along with the derived sampling distribution (i.e., *F*-distribution) for testing hypothesis as well as building confidence interval on intraclass correlation estimates. Such proposed statistical procedures can be easily used and applied in any large-scale or small-scale data sets, whereas small total sample size and small within group size and missing data are limitations on intraclass correlation estimation in terms of precision and accuracy.

Keywords: Intraclass Correlation Coefficient, Cluster Randomized Design, Multilevel Structure, Hierarchical Linear Modeling, Evidence-based Practice Models

This dissertation is dedicated to Mom and
Dad (, both of whom graciously and
patiently tolerated me then and now) ~

Through all the years,

Thank you for always always believing in me
(that this would someday be completed)!

ACKNOWLEDGEMENTS

I like to thank the support of my dissertation/academic advisor, Dr. Kimberly Kelly, and my committee members, Drs. Richard Houang, Gloria Lee, and Sukyeong Pi. This dissertation is the final product of my (long and winding) PhD journey at Michigan State, and it cannot be done without two (separate but equally important) groups' proper training – my MQM (measurement & quantitative methods) and PE (project excellence at rehab counseling). I am very fortunate to have not just one (major of MQM) but two (aficionado of rehab counseling as well) unique experiences (within which there're challenges, difficulties, happiness and joys to make me grow as who I am today) on the special educational trip to the goal line (a doctorate degree). Although I did not attend the graduation ceremony, I was truly inspired by Kirk Cousins who delivered a passionate commencement speech (MSU, Spring 2019; https://www.wkar.org/post/kirk-cousins-may-3-2019-michigan-state-university-commencement-address#stream/0), addressing that:

"Through it all, enjoy the journey ... let us rejoice and be glad in it ... don't just deliver, over-deliver ... see life through a window, not a mirror .... and choose to be a great decision maker."

At the end of the day, I can tell myself, "While chasing/fighting the three letters (phd), I didn't forget to stop and smell the roses along the way (to enjoy enough the tough road thru paradise)." And also "The Lord blessed my time here in ways I never thought possible (God was preparing us for great things, wasn't He?)" To sum up, "It's good to have an end to journey toward; but it's the journey that matters, in the end." Go Green, Go White, Go MQM, and Go PE!

PREFACE

The history of intraclass correlation can be traced back to the last century that Sir Ronald A. Fisher introduced it to research communities as a new tool for measuring the level of similarity within a group. Since then, the intraclass correlation has been used as one of the most important statistical tools in scientific inquiries. In education, for example, it is often to use the intraclass correlation coefficient (or ICC) to measure the degree of intra-cluster resemblance in student educational outcomes (e.g., test scores) between different classrooms or schools. Although the ICC was a great success in the idea of how to measure within-group "correlation," it was not until later that Allan Donner and his colleagues provided a comprehensive and practical framework of the ICC estimation and inference (e.g., point estimates are derived by multivariate normal theory, and hypothesis tests are based on variance components using analysis of variance, ANOVA). In the contemporary era, ICC plays another key role in quantifying the inherent clustering effect size (i.e., within-group variation) in multilevel designs by using hierarchical linear models (HLM). Stephen Raudenbush is a pioneer for the development and application of HLM in education, and he sheds light on how to evaluate the effect magnitude of multilevel structure by ICC. Moreover, Larry Hedges, renowned for his work of meta-analysis in education, finds a novel approach to "empowering (i.e., power analysis)" sampling designs through design effect (i.e., a function of ICC). Lastly, Tenko Raykov gives new insight into strategies for ICC estimates in the complex statistics setting (e.g., a categorical outcome variable) for HLM via latent variable models. The goal of this dissertation is to draw together in one place the major ICC developments, then to further develop a new thinking in statistical inquiry of ICC estimation and inference. In addition, the evidence-based paradigm in vocational rehab is another "painted picture" of this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The need for more scientific evidence-based research has been increasingly concerned in 21st century education (Schneider et al., 2007). The use of rigorous methods such as randomized control trial (RCT) and cluster randomized trial (CRT) experiments in particular, is important to not only reinforce sound research but also build a solid basis of evidence-guided knowledge for informing policymakers and practitioners (Menon et al., 2009; Slavin, 2002). Under The Every Student Succeeds Act of 2016 (amended after No Child Left Behind), the U.S. Department of Education (2016) wrote the new guidelines of implementation of scientific research. Specifically, as for use of evidence-based interventions, researchers need to be guided by auxiliary research evidence from previous studies in order to conduct scientifically rigorous research as well as promote better and effective outcomes in education, according to the statistical standards and guidelines for the National Center for Education Statistics at The What Works Clearinghouse (https://ies.ed.gov/ncee/wwc/). With that goal in mind, RCTs and CRTs are often highly suggested by federal education research agencies, such as Institute of Education Sciences and its affiliated centers, and constantly deemed as the gold standard in scientific research and evidence-informed practice, since both RCT and CRT approaches have already been proved successfully in many circumstances for making decisions in education.

One key element to making any meaningful scientific conclusions is to produce evidential base through designs and experiments (Anderson & Shattuck, 2012; Barab & Squire, 2004; Cobb et al., 2003; Odom et al., 2005; Shavelson et al., 2003). For education policy and practice in the 21st century (Slavin, 2008), the pursuit of research soundness has been already

reinforced persistently by means of education legislation, e.g., NCLB Legislation (2002) and ESRA Legislation (2008). The No Child Left Behind Act of 2001 (NCLB), for example, supported scientifically based research involving rigorous and systematic methods to obtain applicable and generalizable knowledge for improving school programs, teaching methods and learning outcomes. Furthermore, The Education Sciences Reform Act of 2002 (ESRA) was proposed to reform education sciences through principles of scientific research such as randomized experiments to measure causal impacts on educational outcomes.

In the era of evidence-based practice (EBP), rehabilitation counseling is also embracing the concepts of best practice and knowledge translation to incorporate scientific advances and changes that have redefined the relationship between impairments and the capability to work (Leahy et al., 2014a). As for the state-federal vocational rehabilitation (VR) services, the public VR agencies are a major force of employment assistance for individuals with disabilities. Recent legislation for The Workforce Innovation and Opportunity Act (WIOA) of 2014, state VR programs have to assist the target disability populations, with educational or vocational training services, to succeed in the labor market and further to compete, with professional competency skills, in the global economy (WIOA Legislation, 2018). Therefore, nowadays the rehabilitation counseling workforce (including all those counselors, educators, practitioners, and researchers) need to work together to embrace the new era of the EBP paradigm to help VR customers improve the accessibility of quality rehabilitation services with informed choices of effective interventions or treatments. Moreover, it is important to use data-driven or evidence-based rehabilitation counseling best practices to improve accountability and outcomes for people with disabilities by conducting systematic reviews and well-designed studies, as a way to get more reliable and valid evidence for translating knowledge and making good decisions in

VR (Chan et al., 2009; Leahy et al., 2009; Leahy & Arokiasamy, 2010; Leahy et al., 2014b).

The evidence-based practice (EBP) has become a new norm today by conducting valid research and gathering reliable data for improving practices and outcomes (Eignor, 2013). In education (including rehabilitation counseling), the EBP research along with well-constructed designs and experiments can provide fundamental and significant improvements over practices. Not only can the proper use of EBP results help make better decisions about individuals (e.g., people with disabilities) and programs (e.g., VR agencies), but it can also provide a successful pathway to gaining broader access to quality education or full employment, according to the standards for research conduct by educational researchers (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014).

Professionals in the field of rehabilitation counseling, such as VR counselors and practitioners, are often expected to integrate clinical judgement skills (including scientific attitude, cognitive complexity, evidence-based practice, and counselor biases) with research evidence via scientific-based methods to make best informed decisions that maximize the well-being outcomes of the clients (e.g., people with disabilities in public VR) (Austin & Leahy, 2015; Menon et al., 2009). The emphasis of best EBP lends VR counselors a significantly renewed impetus, so that they can be more accurate in clinical judgement by getting research-informed knowledge in clinical issues of interventions and outcomes (Chan et al., 2010).

Since the Pearson's correlation coefficient was introduced last century, it has been used as one of the most important statistical tools for scientific inquiries in educational and social research (Agresti & Finlay, 2009; Fisher, 1915; Olkin & Pratt, 1958; Pearson & Lee, 1903; Pearson, 1904; Pearson, 1920; Soper et al., 1917; Student, 1917; Thorndike, 2005). When using

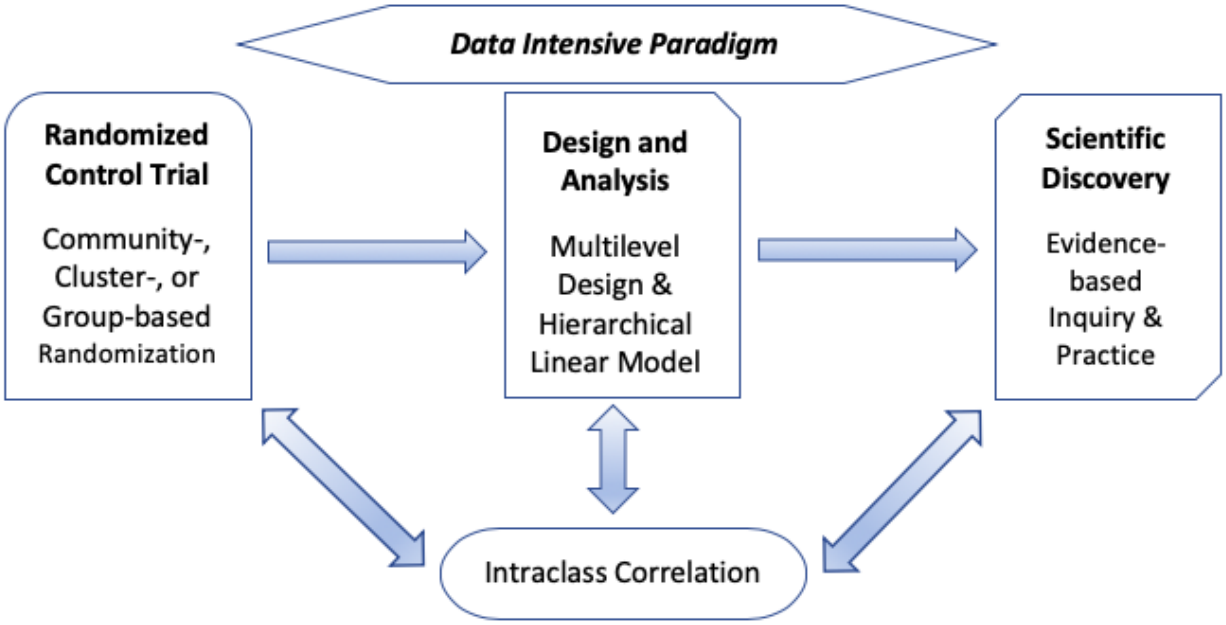
correlation to interpret statistical results, researchers have to be aware of "correlation does not imply causation" and always need to be cautious about post hoc fallacy (Latin: "post hoc, ergo, propter hoc"; English: "after this, therefore, because of this") – whereas, this issue, which has the potential for an informal fallacy, can be rectified by using a well-designed experiment, by means of which researchers are more likely to go the extra mile to obtain valid statistical inference or even causality in studies (Fisher, 1925a, 1942, 1958a, 1958b; Holland, 1986).

Of the different types of effect magnitude measures for the correlation ratio (e.g., Intraclass Correlation, Eta-squared, Omega-squared, R-squared, and Rho-squared indexes), the intraclass correlation (ICC) is a parametric estimator in the random-effect (or mixed-effect) model to quantify the true proportion of total variance accounted for in the outcome variable (Hays, 1994; Raudenbush & Bryk, 2002). Furthermore, the ICC can summarize the clustering effect magnitude (i.e., the relatedness) at a hierarchical design (Note: In statistics, this technique is called the "random" coefficients in multilevel models) (Hays, 1994; Hedges & Olkin, 1985). In this study, one main research goal is to investigate the ICC in hierarchical linear models (HLM) and hierarchical generalized linear models (HGLM) by using the mixed-effect analysis of variance (ANOVA), in order to better understand the ICC's statistical properties on different simulation-based scenarios with respect to complex modeling structures and sampling designs.

Both RCTs and CRTs have been widely viewed as the one of the best EBP approaches (i.e., the gold standard) for appraising and measuring the efficacy and effectiveness of interventions or treatments in educational and social research studies, since not only can such a methodology for designing experiments efficiently identify "how works" and "what works" in relation to the intervention or treatment given by using an experimental design, but it can also effectively provide more robust and valid evidence in EBP for scientific inquiry and research

(Connolly et al., 2018; Menon et al., 2009; Schneider et al., 2007; Sullivan, 2011).

Figure 1.1 Conceptual Flowchart of the Intraclass Correlation Study at Hierarchical Design



The current study is to address the following research questions under an EBP paradigm with the hierarchical design (i.e., CRT) driven by the intraclass correlation coefficient (ICC) analytic strategies (see Figure 1.1 for an illustration of the overall "big picture" concept of ICC). Moreover, this research is to evaluate the statistical performance of ICC in various simulation-based scenarios designed by the complex hierarchical data structures through the existing RSA-911 data set, where clients in VR were represented as real-world connections into computer simulations in the two-level CRT setting (i.e., clients are in level 1, and offices are in level 2). Also note that, in simulations using real data via the RSA-911, the selected variables are incorporated into multilevel modeling (i.e., HLM & HGLM) to represent the pivotal VR relationships between demographic characteristics, rehabilitation services, and employment

outcomes. In order to answer the proposed research questions, a computer simulation study (i.e., the Monte Carlo Method) is conducted using the bootstrapping procedure with the real data of RSA-911 (Note: The bootstrap method is a resampling technique without replacement from a given sample). More details of this computer simulation framework with RSA-911 data are provided in Chapter 4 (Methods and Research Questions) and Chapter 5 (Results).

This study is to address the following three research questions with respect to the ICC.

Research Question 1. Consider RSA-911 data for those people with disabilities served in Michigan in FY 2015. What are the empirical distributions of ICC (estimate, standard error, p-value and 95% confidence limits) for the usable samples of RSA-911 data?

(a). Compare the method performance of statistical estimation and inference among Models 1-4, where Model 1 is fully unconditional, Model 2 is conditional on individual characteristics of gender, minority, age, education and social security insurance benefits, Model 3 is conditional on rehabilitation service predictors (job placement, on-the-job supports and rehabilitation technology), and Model 4 is a combination of Models 2 and 3.

(b). What are the empirical distributions of ICC estimates given by different breaking variables (disability type, disability significance and severity, and previous work experience) for subset analysis under Models 1-4? What are the differences among Models 1-4 in (a) and (b)?

Research Question 2. Given the cluster randomized design structure of RSA-911 data, there are three different "cluster" settings at level 2 - the number of groups = 5, 15, or 25; and there are three "individual" settings at level 1 - the number of subjects = 50, 100, or 150. Based on the bootstrapping procedure by 100 times, what are the empirical distributions of ICC (estimate, standard error, and p-value) in each bootstrap scenario under Models 1-4?

(a). Given by each of bootstrap resampling scenarios (the number of bootstrap repetitions=100), compare the ICC estimates among Models 1-4 and examine which model (from Models 1-4) can provide better statistical performance of ICC estimation and inference.

(b). Evaluate which bootstrap sampling scenarios (based on the number of groups and the number of subjects) can provide more accurate and precise ICC estimates (i.e., less bias and less mean squared error in statistical estimates)? What are the recommended sampling strategies (the number of groups and subjects) for cluster randomized trials using RSA-911 data?

Research Question 3. Comparing the results between Research Question 1 (RQ1: Population Model) and Research Question 2 (RQ2: Resampling Model), which model (in Models 1-4) can provide the best statistical properties of ICC estimation and inference, in terms of statistical bias (expected difference in ICC estimates between RQ1 and RQ2), mean squared error (mean squared deviations in ICC estimates between RQ1 and RQ2), and ICC parameter coverage rate (proportion of true parameter "hits" by 95% confidence interval for ICC, based on the subsamples in RQ2, in comparison with the overall sample result in RQ1)?

The next two chapters are to present both the literature review of statistical methods and applications for intraclass correlation plus the motivation for the study (Chapter 2), as well as the literature of statistical approaches in rehabilitation counseling using RSA-911 data (Chapter 3). The rest of the dissertation is organized as follows. In Chapter 4, it covers a mathematical framework and notation in the proposed methodology for investigating intraclass correlation in multilevel structure. In Chapter 5, it shows the results using the real data set of RSA-911 via an exploratory bootstrap simulation approach to ICC estimation and related statistical inference. Last but not least, simulation results and study findings are discussed in Chapter6.

CHAPTER 2

LITERATURE REVIEW OF STATISTICAL METHODS

In this chapter, a comprehensive introduction to the history of intraclass correlation coefficients (ICC) at experimental designs is provided to serve a basic framework of this study. ICC has been one of the oldest statistical measures since Sir Ronald A. Fisher coined it last century. The fundamental idea of ICC is presented first to show the basic context of intraclass correlation, and then is followed by a series systematic review of its developments in statistical estimation & hypothesis testing, effect size measurement, and Fisher transformation using ICC. In addition, a review of the literature pertinent to the current major developments in ICC by Allan Donner, Larry Hedges, and Tenko Raykov, as well as their proposed analytic strategies for using ICC, are all provided to serve a fundamental basis of the study and then to understand the ICC's statistical phenomenon at multilevel design especially for cluster randomized trials.

2.1 *Fisher Approach*

Since the correlation coefficient was introduced last century, it has been used as one of the most popular and important tools in scientific inquiries including biometrical work as well as social and educational research studies (Fisher, 1915; Olkin & Pratt, 1958; Pearson, 1920; Rodgers & Nicewander, 1988; Soper et al., 1917; Student, 1917). The inheritance of physical and mental characters in human is one classical example to show how powerful this statistical tool can be applied to across all our scientific fields. For example, Pearson and his colleague used U.K. school children data in the late 1800s to investigate a variety of basic human

mechanisms from physical characteristics (e.g., age, body size, stature, and even eye color) to latent or psychic abilities (e.g., mental status or intelligence), and further to compare those measures, using Person product-moment correlation, to understand ancestral heredity, natural inheritance, and family resemblance (Pearson & Lee, 1903; Pearson, 1904).

When using correlation to interpret statistical results, researchers need to be aware of "correlation does not imply causation" and always be cautious about post hoc fallacy (Latin: "post hoc, ergo, propter hoc"; English: "after this, therefore, because of this"), although the issue can be dealt with by a carefully designed experiment (like randomized control trials), and it may help go the extra mile to test causality and further to make a valid statement of causal inference (Fisher, 1958a, 1958b; Holland, 1986). In the experimental field, scientific inquiries can be done synthetically with three key ingredients – replication (for adding precision), randomization (for bringing validity), and control (for reducing interference), and so research workers therefore are able to reach out safely (or "with statistical soundness and completeness") and then obtain fiducial and unchallengeable conclusions (Fisher, 1958a, pp. 409-410); on the other hand, in the observational study, some may be found it useful in the exploratory stages to express a statistical inquiry in the form of a correlation coefficient, but, with the previous cautious statement "correlation is not causation," it is seldom to draw a valid foundation of making causal links rather than simply to produce spurious correlations or even counterfactual connections, due to a reasonable suspicion that, if any, various possible contributory causes of a studied phenomenon cannot be controlled (Fisher, 1925a, Chapter Six "The Correlation Coefficient"), unless researchers use other remedial and modified methods (like quasi-experimentation or regression discontinuity) to circumvent, or at least to alleviate, the difficulty and problem by adjusting "uncontrolled observations (or uncontrollable events)"

with “artificially controlled (or statistically manipulated)” quasi-experimental conditions (i.e., pseudo experimental models) to appropriately but properly estimate causal impacts (i.e., treatment or intervention effects) using Neyman-Rubin Model (Schneider et al., 2007).

In a theoretical perspective, mathematical features (algebraic relationships) and key properties (statistical functions) of Pearson’s correlation coefficient are listed as follows.

Let $(x_1, y_1), \cdots, (x_N, y_N)$ are N pairs of independent samples with bivariate normal with means $[\mu_1, \mu_2]$, variances $[\sigma_1^2, \sigma_2^2]$ and correlation ρ. The frequency can be written in the form

$$f(x, y \mid \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}}$$

, where the correlation ρ may be positive or negative or zero but cannot exceed unity in magnitude (Fisher, 1925a; Roussas, 2002). If one variate has an assigned value (e.g., $x = a$), then by giving x a constant value a, this conditional frequency (i.e., the total frequency above is divided by the frequency with which $x = a$ occurs) can be expressed by a general formula

$$f(y \mid x = a) = \frac{1}{\sqrt{2\pi(1-\rho^2)}\ \sigma_2}\, e^{\frac{-1}{2(1-\rho^2)\sigma_2^2}\left\{y - \left[\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(a-\mu_1)\right]\right\}^2}$$

, where the conditional distribution (y of x given a) is normal with mean $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(a - \mu_1)$ and variance $(1-\rho^2)\sigma_2^2$, and it implies that the total variance of y in the fraction $(1-\rho^2)$

is independent of x, while the remaining variation of y in the fraction ρ^2 is determined by (and calculable from) the value of x (Fisher, 1925a; Mood, Graybill, & Boes, 1974).

The statistical estimation of the correlation is the ratio of the covariance to the geometric mean of the two variances; if $S(x)$ and $S(y)$ represent the deviations of the two variates from their means $[\hat{\mu}_1, \hat{\mu}_2]$, then the correlation coefficient (or product moment) estimator r would be given by

$$r = \frac{S(xy)}{\sqrt{S(x)S(y)}} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_1)(y_i - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{N}(x_i - \hat{\mu}_1)^2 \sum_{i=1}^{N}(y_i - \hat{\mu}_2)^2}}$$

, where the mean estimates $[\hat{\mu}_1, \hat{\mu}_2]$ can be approximated by sample means $[\bar{x}, \bar{y}]$.

By Olkin and Pratt (1958), the probability density of r is derived as

$$f(r) = \frac{2^{N-2}}{\pi\,\Gamma(N-1)}(1-\rho^2)^{N/2}(1-r^2)^{(N-3)/2}\,F\left(\frac{1}{2}, \frac{1}{2}; \frac{(N-1)}{2}; 1-r^2\right)$$

, where $F(\propto, \beta; \gamma; x) = \sum_{k=0}^{\infty} \frac{\Gamma(\propto+k)\Gamma(\beta+k)\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma+k)} \frac{(x)^k}{k!}$ is the hypergeometric function, and the last term therefore can be computed and simplified as $\sum_{k=0}^{\infty} \Gamma^2\left(\frac{N+k}{2}\right) \frac{(2\rho r)^k}{k!}$. It is noteworthy that under the null hypothesis of ρ is true (i.e., $H_O: \rho = 0$), the asymptotic distribution of a sample correlation r is a normal density with mean 0 and variance $(1-r^2)^2$. In the general case of ρ (i.e., $-1 \leq \rho \leq 1$), by using Laplace transformation and Taylor series expansion through on that previous density function of sample correlation r, Olkin and Pratt (1958) derived the

uniformly minimum-variance unbiased estimator (UMVUE: an unbiased estimator that has lower variation error than any other unbiased estimators for all plausible values of the parameter), which is shown to be

$$UMVUE(\rho) \cong r + r\,\frac{1-r^2}{2(N-1)} + r\,\frac{9(1-r^2)^2}{8(N^2-1)} + r\,O(N^{-3}) \approx\ r + r\,\frac{1-r^2}{2(N-3)}$$

Note that there is another simple estimator of correlation coefficient to adjust biased correlation especially for a small sample size $(N < 30)$, according to Kelly (2018) and Flom (2015):

$$r = sgn(r)\sqrt{1 - \frac{(1-r^2)(N-1)}{(N-2)}}$$

, where the formula is resulted from adjusted $R^2 = 1 - (1-R^2)\frac{(N-1)}{(N-P-1)}$ for $P = \#$ predictors.

To test whether a correlation is different from zero (i.e., $H_O: \rho = 0$), the test statistic is

$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}} = r\sqrt{\frac{N-2}{1-r^2}}$$

, which is t-distributed with $\nu = N - 2$ degrees of freedom (Lomax & Hahs-Vaughn, 2012, pp.267-268; Roussas, 2002, pp. 472-473). It is interesting to note that the probability density of a correlation (when $\rho = 0$) can be found using a linear transformation of t-statistic above

by

$$f(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((N-1)/2)}{\Gamma((N-2)/2)} (1-r^2)^{(N/2)-2}$$

, where this density is only true for the case of $\rho = 0$ (independence) (Roussas, 2002, pp. 474).

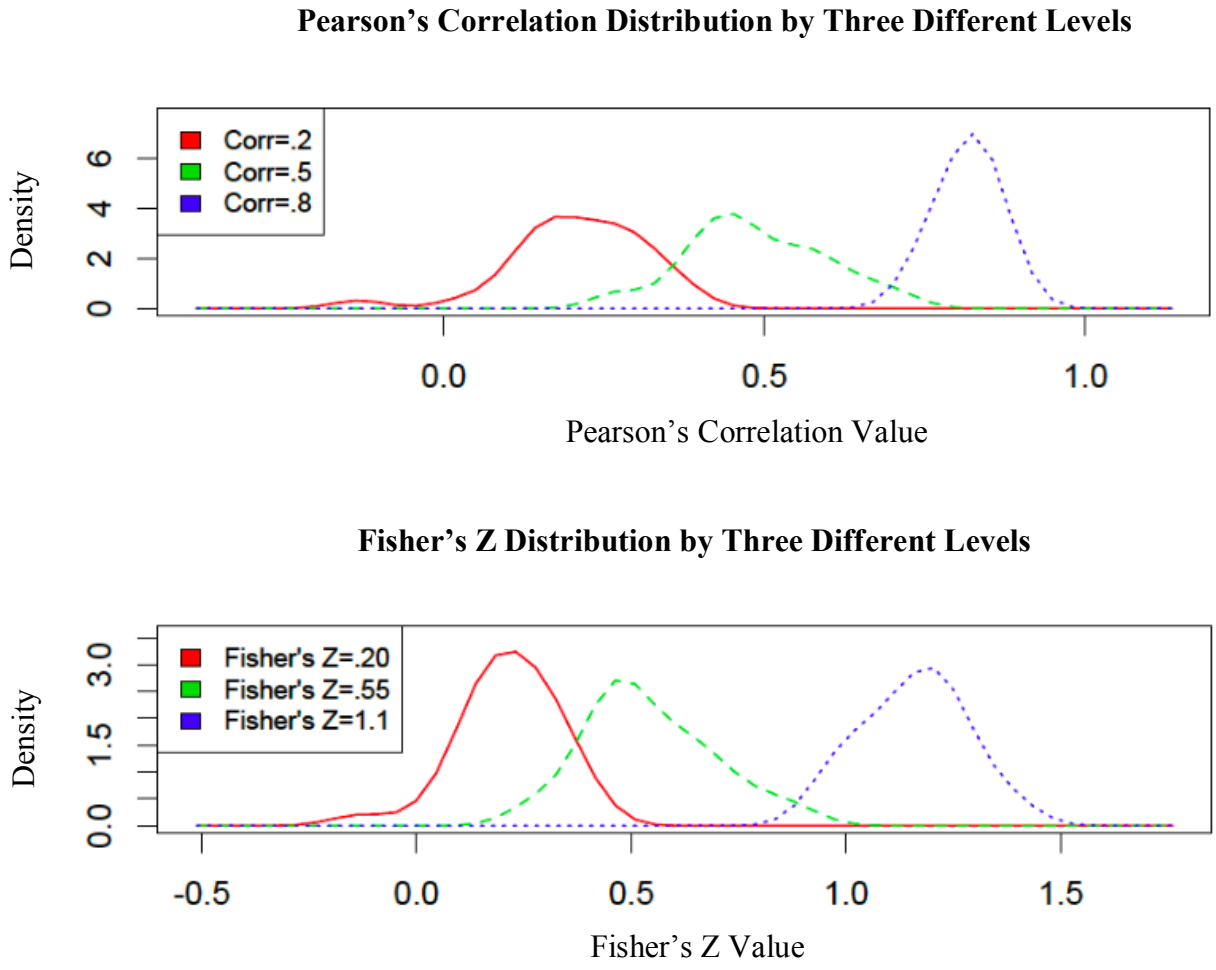
Comparing to the previous t-statistic approach of testing significance of a correlation coefficient, transformed correlations is another way to deal with the issue of testing the significance of an observed correlation coefficient. By using a well-known standard normal Z testing statistic, Fisher (1925a) proposed a more reliable and accurate transformation method that employs the information of a given correlation r to approximate to the standard normal Z distribution in which this test can be carried out without much difficulty in laborious calculation. The Fisher's Z transformation is defined as the formula

$$Z = \frac{1}{2} log_e \left\{ \frac{1+r}{1-r} \right\} = 1/2 \{ log_e(1+r) - log_e(1-r) \} = \sum_{i=1}^{\infty} \frac{1}{1+2(i-1)} r^{1+2(i-1)}$$

, where the statistic value Z ranges from 0 to $\pm\infty$ as the sample correlation r changes from 0 to ± 1 , the Fisher's Z can also be approximated by $r + r^3/3 + r^5/5 + r^7/7 + \cdots$, and the standard error of Z is derived in a simpler form approximately as $\sigma_Z = 1/\sqrt{N-3}$ which is practically independent of any value of correlation in the population from which the sample is drawn. There are three advantages of this transformation of r into Z (Fisher, 1925a, pp. 198-199) : (1) the standard error of Z does not depend on the true value of the correlation ρ , so can provide a true weight for the value of the estimate (i.e., σ_Z is a so-called ancillary statistic

which contains no information about the parameter interest ρ, but sometimes it paradoxically provides "additional" valuable information in statistical inference say like our knowledge of the accuracy and precision of the estimate Z here; Casella & Berger, 2002, pp. 282-284; Cox, 1971; Efron & Hinkley, 1978; Fisher, 1925b, p. 724); (2) although the distribution of r is not normal in small samples and even remains far from normal for large samples with a high correlation (e.g., the correlation ρ is close to ± 1) , the sampling distribution Z still tends to converge to asymptotic normality as the sample size N increases, no matter what the value of the correlation may be (either large or small, positive or negative); (3) while the distribution of r changes rapidly in terms of its shape (i.e., skewness and kurtosis) as the parameter ρ is changed (given by $\rho \in [-1\,, 1]$), the sampling distribution of Z is probabilistically more stable and nearly constant in the form of a symmetrical bell shape (i.e., values are normally distributed) and therefore it would be reasonable to assume a sample correlation by Fisher's Z transformation follows approximate normality with mean the true correlation parameter ρ and variance $1/(N-3)$. Also see below Figure 2.1 for demonstrating the comparisons of the sampling distributions between non-transformed and transformed correlation coefficients at the three different levels (i.e., correlation coefficients ρ are set at 0.2, 0.5, and 0.8, while Fisher's Z are given by 0.20, 0.55, 1.10, respectively). In Figure 2.1, it shows that Fisher's Z distributions are relatively more robust and stable than "raw" distributions of non-transformed Pearson's correlation coefficient across the continuum domain (i.e., $\rho \in [-1\,, 1]$ and $Z \in \mathbb{R}$).

Figure 2.1 Sampling Distributions of Non-Transformed and Transformed Correlations at Three Different Levels



Note. The original idea of this graph (Figure 2.1) comes from Fisher's Z transformation (1925a, p.200). The upper panel demonstrates the sampling distributions of correlation at the levels of r = 0.2, 0.5, and 0.8; and the lower panel shows the respective sampling distributions by Fisher's Z transformation in which the values are shown as z = 0.20, 0.55, and 1.10.

In terms of correlation-based measures, Jacob Cohen (1988, pp. 77-81) proposed that $|r| < 0.2$ (or the threshold of $|r|$ around 0.1) as a small or weak effect, $0.2 < |r| < 0.4$ (or $|r|$ around 0.3) as a medium or moderate effect, and $0.4 < |r| < 0.6$ (or $|r|$ around 0.5) as a large or strong effect, and $|r| > 0.6$ (or $|r|$ around 0.7 or above) as a very large or extremely strong effect, to determine the effect size magnitude of a studied phenomenon of interest (Cohen, 1988; Ellis, 2009; Rosenthal, 1996). It is cautious to note that these standards for correlation thresholds may need to be modified or even re-evaluated & re-justified in different areas of scientific inquiries, especially for the fields other than behavioral and social sciences (such as clinical and social psychology), since J. Cohen (as a clinical and social psychologist) was originally working on this effect-size magnitude research using the data in his field (specifically, unique to psychology and social sciences) for developing "qualitative" descriptors of strength of association with respect to a "quantitative" product-moment r.

In the family of effect size measures of correlation, there are other types of effect size estimates that are calculated based on different variance components (e.g., effect magnitude (EM) = [explained variance] / [total variance] $\equiv \rho^2$, which is translated into plain language – EM is the amount of the explained variance can be accounted for by the total variation within an experimental design model; Cohen, 1988, p. 78.) For instance, the coefficient of determination (aka R-squared, or $R^2 = SS_{Regression}/SS_{Total}$) is widely known and used especially in regression models. In addition, the correlation ratio Eta-squared ($\eta^2 = SS_{Between}/SS_{Total}$) is another form of the squared correlation in analysis of variance (ANOVA) models (Pearson, 1923; Richardson, 2011). Also, Hays (1994) introduced a similar one – the omega-squared index (ω^2) – as a ratio of the relative reduction in uncertainty say about Y due to X, which shows the variance component in Y given by X, and this index can be

described as $\omega^2 = \left(\sigma_Y^2 - \sigma_{Y|X}^2\right)/\sigma_Y^2$. Last but not least, the intraclass correlation coefficient (ICC) is defined as

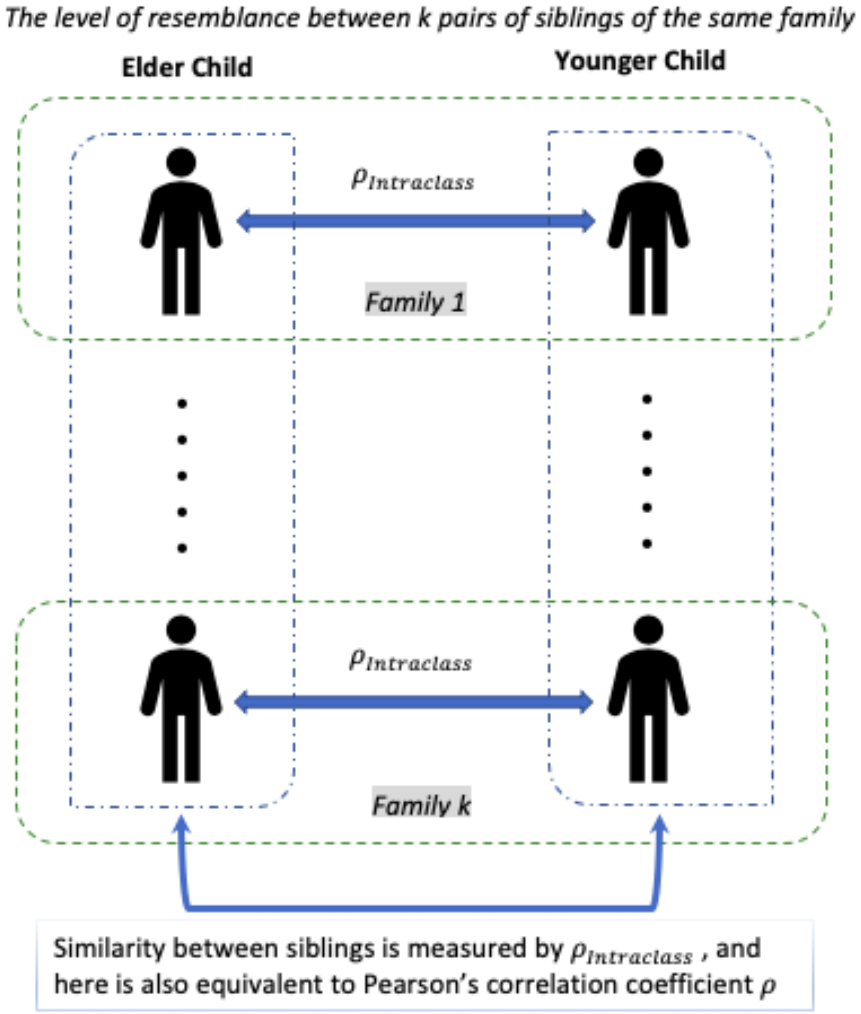
$$\rho_I = \sigma_{Between-Group}^2 / \left(\sigma_{Between-Group}^2 + \sigma_{Within-Group}^2\right)$$

, the formula of which is another idea to quantify the true proportion of variance accounted for in the outcome (by cluster effect) in random-effect mixed models (Hays, 1994; Raudenbush & Bryk, 2002). Note that the intraclass correlation (or the so-called cluster effect) is defined only in the random-effect (esp. random-intercept) models, while the omega-squared index can also be used in the fixed-effect analysis (Hays, 1994, p.535; Hedges & Olkin, 1985, p. 103). In this study, one main focus is to investigate ICC in hierarchical models (mixed effects ANOVA) so as to better understand its properties on different scenarios by design effect and sample size.

Another application of the use of intraclass correlation is to measure the level of similarity or resemblance (Fisher, 1925a; see Figure 2.2 below as an illustration of intraclass correlation). In one case like plant biology fields, the resemblance between leaves or pods on the same tree was studied say by picking 30 seed pods from a number of different 100 trees. In another case of human & family correlation studies, for example, we have a sample of anthropometric measurements of about 1500 pairs of siblings of the same family (e.g., two classes: elder kid vs younger kid); and we may want to calculate correlation between siblings. Here, if an association of interest is based on differences between two classes (or groups) of measurements, then it would be so-called "interclass" correlation that is also equivalent to a typical Pearson's correlation coefficient ρ between two sets of measurements. On the other hand, suppose that all the subjects (e.g., a combination of both older and younger siblings)

belong to the same class (only one group of a single whole study overall) with a common mean and a common standard deviation about that mean for all measurements, and then correlation now is distinguished as "intraclass" correlation (Fisher, 1925a, pp. 211-215).

Figure 2.2 Intraclass Correlation Between Two Classes of Measurements



Note. This illustration of ICC is motivated by an original idea by Fisher (1925a).

In the special case of having two classes of measurements given by *N* pairs of samples $(x_{11}, x_{12}), \cdots, (x_{N1}, x_{N2})$, intraclass correlation is defined as

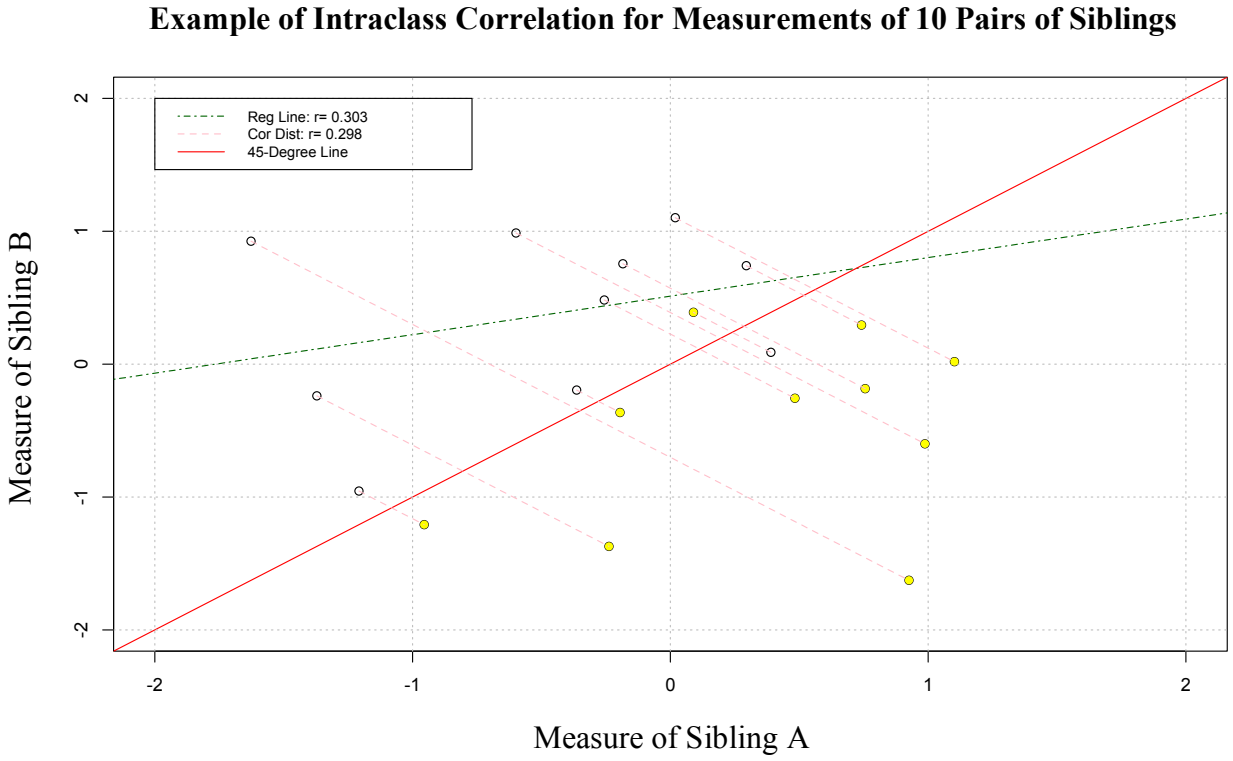
$$\rho_{Intraclass} = \frac{\sum_{i=1}^{N}\{[x_{i1} - \bar{X}][x_{i2} - \bar{X}]\}}{NS^2}$$

, where the common mean is $\bar{X} = (2N)^{-1}\sum_{i=1}^{N}[x_{i1} + x_{i2}]$, and the common variance is $S^2 = (2N)^{-1}[\sum_{i=1}^{N}(x_{i1} - \bar{X})^2 + \sum_{i=1}^{N}(x_{i2} - \bar{X})^2]$. When it considers the general case of having a set of k classes of measurements given by N samples with $[\bar{x}_1, \dots, \bar{x}_N]$ representing a set of means from the k classes in each sample, the general formula of intraclass correlation can be written by

$$\rho_{Intraclass} = \frac{[k\sum_{i=1}^{N}(\bar{x}_i - \bar{X})^2] - NS^2}{NS^2(k-1)}$$

, where the common mean is $\bar{X} = (kN)^{-1}\sum_{i=1}^{N}[x_{i1} + x_{i2} + \cdots + x_{ik}]$, the common variance is $S^2 = (kN)^{-1}[\sum_{i=1}^{N}(x_{i1} - \bar{X})^2 + \sum_{i=1}^{N}(x_{i2} - \bar{X})^2 + \cdots + \sum_{i=1}^{N}(x_{ik} - \bar{X})^2]$, and the range of intraclass correlation values is always positive or should not be less than $-1/(k-1)$. See Figure 2.3 for a geometric interpretation of ICC by illustrating the resemblance of 10 paired observations (i.e., siblings of A's and B's) as to some measure of within-pair association (or intraclass correlation) between the two siblings in the same family. It is interesting to note that the ICC in Figure 2.3 can be geometrically represented as well as numerically approximated by the overall Euclidean distance (or norm) between the paired samples on the standardized scale (i.e., the overall Euclidean length can be defined by the standardized difference between the measures of sibling A and sibling B in the Cartesian coordinate system $\mathbb{R}^2$).

Figure 2.3 Demonstration Example of Intraclass Correlation by Two Classes of Measurements



Note. This illustration comes from the concept of intraclass correlations by Fisher's approach of having the common mean and standard deviation for all the measurements (1925a, Section 38 of Intraclass Correlations and the Analysis of Variance, pp.211-214). The intraclass correlation (or within-pair correlation) can be estimated by the Euclidian distance of the paired measurements between the two related groups of samples (i.e., the true ICC is set at 0.303, and the estimated ICC is given by 0.298 using the standardized length between a pair of measurements from Sibling A and Sibling B).

2.2 *Donner Approach*

In the analysis of family data, it is frequent to use the intraclass correlation coefficient to measure the degree of intra-family resemblance among family members with regard to family health history in quantitative traits of biological or psychological attributes such as human body proportion (e.g., arm's span, leg's length), blood pressure level, and cognitive intelligence (IQ). Donner & Koval (1980a) derived the maximum likelihood estimator (regarding no prior knowledge of statistical estimates) of the intraclass correlation $\rho_{Intraclass}$ using multivariate normal theory in variance component models (assuming unequal group/family sample size).

In statistical theory, suppose one observation on the *j*-th member ($j = 1, \dots, n$) of the *i*-th family ($i = 1, \dots, k$) is used to investigate the intraclass resemblance $\rho_{Intraclass}$ among the class of n samples from each of k families, which can be stated mathematically as

$$Y_{ij} = \mu + a_i + e_{ij}$$

, where Y_{ij} is an observation for which *i* is the index of a family or group factor ($i = 1, \dots, k$) and *j* is an individual member within that family or group factor ($j = 1, \dots, n$), μ is the grand mean of all the observations, a_i is designed as the random effect (identically distributed) with mean 0 and variance σ_a^2 (i.e., NID(0, σ_a^2)), e_{ij} is a random normal error term for *j*-th subject in *i*-th group (i.e., independently and identically distributed with mean 0 and variance σ_e^2; viz., NID(0, σ_e^2)), and both random components, $\{a_i\}$ and $\{e_{ij}\}$, are assumed to be mutually independent.

By summing the additive variance components (i.e., a sum of both between-group and within-group variation is equal to total variation), the variance of Y_{ij} is given by $\sigma_Y^2 = \sigma_a^2 + \sigma_e^2$, and then the intraclass correlation is defined by the Fisher's concept as $\rho_{Intraclass} = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, where this index will be zero when $\sigma_a^2 = 0$, and it will be unity if $\sigma_e^2 = 0$ (assuming that $\sigma_Y^2 > 0$). Notice that the intraclass correlation represents the true proportion of variance attributable to Factor a, and that the intraclass correlation is similar to the omega-squared index (ω^2) in the general form, although the intraclass correlation ($\rho_{Intraclass}$) applies to the random-effect model but the omega-squared index (ω^2) often only to the fixed-effect model (Hays, 1994, p.535).

Equivalently, from a point of view of statistical theory, the intraclass correlation can also be fundamentally defined as the ordinary correlation coefficient between any two observations in the same class (group or family), say Y_{ij} & Y_{ik}, since their statistical relationship holds that

$$Corr(Y_{ij}, Y_{ik}) = E\{[(Y_{ij} - \mu)(Y_{ik} - \mu)]/\sigma_Y^2\} = E(a_i^2)/\sigma_Y^2 = \sigma_a^2/\sigma_Y^2$$

, where $Cov(\{a_i\},\{e_{ij}\}) = Cov(\{a_i\},\{e_{ik}\}) = 0$, and $a_i \sim NID(0, \sigma_a^2)$ (Donner & Koval, 1980a).

Since the traditional method (Fisher's approach) above requires distributional assumptions of observations (based upon multivariate normal theory), it is the analysis of variance (ANOVA) that provides an alternative estimator of intraclass correlation (for relaxing the assumptions) in the classical linear models (Donner & Koval, 1980a). The new

practical method for estimating intraclass correlation is to utilize relevant information in the ANOVA table shown as following (without loss of generosity, it is assumed to be a balanced design with equal group/family size).

Table 2.1 Analysis of Variance (ANOVA) for Intraclass Correlation (ICC) Calculations

Source of Variation	Degree of Freedom (DF)	Sum of Squares (SS)	Mean Squares (MS)	F Statistic
Among Groups	k-1	SSA	MSA	MSA / MSW
Within Groups	k(n-1)	SSW	MSW	
Total	kn-1	SST		

, where the between-group variation SSA = $\sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2$, the within-group variation SSW = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij}-\overline{Y}_{i\cdot}\right)^2$, the total variation SST = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij}-\overline{Y}_{\cdot\cdot}\right)^2$, the mean squares among groups MSA = SSA / DF(Among Groups) = SSA / (k-1) = $\sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot})^2/(k-1)$, the mean squares within groups (or the mean squared error) MSW = SSW / DF(Within Groups) = SSW / [k(n-1)] = $\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij}-\overline{Y}_{i\cdot}\right)^2/[k(n-1)]$, the between-group degrees of freedom DF(Among Groups) is $k-1$ (for k = the number of groups), the within-group degrees of freedom DF(Within Groups) is $k(n-1)$ (for n = the number of within-group subjects).

It is interesting to note that, by Hays (1994, pp. 533-535), the expectation of mean square among groups E[MSA] = $n\sigma_a^2+\sigma_e^2$, and that the expectation of mean square within groups E[MSE] = σ_e^2 (i.e., MSE is an unbiased estimate of error variance; Hays, 1994, p.532). Therefore, the intraclass correlation estimator can be indirectly obtained in such a way (via ANOVA) that:

$$\rho_{Intraclass} = \sigma_a^2 \Big/ \sigma_Y^2 = E[MSA] - E[MSE] \Big/ E[MSA] - (n-1) \times E[MSE]$$

, where the total variance consists of two independent variance components and hence is given by $\sigma_Y^2 = (\sigma_a^2 + \sigma_e^2)$ for $\sigma_a^2 = \{E[MSA] - E[MSE]\}/n$ and $\sigma_e^2 = E[MSE]$; the best estimate of the total variance ($\hat{\sigma}_Y^2$) is to use the estimates of group variance ($\hat{\sigma}_a^2$) and error variance ($\hat{\sigma}_e^2$), so that $\hat{\sigma}_Y^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2 = [MSA - (n-1) \times MSE]/n$. Also notice: an unbiased estimate of group variance may be found $\hat{\sigma}_a^2 = 0$ when MSE is greater than or equal to MSA (Hays, 1994, p.534).

For an unbalanced "natural" design with unequal family or group size n_i (for $i = 1, \ldots, k$) in ANOVA, the common family or group size n_0 is calculated for representing the mean within-group individuals, and the intraclass correlation coefficient (Donner & Koval, 1982) is given by

$$\rho_{Intraclass} = (MSA - MSW) \Big/ [MSA + (n_0 - 1)MSW]$$

, where $n_0 = [N - \sum_{i=1}^{k} n_i^2/N]/(k-1)$ and N is defined by the number of total sample size (i.e., $N = \sum_{i=1}^{k} n_i$). Also note that, by Donner & Koval (1980a), the mean within-group subjects can be alternatively calculated by $n_0' = \bar{n} - \sum_{i=1}^{k} (n_i - \bar{n})^2/[N(k-1)]$, where the approximate group size $\bar{n} = \sum_{i=1}^{k} n_i/k = N/k$, and this latter formula of the average within-group size (n_0') is mathematically equivalent to the former (n_0), yet the computation (n_0') is more laborious. Since $\tilde{\sigma}_a^2 = (MSA - MSW)/n_0'$ and $\tilde{\sigma}_e^2 = MSW$ are deemed, respectively, as

the unbiased estimates of σ_a^2 and σ_e^2, it is intuitive and straightforward to find the estimator of intraclass correlation via the Fisher's definition:

$$\rho_{Intraclass} = \tilde{\sigma}_a^2/(\tilde{\sigma}_a^2 + \tilde{\sigma}_e^2) = {(MSA - MSW)}\Big/{[MSA + (n_0' - 1)MSW]}$$

, where it is equivalent to the previous formula due to $n_0' = n_0$ (Donner & Koval, 1980a).

As for statistical testing of intraclass correlation, by Donner & Koval (1980a), there is a test of significance for the estimate of intraclass correlation in analysis of variance using *F*-distribution with $k - 1$ and $N - k$ degrees of freedom at the chosen level of significance, with respect to testing the hypotheses $H_o: \rho_{Intraclass} = 0$ vs. $H_a: \rho_{Intraclass} > 0$. A significant *F* testing statistic value (i.e., $\rho_{Intraclass} > 0$) implies that members of the same group tend to be more alike and similar to each other with respect to the attribute or characteristic in question than those from a different group, and also that the estimated intraclass correlation coefficient shows the idea of the true proportion of variance accounted for in the population by that factor of interest (e.g., families or groups).

For the sake of another mathematical and statistical expression of the intraclass correlation index, the intraclass correlation coefficient can be re-defined using the quantity $\theta = \sigma_a^2 / \sigma_e^2$ as

$$\rho_{Intraclass} = \theta/(1 + \theta)$$

, where there is a basic statistical assumption of the normal distribution for the random effect ($a_i \sim NID(0, \sigma_a^2)$) and the error term ($e_{ij} \sim NID(0, \sigma_e^2)$) (Hays, 1994, p.535). Further, in linear

modeling theory (Hays, 1994, pp.535-536; Kutner et al., 2005, pp.1040-1041; Stapleton, 2009, p.285), the testing statistic of the proposed intraclass correlation estimator can be shown that

$$F_0 = {MSA/(n \times \sigma_a^2 + \sigma_e^2)}\Big/{MSW/\sigma_e^2} = \frac{MSA}{MSW}\left(\frac{1}{1+n\times\theta}\right)$$

, where this proposed method is mainly based on the random-effect ANOVA with a balanced design, and it follows an F distribution with $df_1 = k-1$ and $df_2 = k(n-1) = N-k$ degrees of freedom, so that a $100(1-\alpha)\%$ confidence interval on $\theta = \sigma_a^2\ /\sigma_e^2$ can be obtained by

$$1-\propto = P\left(F_{\frac{\alpha}{2},df_1,df_2} \le F_0 \le F_{1-\frac{\alpha}{2},df_1,df_2}\right)$$

$$= P\left(\frac{1}{n}\left[\frac{F^*}{F_{1-\frac{\alpha}{2},df_1,df_2}} - 1\right] \le \theta \le \frac{1}{n}\left[\frac{F^*}{F_{\frac{\alpha}{2},df_1,df_2}} - 1\right]\right)$$

, where $F^* = MSA/MSW$ is the sample F ratio value in ANOVA table. By the algebraic relationship $\rho_{Intraclass} = \theta/(1+\theta)$, the corresponding interval for intraclass correlation is

$$P\left(\frac{F^* - F_{1-\frac{\alpha}{2},df_1,df_2}}{F^* + (n-1)\times F_{1-\frac{\alpha}{2},df_1,df_2}} \le \rho_{Intraclass} \le \frac{F^* - F_{\frac{\alpha}{2},df_1,df_2}}{F^* + (n-1)\times F_{\frac{\alpha}{2},df_1,df_2}}\right) = 1-\alpha$$

, where this confidence limit, with confidence coefficient $1-\alpha$, for intraclass correlation

$\rho_{Intraclass}$ represents the degree of total variability accounted for by the mean differences among different factor levels (or the effect of the extent of variation between groups or families in the analysis of family data). Note that this interval estimate (for either θ or $\rho_{Intraclass}$) may not be very precise, if it results from a relatively small sample size, or if variance components are much more difficult (e.g., relatively low reliability in measurements) to be estimated precisely than means. Also note that it may occasionally happen that the lower limit of the confidence interval for either θ or $\rho_{Intraclass}$ is negative, but since this ratio (θ or $\rho_{Intraclass}$) normally should not be negative, the usual practice is to replace the "negative" lower limit with the best value to the zero lower bound – that is, simply, zero in this case.

The maximum likelihood estimator of intraclass correlation can be derived by using a theory of multivariate normal distribution (with the common mean and variance-covariance structure). Let $\mathbb{Y}_i = (Y_{i1}, \dots, Y_{in})$ represent measurements taken on the i-th groups ($i = 1, \cdots, k$), each consisting of n subjects, with a total size $N = k \times n$. Assume this n-variate follows a multivariate normal

$$\mathbb{Y}_i \sim \mathbb{N}_n(\mathbb{m}_i, \mathbb{V}_i) \quad , \quad for\ i = 1, \cdots, k$$

or equivalently, the (n-variate normal) probability density function is given by

$$f(\mathbb{Y}_i) = (2\pi)^{-n/2}|\mathbb{V}_i|^{-1/2} exp\left[-\frac{1}{2}(\mathbb{Y}_i - \mathbb{m}_i)'\mathbb{V}_i^{-1}(\mathbb{Y}_i - \mathbb{m}_i)\right]$$

, where the mean vector is $\mathbb{m}_i = (\mu, \dots, \mu)'_{1\times n}$ for a common mean μ across all groups, and the

variance-covariance matrix is $\mathbb{V}_i = \begin{pmatrix} \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \ddots & \vdots \\ \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix}_{n\times n}$ for the diagonal element $= \sigma^2$ (or a common variance across groups) and the off-diagonal element $= \rho\sigma^2$ (or a common covariance over groups), $|\mathbb{V}_i|$ denotes the determinant of $\mathbb{V}_i$ (i.e., the scaling factor in matrix algebra), and i is the index of groups for $i = 1, \cdots, n$. In a balanced design (the common correlation model), the estimate of intraclass correlation ρ can be obtained by using Pearson product-moment correlation (Donner & Koval, 1980a), and the explicit form of the estimator can be expressed by

$$\rho_{Pearson} = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l\neq j}^{n}\left(Y_{ij} - \bar{Y}\right)(Y_{il} - \bar{Y}) \Big/ [N(n-1)S_Y^2]$$

, where $\bar{Y}$ and S_Y^2 represent the common sample mean and variance, respectively, and can be computed across all observations N using the concept of intraclass correlation by Fisher (1925a).

And, by a large sample theory (asymptotic normality), the variance of the proposed estimator is

$$Var(\rho_{Pearson}) = 2(1-\rho)^2[1+(k-1)\rho]^2 \Big/ n(n-1)k \quad .$$

Note that when a balanced design is considered (i.e., $n_i = n$ for all $i = 1, \ldots, k$), this estimator $\rho_{Pearson}$ is also equivalent to the result of the maximum likelihood estimate (MLE) of intraclass correlation (i.e., the multivariate normal density is taken by the maximum

likelihood method).

On the other hand, for an unbalanced design, the asymptotic (large sample) variance of the proposed estimator $\rho_{Pearson}$ is given by

$$Var(\rho_{Pearson}) = {2N(1-\rho)^2}\Big/{N\sum_{i=1}^{k}[n_i(n_i-1)V_iW_i^{-2}] - \rho^2[\sum_{i=1}^{k} n_i(n_i-1)W_i^{-1}]}$$

, where the sampling weights are $V_i = 1 + (n_i - 1)\rho^2$ & $W_i = 1 + (n_i - 1)\rho$, total sample $N = \sum_{i=1}^{k} n_i$, and Pearson correlation is used as the estimate of ρ (Donner & Koval, 1982).

In addition, as for the estimators of μ and σ^2, the MLE solutions can be found by

$$\hat{\mu}_{MLE} = {\sum_{i=1}^{k}(n_i\overline{Y}_i/W_i)}\Big/{\sum_{i=1}^{k}(n_i/W_i)} = {\sum_{i=1}^{k} W_i^{-1}\left(\sum_{j=1}^{n_i} Y_{ij}\right)}\Big/{\sum_{i=1}^{k}(n_i/W_i)} \quad ,$$

and

$$\hat{\sigma}^2_{MLE} = [N(1-\rho)]^{-1}\left\{\left[\sum\nolimits_{i=1}^{k}\left(\frac{W_i-\rho}{W_i}\right)\sum\nolimits_{j=1}^{n_i}\left(Y_{ij}-\mu\right)^2\right] - \rho\left[\sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{n_i}\sum\nolimits_{l\neq j}^{n_i}\left(\frac{(Y_{ij}-\mu)(Y_{il}-\mu)}{W_i}\right)\right]\right\} \quad .$$

Hence, with $\hat{\mu}_{MLE}$ and $\hat{\sigma}^2_{MLE}$, the MLE of intraclass correlation in this case can be computed as

$$\hat{\rho}_{MLE}$$

$$= \left\{\sum_{i=1}^{k}(n_i-1)^{-1}\left[\sum_{j=1}^{n_i}\sum_{l\neq j}^{n_i}(Y_{ij}-\hat{\mu}_{MLE})(Y_{il}-\hat{\mu}_{MLE})\right]\right\} \Big/ \left\{\sum_{i=1}^{k}(n_i-1)^{-1}\left[\sum_{i=1}^{n_i}(Y_{il}-\hat{\mu}_{MLE})^2\right]\right\}.$$

Alternatively, it is equivalent to

$$\hat{\rho}_{MLE} = \hat{\sigma}_{MLE}^{-2} \times \left\{\sum_{i=1}^{k}(n_i-1)^{-1}\left[\sum_{j=1}^{n_i}\sum_{l\neq j}^{n_i}(Y_{ij}-\hat{\mu}_{MLE})(Y_{il}-\hat{\mu}_{MLE})\right]\right\} \quad .$$

Note that Karlin et al. (1981) derived this MLE of intraclass correlation in an unbalanced design (by using invariance property of MLEs; Casella & Berger, 2002, p.320, "If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$."), although Donner & Koval (1980b) used a different approach to solving the MLE of ρ by numerically "minimizing" the multivariate log-likelihood function (the logarithm of n-variate normal density) with a scaling factor of -2 :

$$-2logL(\rho|Y,\hat{\mu}_{MLE},\hat{\sigma}_{MLE}^2)$$

$$= N(1+log\hat{\sigma}_{MLE}^2+log2\pi)+(N-k)log(1-\rho)+\sum_{i=1}^{k}logW_i$$

, where this optimization method takes differentiation with respect to ρ to find the MLE.

2.3. *Hedges Approach*

Hedges used intraclass correlation to summarize the information of variance components in multilevel structure of 2-Level, 3-Level, and 4-Level hierarchical design (Hedges et al., 2012; Hedges & Hedberg, 2013). Further, intraclass correlation has been considered as an important tool/statistic to provide design effect parameters for statistical planning (power analysis) in experimental design and survey sampling (e.g., randomized controlled trials or large-scale experiments in education settings). In hierarchical linear models, intraclass correlation plays a key role in quantifying the amount of inherent clustering effects (i.e., within-cluster variation) in multilevel data. Look back at the development of ICC in hierarchical designs. The ICC was first introduced by Fisher (1925a), who created the oldest measure for within-group correlation and provided a significance testing procedure in experimental designs (such as RCTs and CRTs). Later on, Raudenbush (1997) built on hierarchical linear models in education to evaluate the clustering effect of multilevel data structure through ICC. Furthermore, Hedges used the meta-analytic framework to rethink the ICC by using design effect to improve multilevel designs in education and social research.

The Hedge's theoretical framework of intraclass correlation in multilevel design (like a cluster randomized trial, CRT) using hierarchical linear model (HLM) is:

In a two-level HLM, suppose that the variance components associated with fully unconditional model (no covariates at any level of the model). Let σ_1^2 and σ_2^2 be the variance components at Level 1 and Level 2, respectively, and s_1^2 and s_2^2 be the MLEs of σ_1^2 and σ_2^2, respectively. Let the variances of s_1^2 and s_2^2 be $v_1 \equiv Var(s_1^2)$ and $v_2 \equiv Var(s_2^2)$,

respectively. Without loss of generality, suppose that $\sigma_1^2 \approx s_1^2$ (note: in most large-scale studies by hierarchical design, the Level-1 variance component is usually known, i.e., σ_1^2 is a given constant and $v_1 \approx 0$, or can most likely be estimated precisely, i.e., $s_1^2 \approx \sigma_1^2$, since there are many Level-1 units that provide sufficient information for estimation; Hedges et al., 2012.) Let m denote the number of groups or clusters (Level-2 units) and n_i denote the number of Level-1 units in the i-th Level-2 unit of group or cluster. When the study is a balanced design analysis (i.e., $n_1 = n_2 = \cdots = n_m = n$), the intraclass correlation in the two-level HLM model is

$$\rho_{Intraclass} = {\sigma_2^2}\Big/{(\sigma_1^2 + \sigma_2^2)}$$

, and the intraclass correlation estimator (based on cluster random samples) is given by

$$\hat{\rho}_{Intraclass} = r_{Intraclass} = {s_2^2}\Big/{(s_1^2 + s_2^2)}$$

, then the asymptotic variance (based on large sample theory and delta method) is shown by

$$Var(\rho_{Intraclass}) = {[(1-\rho)^2 v_2]}\Big/{(\sigma_1^2 + \sigma_2^2)^2} = \sigma_T^{-4}[(1-\rho)^2 v_2]$$

, where the total variance component is $\sigma_T^2 = \sigma_1^2 + \sigma_2^2$, and the variance of s_2^2 is $v_2 \equiv Var(s_2^2)$ which is the variance (or squared standard error) estimate of the Level-2 variance

component. As for the estimate of the variance of $\hat{\rho}_{Intraclass}$ (i.e., sampling variability of the sample ICC), the large sample variance is given by

$$Var(\hat{\rho}_{Intraclass}) = {[(1-r)^2 v_2]}\Big/{(s_1^2 + s_2^2)^2} = s_T^{-4}[(1-r)^2 v_2]$$

, where r is the intraclass correlation estimate $\hat{\rho}_{Intraclass}$ (or $r_{Intraclass}$), and the variance of s_2^2 is defined by $v_2 = 2[\sigma_1^2 + n\sigma_2^2]^2/[n^2(m-1)] = 2\sigma_T^4[1+(n-1)\rho]^2/[n^2(m-1)]$, so that the estimate is $v_2 = 2s_T^4[1+(n-1)r]^2/[n^2(m-1)]$ for $s_T^2 = s_1^2 + s_2^2$. (Note: the assumption of $v_1 \approx 0$, or $\sigma_1^2 \approx s_1^2$, is imposed on the large-sample variance of intraclass correlation estimates.)

Fisher (1925a, p.220) derived a similar formula (large sample variance) for the intraclass correlation in a balance design (note: Fisher did not consider the assumption of $v_1 \approx 0$):

$$Var(\rho_{Intraclass}) = {\{2(1-\rho)^2[1+(n-1)\rho]^2\}}\Big/{[n(n-1)(m-1)]} \quad .$$

Donner & Koval (1980b) showed the large sample variance of intraclass correlation in an unbalanced design (note: Donner & Koval did not consider the assumption of $v_1 \approx 0$) as

$$Var(\rho_{Intraclass}) = \frac{[2N(1-\rho)^2]}{\left\{N\sum_{i=1}^{m}\frac{n_i(n_i-1)[1+(n_i-1)\rho^2]}{[1+(n_i-1)\rho]^2} - \rho^2\left[\sum_{i=1}^{k}\frac{n_i(n_i-1)}{1+(n_i-1)\rho}\right]^2\right\}} \quad .$$

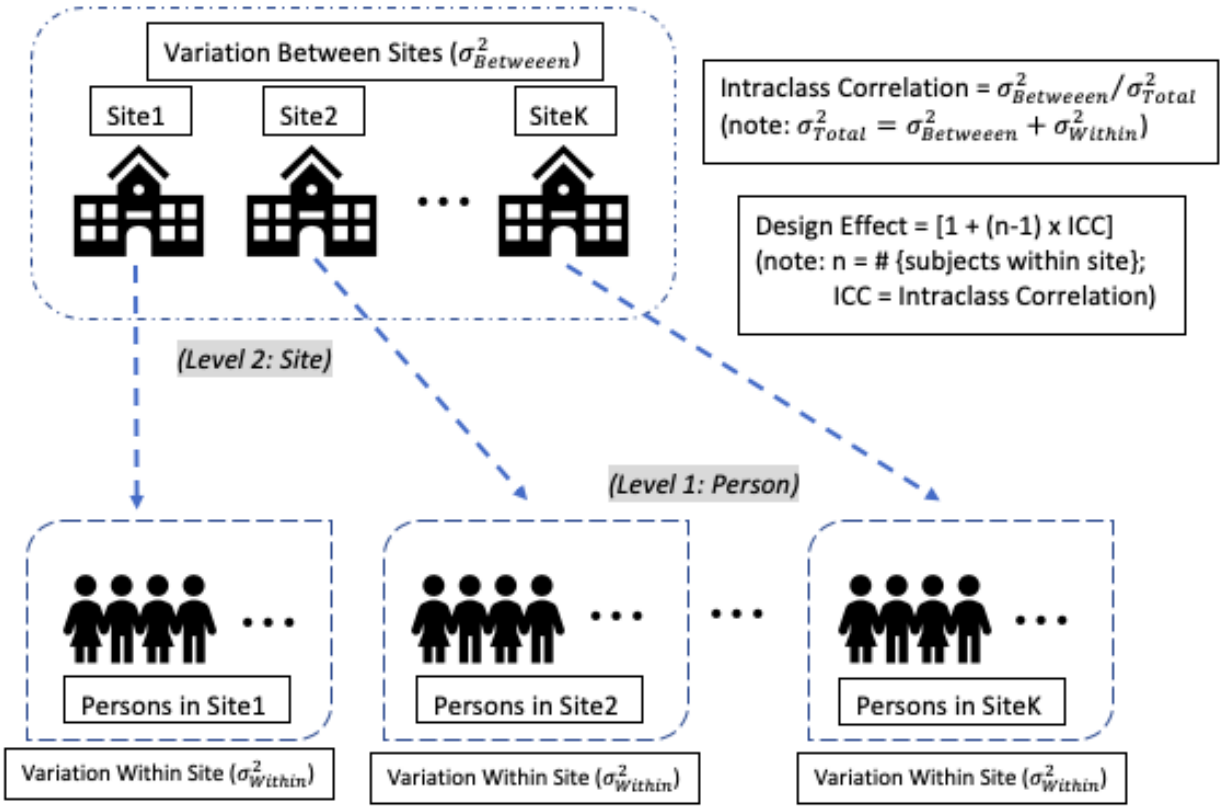
In a cluster (or group) randomized design, researchers often operate interventions or assign treatments at a group level (say Level-2 such as classrooms, schools, or sites) rather than at an individual level (say Level-1 for individual subjects like students) for some practical reasons that it is sometimes too expensive (or even not feasible) to work on interventions to each subject but rather than deal with an entire intact group (e.g., a whole community, school, worksite, or family). Therefore, cluster-randomized trials (or group-randomized experiments) recently have become more and more important and popular in educational and social research studies for effectively and economically evaluating educational and social interventions (Donner et al., 1981; Hauck et al., 1991; Hedges & Hedberg, 2007; Klar & Donner, 2015). For example, a research investigator could save money (or increase the effectiveness of cost) by using group interventions, e.g., CRTs, instead of individual ones like RCTs (Tachibana et al., 2018). Also note that researchers find CRTs are more suitable than RCTs for the construction of economically-efficient and economically-productive samples that have the desired statistical properties (Connelly, 2003).

In a theoretical framework of cluster sampling experiments (i.e., cluster-randomized trials), suppose a sample of subjects are collected from m clusters (or organizational units such as classrooms, schools, or district sites) of a group size n which are assigned to an intervention (or a treatment group) with randomization. In this cluster sampling design, the classical simple random sampling approach doesn't hold (i.e., all $m \times n$ individual samples are not independent to each other, but rather are highly dependent on the cluster to whom a subject, he or she, belongs or is assigned; Lohr, 1999, Chapter 2 Simple Probability Samples & Chapter 5 Cluster Sampling with Equal Probability). Therefore, the sampling distribution of a statistic using cluster samples needs to take into account both between-group correlation and

within-group variation at the same time in analysis. Suppose that in this cluster sampling structure, the total variance σ_T^2 consists of a within-cluster variance σ_W^2 and a between-cluster variance σ_B^2, i.e., $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$. Then, comparing with the formula of the population mean variance estimator for a simple random sample $\sigma_T^2/_{m \times n}$, the population average variance for an individual sample (from m clusters with size n) is shown as $[1 + (n-1)\rho]\sigma_T^2/_{m \times n}$, where the intraclass (or sometimes called intra-cluster) correlation coefficient is $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2) = 1 - \sigma_W^2/(\sigma_B^2 + \sigma_W^2)$ which provides a statistical measure of homogeneity within the clusters (i.e., if the clusters are perfectly homogeneous, then $\sigma_W^2 = 0$ and $\rho = 1$), and the design effect (DE) or variance inflation factor (VIF) is defined as $[1 + (n-1)\rho]$ (Donner et al., 1981; Lohr, 1999, pp.138-140). Note that clustering has more variation than simple random sampling by a factor of DE (or VIF>1) due to the major part of cluster-to-cluster variability plus the minor portion of within-cluster variance (i.e., samples in different clusters often vary more than those samples in the same cluster). See Figure 2.4 as an example of 2-level hierarchical structure with regard to intraclass correlation and design effect.

In experimental design, statistical planning for sample size determination and power calculation is critical for researchers to better produce evidence-based conclusions by rigorously detecting true effects at the desired level of significance. Traditionally, the experimental planning approach of sample and power computation considers the classical assumption of simple random samples. Therefore, power analysis for cluster sampling design or group randomized experiments need to use intraclass correlation coefficient along with non-centrality parameters (of F-distribution) to account for variability in multilevel design (e.g., between-group and within-group variations) (Cohen, 1992; Hedges & Hedberg, 2007, 2013; Raudenbush, 1997; Rutterford et al., 2015).

Figure 2.4 Intraclass Correlation & Design Effect in 2-Level Hierarchical Linear Model



Note. Each level has its own variation, where variation between sites is sigma-square of between, and variation within site is sigma-square of within, and the total variation is the sum of these two, i.e., "sigma-square of between" + "sigma-square of within".

In a two-level hierarchical design structure (i.e., individuals are at the level 1, and groups or clusters at the level 2), the unconditional model (involving with no covariates) is written by

$$\begin{aligned} (Level\ 1) \qquad & Y_{ij} = \beta_{0i} + \varepsilon_{ij} \\ (Level\ 2) \qquad & \beta_{0i} = \pi_{00} + \varsigma_i \\ (Overall) \qquad & Y_{ij} = \pi_{00} + \varsigma_i + \varepsilon_{ij} \end{aligned}$$

, where Y_{ij} represents an outcome for the j-th individual subject (at the level 1) in the i-th cluster group (at the level 2), π_{00} is a grand mean outcome $\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}/(m \times n)$, ε_{ij} is a random error term at the level 1 (i.e., $\varepsilon_{ij} \sim NID(0, \sigma_W^2)$) corresponding to the j-th person in the i-th group, ς_i is a random effect (i.e., $\varsigma_i \sim NID(0, \sigma_B^2)$) associated with the i-th cluster (or a random error term at the level 2), the within-group (between-person) variance component is given by $Var(\varepsilon_{ij}) = \sigma_W^2$, the between-group variance component is given by $Var(\varsigma_i) = \sigma_B^2$, and the random error terms at the level 1 and level 2 are not correlated (i.e., $Cov(\{\varsigma_i\}, \{\varepsilon_{ij}\}) = 0$).

The (unconditional) intraclass correlation coefficient associated with the unconditional model is

$$\rho = {\sigma_B^2}\Big/{(\sigma_W^2 + \sigma_B^2)} = {\sigma_B^2}\Big/{\sigma_T^2}$$

, where the (unconditional) total variance is defined as $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$, σ_W^2 and σ_B^2 represents the error variances corresponding to the within- and between-group random variation,

respectively.

In a hierarchical design (such as cluster-randomized experiment) involving statistical adjustment by covariate(s), the (covariate-adjusted, or conditional) intraclass correlation is defined by

$$\rho_A = {\sigma_{AB}^2}\Big/{(\sigma_{AW}^2 + \sigma_{AB}^2)} = {\sigma_{AB}^2}\Big/{\sigma_{AT}^2}$$

, where the (covariate-adjusted) total variance is defined as $\sigma_{AT}^2 = \sigma_{AW}^2 + \sigma_{AB}^2$, σ_{AW}^2 and σ_{AB}^2 represents the "adjusted" residual variances (or random-effect variance components adjusted by covariates) corresponding to the within- and between-group random variation, respectively.

In order to evaluate the relative efficiency between unconditional and conditional hierarchical models, Hedges & Hedberg (2007) proposed two statistical auxiliary quantities

$$\eta_B^2 = {\sigma_{AB}^2}\Big/{\sigma_B^2}$$

and

$$\eta_W^2 = {\sigma_{AW}^2}\Big/{\sigma_W^2}$$

, where η_B^2 indicates the proportion of between-group variance remaining, and η_W^2 indicates the proportion of within-group variance remaining. Note that these two measures, along with $R_B^2 = 1 - \eta_B^2$ and $R_W^2 = 1 - \eta_W^2$, are useful to provide information of statistical variation for power and sample size computations, where R_B^2 and R_W^2 are defined as the proportion of

between-group and within-group variance explained by covariate(s) in hierarchical design, respectively.

2.4. *Raykov Approach*

In classical test theory (CTT), a given test score (X) consists of two parts – the true score (T) and the measurement error (E) (Raykov & Marcoulides, 2011, pp.117-118); hence, the relationship can be mathematically described as $X = T + E$, where the true score variance is $Var(T) \equiv \sigma_T^2$, the error variance is $Var(E) \equiv \sigma_E^2$, plus the true score and error score are assumed to be mutually independent, i.e., $\rho_{T,E} \equiv Corr(T, E) = 0$. According to the CTT equation, reliability coefficient (ρ_X) is the ratio of the true score variance to observed score variance, and can be expressed as $\rho_X \equiv \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2$, which is equivalent to a similar idea of the R^2 index in regression analysis when predicting true score from observed score. Moreover, it is interesting to note that the standard error of measurement (SEM) is $\sigma_E = \sigma_X\sqrt{1-\rho_X}$ (Raykov & Marcoulides, 2011, pp.137-145). Thereby, within the CTT framework, it appears a strong connection between reliability coefficient and intraclass correlation coefficient in terms of statistical concepts and mathematical definitions (i.e., both share common ground to utilize variance accounted for).

By the latent variable modeling (LVM) approach (Bartholomew, 1987), Raykov & Penev (2010) showed a procedure to evaluate reliability coefficients (such as point and interval estimators) in 2-level HLM unconditional and conditional models, and further derived

standard error (SE) estimates for reliability coefficients with logit transformation (i.e., $\hat{k} \equiv logit(\rho_X) = ln[\rho_X/(1-\rho_X)]$) via Taylor series expansion method (aka Delta method) as $SE(\hat{k}) = SE[logit(\rho_X)] = SE(\rho_X)/[\rho_X(1-\rho_X)]$, which can lead to an $100(1-\alpha)\%$ large sample confidence interval using the standard normal Z distribution by $\hat{k} \pm Z_{(1-\alpha)} \times SE(\hat{k})$ where $\alpha \in (0,1)$, and $\hat{k}$ is a logit-transformed reliability coefficient, and $SE(\hat{k})$ is the error measurement.

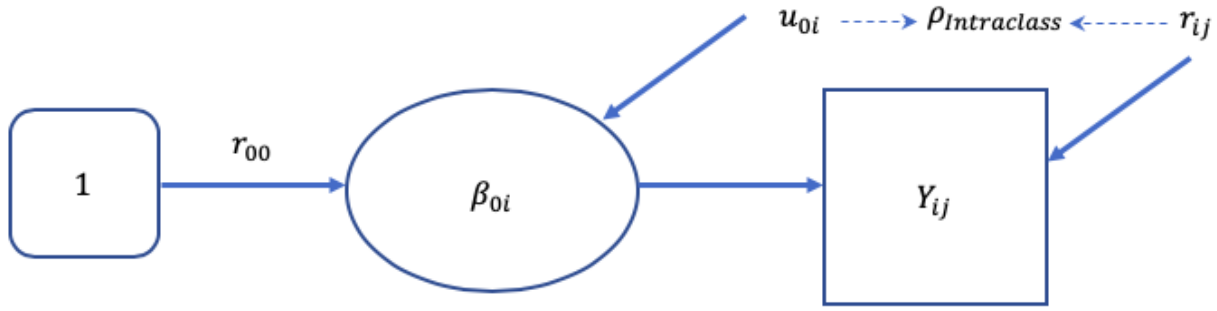
As for intraclass correlation coefficients (ICC) in hierarchical designs (e.g., two-level models) within the LVM framework (Bartholomew et al., 2011), Raykov (2011) used the restrictive maximum likelihood (REML) estimators to find ICC in the two-level HLM structure (aka factorial random-effect ANOVA):

$$\begin{array}{ll} (Level\ 1\ Unit) & Y_{ij} = \beta_{0i} + r_{ij} \\ (Level\ 2\ Unit) & \beta_{0i} = r_{00} + u_{0i} \\ (Overall\ Unit) & Y_{ij} = r_{00} + u_{0i} + r_{ij} \end{array}$$

, where Y_{ij} represents a response outcome score for the j-th individual subject (at the level 1; $j = 1, \dots, n_i$) in the i-th cluster group (at the level 2; $i = 1, \dots, k$), r_{00} is the grand mean, r_{ij} is a random error term at the level 1 and assumed to be normally distributed with mean 0 and within-group variance σ_W^2 (i.e., $r_{ij} \sim NID(0, \sigma_W^2)$) corresponding to the j-th person in the i-th group, u_{0i} is a random effect and assumed to be normally distributed with mean 0 and between-group variance σ_B^2 (i.e., $u_i \sim NID(0, \sigma_B^2)$) associated with the i-th cluster's random deviation term at the level 2, and the random error terms at the level 1 and level 2 are

supposed to be mutually uncorrelated (i.e., $Corr(\{u_{0i}\},\{r_{ij}\}) = 0$). In this LVM framework, the ICC is defined as the ratio of between-group variance to observed total variance $\rho_{Intraclass} = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$, where the within-group variance is $Var(r_{ij}) = \sigma_W^2$, and the between-group variance is $Var(u_{0i}) = \sigma_B^2$. The visualization of this LVM modeling approach using a path diagram is shown in Figure 2.5.

Figure 2.5 Latent Variable Model for Estimation of Intraclass Correlation in 2-Level Design



Note. The path diagram is inspired by the visualization of 2-level random coefficient models in the book of statistical multilevel modeling (Muthén & Muthén, 2012, Chapters 9 & 10).

With the invariance property of MLE for the variance estimates in LVM, the ICC is given by $\hat{\rho}_{Intraclass} = \hat{\sigma}_B^2/(\hat{\sigma}_B^2 + \hat{\sigma}_W^2)$, where $\hat{\sigma}_B^2$ and $\hat{\sigma}_W^2$ are the between- and within-group variation estimates, respectively, obtained by the REML method in the two-level LVM model. Note that according to Casella & Berger (2002, p.320), the invariance property of MLEs is stated as follows: If $\hat{\theta}$ is the MLE of θ, then for any one-to-one function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$. As for hypothesis testing, the test statistic for intraclass correlation is

given by a standard normal Z distribution for the pivotal quantity

$$Z_0 = (\hat{\rho}_{Intraclass} - \rho_0)/SE(\hat{\rho}_{Intraclass})$$

is used to test the simple hypotheses $H_0: \rho_{Intraclass} = \rho_0$ vs $H_a: \rho_{Intraclass} \neq \rho_0$ (i.e., a two-tailed test at the significance level of α), or $H_0: \rho_{Intraclass} = \rho_0$ vs $H_a: \rho_{Intraclass} \geq \rho_0$ (i.e., a one-tailed test at the α level), albeit this analytic strategy may only work for the large sample case, plus the lower bound of an interval estimation for $\rho_{Intraclass}$ by this method may reach out below zero (i.e., an out-of-bounds value from the valid domain of ICC $[0,1]$).

The LVM procedure can also be extended and used to evaluate ICC at two-level designs with discrete response variables (Raykov & Marcoulides, 2015a). Suppose the same two-level LVM setting above, but assume that the observed outcome score Y_{ij} is recorded on a categorical scale (i.e., a discrete variable for the j-th unit at the level-1 of individual subject ($j = 1, \ldots, n_i$) in the i-th unit at the level-2 of cluster group ($i = 1, \ldots, k$)). In this situation with categorical responses, the traditional approach of ICC estimation (which presumes the outcome is continuous) needs to be modified by the following modification procedure via the LVM framework (Raykov & Marcoulides, 2011, Chapter 10 Introduction to Item Response Theory). First, consider the underlying latent structure "behind" an observed response variable (k possible cateogries) as

$$Y_{ij} = \begin{cases} 1, & if \quad \tau_0 < Y_{ij}^* \leq \tau_1 \\ 2, & if \quad \tau_1 < Y_{ij}^* \leq \tau_2 \\ & \vdots \\ k, & if \quad \tau_{k-1} < Y_{ij}^* \leq \tau_k \end{cases}$$

, where Y_{ij}^{*} ($\forall\ i, j$) plays an important role of a continuous latent variable (i.e., $Y_{ij}^{*} \in \mathbb{R}$), which is not only linked with the observed measure Y_{ij} by a one-to-one linear transformation from one domain (latent space) to another (real space), but also used to assign a specific categorical value through the given thresholds points from $\{\tau_0, \tau_1, \ldots, \tau_{k-1}, \tau_k\}$ (note: each threshold or cut-off point is a real number, and it holds that $-\infty = \tau_0 < \tau_1 < \cdots < \tau_{k-1} < \tau_k = \infty$) (Raykov & Marcoulides, 2015b).

Given this underlying latent structure above, the ICC estimator for a binary outcome (a special case of categorical outcome variables; Raudenbush & Bryk, 2002, p.334) can be derived by

$$\rho_{Intraclass} = {\sigma_B^2}\Big/{(\sigma_B^2 + \pi^2/3)}$$

, where σ_B^2 is the between-group variation, and π is a mathematical constant $3.14159\cdots$ (note: the standard logistic distribution, with location $= 0$ and shape $= 1$, has a variance of $\pi^2/3$). Also notice that this ICC estimator for the dichotomous outcome case (say, 0 or 1) makes a strong assumption that the within-group variance σ_W^2 is held as a constant of $\pi^2/3$ over all individual subjects and groups; yet, this assumption of "a given constant within-group variance" is often not met in real-life data, and so the modified analytic strategies are needed for building non-constant within-group variances (which are data-driven and more flexible for a real world situation) into hierarchical generalized linear models (HGLM).

Furthermore, the standard error of the ICC above (for the binary response case) can be approximately derived via Delta method (Raykov & Marcoulides, 2004; Hedges et al., 2012),

which is given by

$$SE(\hat{\rho}_{Intraclass}) \approx \sqrt{(1-\hat{\rho}_{Intraclass})^2 \Big/ \hat{\sigma}_T^4 \times Var(\hat{\sigma}_B^2)} = \sqrt{(1-\hat{\rho}_{Intraclass})^2 \Big/ \hat{\sigma}_T^4} \times SE(\hat{\sigma}_B^2)$$

, where $\hat{\rho}_{Intraclass}$ is the ICC estimate, the total variance estimate is $\hat{\sigma}_T^2 = \hat{\sigma}_B^2 + \pi^2/3$ (assuming the within-group variance is a constant of $\pi^2/3$), and $\hat{\sigma}_B^2$ is the between-group variance estimate.

CHAPTER 3

LITERATURE IN REHABILITATION COUNSELING

This chapter presents literature of EBP in rehabilitation counseling using the RSA-911.

The state vocational rehabilitation (VR) agencies collect and report summary data in a federally mandated format called the Rehabilitation Services Administration (RSA) Case Service Report, aka the RSA-911 (Schwanke & Smith, 2004). The RSA-911 provides researchers in the field of rehabilitation counseling an open playground and additional resource for deep learning and data mining. Not only does the RSA-911 allow multi-faceted explorations of complex issues about people with disabilities in VR, but rehabilitation researchers can also probe extensively into big data to examines the hidden components or latent factors contributed to successful VR outcomes (Pi & Thielsen, 2011). Moreover, rehabilitation practitioners and scholars can take full advantage of the RSA-911 data to develop evidence-based practices, particularly for individual-level and employment-focused interventions, effective strategies, as well as best practices to promote independent living and positive outcomes for individuals with disabilities (Fleming et al., 2013).

With EBP as a cookbook approach to rehabilitation counseling (Kosciulek, 2010), it provides the fundamental framework for rehabilitation counseling practitioners that incorporates the available scientific evidence with the expertise of clinical judgement skills to make best decisions about interventions, services, or treatments for people with disabilities. In this manner, EBP guidelines also suggests rehabilitation counselors to identify relevant literature and systematic research, to assess different available information resources such as the RSA-911 data, and to constitute “best available evidence” on rehabilitation services for people

with disabilities. So, with the data-driven framework using information on RSA-911, which research method or statistical approach can provide insights to work best for whom (target populations), how (intervention or treatment programs), and under what condition (rehabilitation support or other types of services)? This literature review surveys recent academic knowledge on those key questions and provides a firm foundation to this study.

The following is a summary of literature review of statistical methods using the RSA-911.

3.1. *Multilevel Analysis*

Hierarchical data structures are often seen in educational and social research studies. For example, in rehabilitation counseling, VR clients are grouped into organizational buildings and structures or field offices, which are nested into different local districts, and local districts can be nested into states or regions, and so on. So, it is important to take into account all these hierarchical data structures and topological data relationships by using multilevel analysis (hierarchical linear models). Note that conventional regression models often underperform statistical estimation and inference (e.g., inflation of standard errors, and relative bias in ICC) in hierarchically structured data due to non-normal residuals resulted from the interrelation between subjects (which somewhat leads to violation of the important assumptions of independence, homogeneity and normality) (Maas & Hox, 2004; Raudenbush & Bryk, 2002).

Chan and his colleagues (2014) used RSA-911 data in FY 2005 (before the economic recession) and FY 2009 (after the economic recession) to study the impact of the contextual

factor of state unemployment rate, and its impact on the employment opportunities and outcomes in VR. By the (2-level) hierarchical (generalized) linear modeling approach, they found state unemployment rate (the contextual variable) was having a significant moderation effect on the relationship between personal factors (demographic and disability variables) and competitive employment.

Alsaman & Lee (2017) examine the relationships between contextual factors, individual factors, and employment outcomes of transition youth with disabilities in VR using the RSA-911 in FY 2013 by the 2-level hierarchical generalized linear modeling. They found state unemployment rates were having the indirect interaction impacts on the relationships between individual characteristics, rehabilitation services, and successful employment. For example, the state unemployment rate increased, the disparity in successful VR closure decreased across some types of disabilities such as intellectual disabilities, TBI, or youth with autism and other communicative disabilities (in comparison to the reference group of physical disabilities).

Pi (2006) constructed the 2-level hierarchical structure model with the micro- and macro-level factors related to VR outcomes using RSA-911 in FY 2002. Results showed the micro-level variables (i.e., age, education, minority, SSI/DI, disability significance, services – rehabilitation technology, job placement assistance, on-the-job-support, and diagnosis & treatment) were more related to rehabilitation outcomes than the macro-level variables (i.e., counselors who met CSPD requirements, proportion of clients with significant disabilities, unemployment rate, proportion of minority population). Note: CSPD=Comprehensive System of Personnel Development.

3.2. *Structural Equation Model*

The structural equation modeling (SEM) with latent constructs (unobserved factors) and manifest variables (truth realizations) is one type of structural causal modeling (statistical models for causation) that is built (through a path diagram for visualization) to identify the underlying factor structure explaining the direct and/or indirect effects of latent constructs and their inter-relationships on outcomes of interest (Raykov & Marcoulides, 2006). In the VR context, SEM can be used to understand complex theoretical models (or EBP) and to find important predictive associations (using latent factor analysis) among individual characteristics, rehabilitation services, and employment outcomes (Austin & Lee, 2014).

Kosciulek & Merz (2001) conducted structural analysis of consumer-directed theory of empowerment for consumers with disabilities in the community rehabilitation program.

Chan et al. (2007) provided an overview of the basic concepts and applications of SEM (e.g., confirmatory factor analysis) in counseling, psychology, and rehabilitation research.

Austin & Lee (2014) built a structural equation model of VR services (consisting of job-related and person-related factors) via RSA-911 in FY 2009, to study predictors of employment outcomes in VR for people with intellectual and co-occurring psychiatric disabilities. The study found job-related services such as job placement, job search, job readiness, and on-the-job support, were to significantly predict competitive employment outcomes.

3.3. *Classification Tree Model*

The tree model is a data-mining technique via the classification method of CHAID – *Chi-squared Automatic Interaction Detection* algorithm – to explore hidden relationships and predictive information in a large database (Tan et al., 2005). In the classification tree procedure, the tree-based model is designed to classify all subjects into homogeneous subgroups by their attributes. Additionally, the "exhaustive" classification procedure is quite useful to uncover the complex multivariate system like the VR process by providing useful "grouping" information.

Rosenthal et al. (2007) used the data mining approach via RSA-911 data in FY 2001 to examine factors (i.e., services) affecting outcomes in the VR process for individuals suffering psychiatric disabilities. Results showed receiving job placement services was found to be the most important variable and had a positive effect for the target population in VR.

Schoen (2010), and Schoen & Leahy (2012) conducted an examination of demographics, services, and employment outcomes for people with spinal cord injury in VR between FY 2004 and FY 2008 by data mining models via RSA-911 data. Findings suggested the most significant predictors of employment were level of education attained, cost of purchased services, days from application to closure, rehabilitation technology, job placement assistance, and job supports.

Lee and his colleagues (2012), and Lee (2014) tried to discover the VR evidence-based best practices using a data mining approach of decision (or classification) tree models through the RSA-911 data in FY 2011 and FY 2013, respectively, to study the inter-relationships of VR measurements between services delivery, personal backgrounds and rehabilitation outcomes for

people with disabilities in State of Michigan.

3.4. *Other Methods such as Social Network Analysis and Spatial Analysis*

Spatial analysis is a type of geographical/locational analysis (statistics) which seeks to explain patterns of human behavior (e.g., rehabilitation outcomes) and its spatial expression (residential areas). The geostatistical model can predict the spatial patterns (using geographical information) in the complex networks or systems (like RSA-911) for spatial decision-making support and solving geographic issues in planning and policy development (Mayhew, 2015).

Sink et al. (2014) developed location theory in VR to study effectiveness of service delivery and consumption for persons with disabilities using the geographic information system (GIS) and data from West Virginia Division of Rehabilitation Services (including RSA-911 and Census data). The findings supported the value of public VR field office or facility location and its effectiveness and efficiency for people with disabilities to achieve or maintain employment.

Social network analysis is the process of investigating social structures through the use of graph and network theory. The social networking model characterizes individual links or ties (relationships or interactions) within a networked structure (such as the VR system). One key feature of this social network analysis is visual representation (via sociograms) which provides pivotal information about attributes within a network (e.g., positive or negative relationships between services and outcomes in the VR network data) (Schneider, 2018).

Ditchman et al. (2018) applied social network analysis, via the RSA-911 data in FY 2009, to examine service patterns and their relationships with employment outcomes for

transition-age individuals with autism spectrum disorder (ASD). By social network analysis, six core VR services were found positively linked with a better employment outcome, including: assessment, counseling & guidance, job placement, job search, job support and transportation.

3.5 *Justification for Covariates Used in Multilevel Analysis*

The Rehabilitation Act of 1973 (and its Amendments of 1986, 1992) was legislated with the goal of providing individuals with disabilities with equal opportunities to achieve employment, independent living, and self-sufficient as the general population without disabilities. Under the law, state VR programs are to help people with disabilities to obtain or maintain employment through rehabilitation services, which may include but not limited to assessment, vocational rehabilitation counseling & career guidance, educational training (e.g., colleges or universities), job coaching, job placement services, on-the-job support training, transportation and miscellaneous services (see Appendix A for the definitions of VR variables used in the study; Rehabilitation Services Administration Policy Directive, 2013). Many research studies have been conducted to examine the relationships between various factors (i.e., individual characteristics, VR services, VR counselors, and environmental factors) and rehabilitation outcomes. Based on a systematic review on VR outcomes in relation to VR factors, previous rehabilitation studies confirms the VR variables of interest in this study (including individual characteristics, employment backgrounds, rehabilitation services) are all supported by the VR foundations with the significance of associations with successful employment outcomes for people with disabilities (Alsaman & Lee, 2016; Bolton et al., 2000; Chan et al., 2014; Dutta et al., 2008; Moore et al., 2000, 2001, 2002a, 2002b, 2004).

CHAPTER 4

METHODS AND RESEARCH QUESTIONS

In this chapter, it provides analytic strategies of experimental planning for cluster (or group) randomized design structure with respect to power & sample size calculations using intraclass correlation coefficient (or ICC) via hierarchical linear model (HLM) and hierarchical generalized linear model (HGLM). By the bootstrapping simulations (Givens & Hoeting, 2012; Rizzo, 2007), the methods are proposed to evaluate statistical performance of ICC, in terms of relative bias, estimation error, and inference on parameter, via HLM & HGLM using the real data set of RSA-911 from the U.S. Department of Education and Labor. In the RSA-911 data of this study, the target population focuses on those people with disabilities who had been served in Michigan in fiscal year (FY) 2015. In addition, the two-stage sampling approach is used to generate the simulated data sets with the cluster-randomized design structure, where individual subject (person with disability) is for Level 1 and structure (rehabilitation office) is for Level 2.

4.1 *Research Methods*

Three proposed ICC estimation methods are shown for different statistical settings and experimental design purposes using multilevel models:

Method 1 – the ICC estimator (via Pearson correlation & F of ANOVA) given by a balance design (equal size of n individual subjects across k groups) is shown in Equations 1

and 2:

$$\rho_{Intraclass} = \left\{\left[k\sum_{j=1}^{n}\left(\bar{x}_j - \bar{X}\right)^2\right] - nS^2\right\} \Big/ nS^2(k-1) \quad (1)$$

and

$$\rho_{Intraclass} = (MSA - MSW) \Big/ [MSA + (n-1)MSW] \quad (2)$$

, where n is group sample size, j is the index of samples, the among-group mean is $\bar{x}_j$ (from the j-th sample over all k groups), the common mean is $\bar{X}$, the common variance is S^2, MSA and MSW are Mean Squares Among and Mean Squares Within from ANOVA, respectively.

Method 2 – the ICC estimator (via Pearson correlation & F of ANOVA) given by an unbalance design (unequal size of n_i individual subjects across k groups, for $i = 1, \cdots, n$) is shown in Equation 3:

$$\rho_{Intraclass} = (MSA - MSW) \Big/ [MSA + (n_0 - 1)MSW] \quad (3)$$

, where $n_0 = [N - \sum_{i=1}^{k} n_i^2/N]/(k-1)$ is the "adjusted" group sample size for ICC estimation, and N is the total sample size (i.e., $N = \sum_{i=1}^{k} n_i$), MSA and MSW are Mean Squares Among and Mean Squares Within from ANOVA, respectively.

Note that Pearson correlation estimate requires numerical approximation of -2 log likelihood.

Method 3 – find auxiliary information (based on the ICC estimate from Method 1 or 2) for experimental planning in designs (design effect and minimum detectable effect size with respect to desired power & required sample size):

(a) Design effect (DE), or variance inflation factor (VIF), is defined in Equation 4 as

$$DE = VIF = [1 + (n - 1)\rho] \tag{4}$$

, where the intraclass correlation coefficient is $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$, or alternatively $\rho = 1 - \sigma_W^2/(\sigma_B^2 + \sigma_W^2)$, which provides a statistical measure of homogeneity within the clusters, n is group sample size for a balanced design case (or, alternatively, n_0 can be substituted for n in an unbalanced design).

(b) The unconditional intraclass correlation coefficient is shown in Equation 5:

$$\rho = {\sigma_B^2}\Big/{(\sigma_W^2 + \sigma_B^2)} = {\sigma_B^2}\Big/{\sigma_T^2} \tag{5}$$

, where the unconditional total variance is $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$, σ_W^2 and σ_B^2 represent the error variances corresponding to the within- and between-group variation, respectively.

In a hierarchical design, such as cluster-randomized experiment, involving statistical adjustment by covariate(s), the conditional (or covariate-adjusted) intraclass correlation is described in Equation 6:

$$\rho_A = {\sigma_{AB}^2}\Big/{(\sigma_{AW}^2 + \sigma_{AB}^2)} = {\sigma_{AB}^2}\Big/{\sigma_{AT}^2} \quad (6)$$

, where the covariate-adjusted total variance is $\sigma_{AT}^2 = \sigma_{AW}^2 + \sigma_{AB}^2$, σ_{AW}^2 and σ_{AB}^2 represent the random-effect variance components, adjusted by covariates, corresponding to the within- and between-group random variation, respectively.

(c) Four proposed statistical auxiliary quantities for evaluating the relative efficiency between unconditional and conditional hierarchical models, are shown as follows. The first two for measuring "variance remaining" are described in Equations 7 and 8:

$$\eta_B^2 = {\sigma_{AB}^2}\Big/{\sigma_B^2} \quad (7)$$

and

$$\eta_W^2 = {\sigma_{AW}^2}\Big/{\sigma_W^2} \quad (8)$$

, where η_B^2 indicates the proportion of between-group variance remaining, and η_W^2 indicates the proportion of within-group variance remaining.

The other two supplementary measures for variance explained by covariates (also serving the complementary side of measurements in Equations 7 and 8) are described below in Equations 9 and 10:

$$R_B^2 = 1 - \eta_B^2 \quad (9)$$

and

$$R_W^2 = 1 - \eta_W^2 \quad (10)$$

, where R_B^2 and R_W^2 are defined as the proportion of between-group and within-group variance explained by covariate(s) in hierarchical design, respectively.

4.2 *Proposed Models*

Four hierarchical modeling structures (Models 1-4 as shown below) are considered in the study to test the proposed methods. And three "breaking variables" – significance of disability (yes/no), type of disability (nominal measure with 10 categories), and previous work experience (yes/no) – are included in all four models for separate (subgroup-specific) analyses by breaking down the whole sample into different subsets based on the shared characteristics.

Model 1 – Unconditional Model (no covariate-adjusted)

Model 2 – Conditional Model (covariate-adjusted by Covariate Set 1)

Covariate Set 1 consisting of demographic characteristics includes: (a) gender (male or female); (b) minority (yes or no); (c) age (continuous measure); (d) SES by social security and/or insurance benefits (yes or no); (e) educational background (ordinal measure).

Model 3 – Conditional Model (covariate-adjusted by Covariate Set 2)

Covariate Set 2 consisting of VR service variables includes: (a) job placement assistance (binary; received or not received); (b) on-the-job supports (binary; received or not received); and (c) rehabilitation technology (binary; received or not received).

Model 4 – Conditional Model (covariate-adjusted by Covariate Set 3)

Covariate Set 3 combines both Covariate Sets 1 and 2 altogether into one set.

There are two different VR outcomes used in simulation analyses – (1) competitive employment outcome (yes/no); and (2) weekly earnings (a continuous measure) = rehabilitation outcome (a dichotomous 0 or 1 measure) X weekly income (a continuous measure), where the weekly earnings can also be deemed as an indicator of quality of employment outcomes achieved at exit in the VR (Chan et al., 2016; O'Neill et al., 2015).

Note. The total number of all combinations of analyses (4 Models X 2 Outcomes) = 8.

4.3 *Research Questions*

Our proposed methods are used to address the following research questions:

In order to evaluate the simulation results, descriptive statistics of ICC are provided to answer Research Question 1 (RQ1) & Research Question 2 (RQ2) below. In addition, statistical performance (precision and accuracy) of ICC under the designated conditions using randomized cluster samples is examined by statistical bias (or average bias) and its error variance (or mean square error) to answer Research Question 3 (RQ3) below. Furthermore, the usable samples in

the "whole" data set of RSA-911 are used as a collection of the true parameters of ICC in RQ1; then, in the bootstrapping computations (Ross, 2013), the full data set of RSA-911 is resampled 100 times (number of bootstrapping repetitions=100) under the given sampling conditions for ICC estimation using the "bootstrap" samples in RQ2. At the end, by comparing the differences in ICC estimates between RQ1 and RQ2, it shows which one of estimation methods, designated models, and sampling conditions, can provide the best results of statistical performance of ICC estimation and inference at multilevel design with randomized cluster samples (RQ3).

Research Question 1 (RQ1): What are the "true" intraclass correlation values (ICC estimate, standard error, p-value and 95% confidence limits) in the usable samples given by the breaking variables for subset analysis (Models 1-4)? How are ICC estimates distributed in Models 1-4? What are the differences in the ICC estimates among Models 1-4?

Research Question 2 (RQ2): Given the designated cluster randomized structure (i.e., the number of groups = 5, 15, 25; the number of subjects = 50, 100, 150), what are the intraclass correlation estimates (ICC estimate, standard error, and p-value) using the "bootstrap" samples (the number of bootstrap repetition=100) given by breaking variables under Models 1-4?

Research Question 3 (RQ3): Comparing the results between Research Question 1 (population model) and Research Question 2 (bootstrap subsample model), which modeling structure (Models 1-4) can provide the best statistical properties of ICC estimation and inference, in terms of statistical bias (mean difference in ICC estimates between RQ1 and RQ2), mean squared error or mean squared deviation (average squared difference in ICC estimates between RQ1 and RQ2), and parameter coverage rate (proportion of true parameter "hits" by 95% confidence interval for ICC, using the results of RQ1 and RQ2)?

4.4 *Description of RSA-911 Data*

The RSA-911 data in FY 2015 (which RSA-911 is supporting information by state VR agencies for rehabilitation services administration by the U.S. Department of Education) is used to test the proposed methods for the ICC in different simulation scenarios of multilevel structure models. As for the foundations of evidence-based rehabilitation, the target population is defined as the entire group of the "usable" clients (who were having an individualized plan for employment, IPE, and had been receiving VR services already by their IPE) from the public VR program in the State of Michigan. There are 33 VR office structures in Michigan that are used as an indicator of level-2 units in HLM & HGLM analyses in the simulation study.

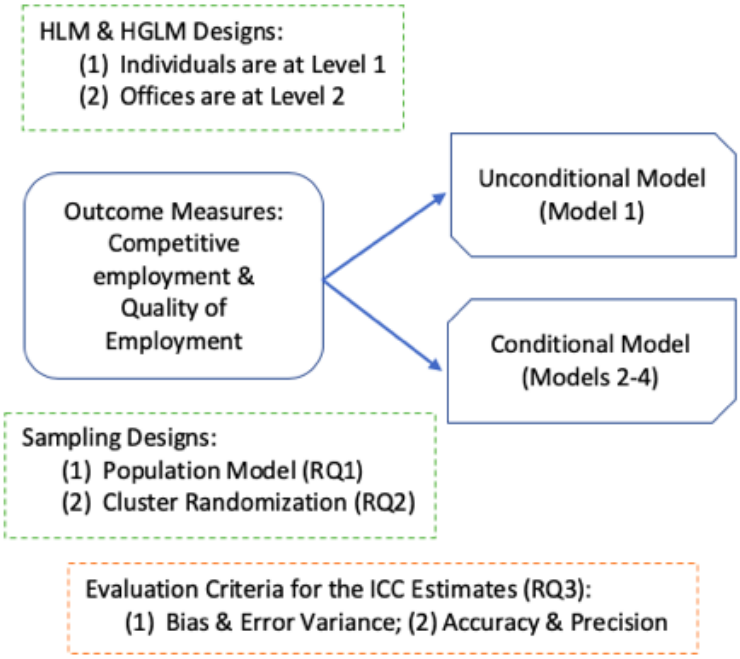
4.5 *Simulation and Analysis Plan*

To address the proposed research questions, a simulation study via the existing RSA-911 data (representing a complex system in a real-world situation) is conducted by 2-level hierarchical design modeling, where individual is on level-1, and office is on level-2. Two types of the proposed hierarchical models are considered in analyses: (1) unconditional model (without covariates) is designed by Model 1; and (2) conditional model (with covariates) is given by Models 2-4. To test proposed multilevel designs and their modeling structures in different sampling scenarios (i.e., "full data set" versus "three schemes of cluster samples"), we apply a simulation analysis to compare the results between unconditional and conditional

models, with respect to four different sampling schemes (i.e., the "population" model by the full data set in RQ1, plus three different cluster sampling procedures in RQ2). Furthermore, in test design and evaluation, three outcomes of interest in the study (rehabilitation outcome, competitive employment, and quality of employment) are used to examine the statistical performance (effectiveness analyses) of the proposed models and the simulation results, in terms of statistical bias, error bias, and accuracy & precision (in RQ3).

A graphic overview of the simulation process in the study is shown as a workflow chart below in Figure 4.1.

Figure 4.1 A Workflow Diagram of Simulation-based Exploration and Evaluation for the ICC



In computer simulations via the RSA-911, the statistical software R (Linear Mixed

Model *lmer* and Generlized Linear Mixed Model *glmer* in the package of *lme* or *lme4*), IBM SPSS (Mixed Effect Model by *MIXED*; Generalized Linear Mixed Model by *GENLINMIXED*; Variance Component Analysis by *VARCOMP*), SAS (Mixed Effect Modeling through *Proc Mixed*; Generalized Linear Mixed Model via *Proc Glimmix*), and Stata (Multilevel Mixed Model through *Xtmixed* or *Mixed*) are used for conducting statistical analysis and outcome performance evaluation for simulation results of ICC estimation and statistical inference.

4.6 *Theoretical Framework of HLM and HGLM in 2-Level Cluster Randomized Design*

This section provides mathematical details of multilevel modeling structures used in the study.

4.6.1. *HLM in 2-Level Cluster Randomized Structure via RSA-911*

In the two-level hierarchical design structure (i.e., individuals are at the level 1, and offices are at the level 2), the unconditional model (involving with no covariates) is described in Equation 11:

$$\begin{array}{ll} (Level\ 1) & Y_{ij} = \beta_{0i} + \varepsilon_{ij} \\ (Level\ 2) & \beta_{0i} = \mu_{00} + \xi_i \\ (Overall) & Y_{ij} = \mu_{00} + \xi_i + \varepsilon_{ij} \end{array} \qquad (11)$$

, where Y_{ij} represents an outcome for the j-th individual subject (at the level 1; $j = 1, \dots, n_i$) in the i-th office (at the level 2; $i = 1, \dots, m$), μ_{00} is a grand mean outcome that can be estimated by $\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij}/(m \times n)$, ε_{ij} is a random error term (or individual variation) at the level 1 (i.e., $\varepsilon_{ij} \sim NID(0, \sigma_W^2)$) corresponding to the j-th person in the i-th group, ξ_i is a random effect (i.e., $\xi_i \sim NID(0, \sigma_B^2)$) associated with the i-th office (or cluster variation at the level 2), the within-cluster (i.e., between-person) variance component is given by $Var(\varepsilon_{ij}) = \sigma_W^2$, the between-cluster variance component is given by $Var(\xi_i) = \sigma_B^2$, and the random error terms at the level 1 and level 2 are assumed to be not mutually correlated (i.e., $Cov(\{\xi_i\}, \{\varepsilon_{ij}\}) = 0$).

When a covariate (e.g., age groups) used in the hierarchical design, the conditional model (involving with one covariate centered at the group mean) is written in Equation 12:

$$\begin{aligned} (Level\ 1) \quad & Y_{ij} = \beta_{0i} + \beta_{1i}(X_{ij} - \bar{X}_{i\cdot}) + \varepsilon_{ij} \\ (Level\ 2) \quad & \beta_{0i} = \mu_{00} + \mu_{01}\,\bar{X}_{i\cdot} + \xi_i \\ (Overall) \quad & Y_{ij} = \mu_{00} + \mu_{01}\,\bar{X}_{i\cdot} + \beta_{1i}(X_{ij} - \bar{X}_{i\cdot}) + \xi_i + \varepsilon_{ij} \end{aligned} \tag{12}$$

, where the covariate model uses group (office) mean centering for reducing correlation between groups (Paccagnella, 2006; Raudenbush & Bryk, 2002), the Level 1 model is for the j-th person ($j = 1, \dots, n_i$) and the Level 2 is for the i-th group ($i = 1, \dots, m$), X_{ij} is the covariate for the j-th individual subject in the i-th office, $\bar{X}_{i\cdot}$ is group mean for the i-th group, ξ_i is a

random effect of the i-th office (a random residual at Level 2), ε_{ij} is an individual error term for the j-th person (a random residual at Level 1), β_{1i} is the common covariate's slope (assuming all covariate's slopes are equal across offices), μ_{00} is grand mean, and independence between errors at levels 1 and 2.

4.6.2. *HGLM in 2-Level Cluster Randomized Structure via RSA-911*

Suppose that Y_{ij} is a binary outcome variable for the j-th individual subject (at the level 1; $j = 1, \dots, n_i$) from the i-th cluster (office). In the 2-level cluster randomized trial, the 2-level hierarchical generalized linear model, HGLM, (involving with no covariates) is given in Equation 13:

$$logit(Y_{ij}) = log[P_{ij}/(1 - P_{ij})] = \mu_{00} + \xi_i + \varepsilon_{ij} \quad (13)$$

, where Y_{ij} denote a dichotomous outcome (coded as zero or one) for the j-th individual subject (at the level 1; $j = 1, \dots, n_i$) from the i-th office (at the level 2; $i = 1, \dots, m$), μ_{00} is grand mean, ε_{ij} is an individual error term at the level 1 (i.e., $\varepsilon_{ij} \sim NID(0, \sigma_W^2)$) corresponding to the j-th person in the i-th group, ξ_i is a random effect (i.e., $\xi_i \sim NID(0, \sigma_B^2)$) associated with the i-th group (or office variation at the level 2), the within-group variance is given by $Var(\varepsilon_{ij}) = \sigma_W^2$, the between-group variance is given by $Var(\xi_i) = \sigma_B^2$, and the random error terms at the level 1 and level 2 are assumed to be not mutually independent (i.e., $Cov(\{\xi_i\}, \{\varepsilon_{ij}\}) = 0$).

When a covariate (e.g., minority groups) used in the 2-level generalized hierarchical design, the conditional model (involving with one covariate centered at the group mean) is written in Equation 14:

$$\begin{aligned} logit(Y_{ij}) &= log[P_{ij}/(1-P_{ij})] \\ &= \mu_{00} + \mu_{01}\,\bar{X}_{i\cdot} + \beta_{1i}(X_{ij} - \bar{X}_{i\cdot}) + \xi_i + \varepsilon_{ij} \end{aligned} \tag{14}$$

, where the generalized or binary covariate model is centered by cluster (office) mean, Level 1 is denoted for the j-th person ($j = 1, \dots, n_i$) and Level 2 is denoted for the i-th group ($i = 1, \dots, m$), X_{ij} is a covariate for the j-th person in the i-th group, $\bar{X}_{i\cdot}$ is group mean for the i-th cluster, ξ_i is a random effect of the i-th office (a residual at Level 2), ε_{ij} is an individual error for the j-th subject (a residual at Level 1), β_{1i} is the common covariate's slope (assuming covariate's slopes are not the same across office structures), μ_{00} is grand mean, and random errors at levels 1 and 2 are assumed to be mutually independent (Klar & Donner, 2001; Raudenbush & Bryk, 2002).

CHAPTER 5

RESULTS

5.1 *Data Source and Sample Characteristics*

This study used the real data set of Rehabilitation Services Administration, RSA-911, in FY 2015 to examine and verify the proposed analytic methods of intraclass correlation (ICC) estimation and related inferential statistics (e.g., confidence interval and p-value) in different types of scenarios with respect to hierarchical design and modeling structure. The target samples are selected from people with disabilities who had been receiving services in the Michigan Rehabilitation Services Programs for vocational rehabilitation and supported employment. Note that in order to select usable samples for data simulations, this study only includes those samples having an individualized plan for employment (IPE) for services in vocational rehabilitation (VR), while all other subjects (ineligible for VR or not having an IPE) are excluded from the target samples and not considered further in data analysis for ICC calculations. In simulation analysis of the study, the target sample is of size N=17,633, while the usable sample size is n=11,819 for ICC estimation and inference. By hierarchical design & model considerations (i.e., individuals are on Level 1 and offices are on Level 2), all usable samples are distributed across 33 office units statewide in Michigan (see Tables B.1 and B.2 and Figure B.1 in Appendix B for an illustration of the hierarchical spatial data structure for usable samples in Michigan from RSA-911). Individual characteristics of the usable samples

are described in Tables 5.1, 5.2 and 5.3 for more details.

Table 5.1 Individual Characteristics of the Usable Samples (*n*=11,819)

Demographic Background	Frequency	Percentage
Gender		
Female	5,069	42.90%
Male	6,750	57.10%
Age		
Younger than 22	3,771	31.91%
Ages 22-40	2,905	24.58%
Ages 40-64	4,734	40.05%
Older than 65	409	3.46%
Minority		
Yes (Non-Whites)	7,757	65.63%
No (Whites)	4,062	34.37%
Education		
Elementary or Secondary	3,177	26.88%
Special Education	840	7.11%
High School	5,075	42.94%
College Above	2,727	23.07%
Social Security Benefits		
No	9,168	77.60%
Yes	2,651	22.40%
Total	11,819	100.00%

Note1. Minority group is defined as the non-white populations (e.g., Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islanders). Non-minority is defined as the white population (i.e., an individual's origins come from Europe, the Middle East or North African, according to the RSA-911 Report Manual; also see Appendix A).
Note2. Mean of Age = 36.4, and Standard Deviation of Age = 16.3.

Table 5.2 Disability & Rehabilitation Characteristics of the Usable Samples (*n*=11,819)

Disability & Rehabilitation Information	Frequency	Percentage
Type of Disability		
VI: Visual Impairments	87	0.70%
HI: Hearing Impairments	1,989	16.80%
PI: Physical Impairments	2,154	18.20%
LD: Learning Disability	2,276	19.30%
ADHD	443	3.70%
ID	652	5.50%
TBI	132	1.10%
ASD: Autism	436	3.70%
MI: Mental Illness	3,073	26.00%
SA: Substance Abuse	577	4.90%
Significance of Disability		
No	1,259	10.65%
Yes	10,560	89.35%
Previous Work Background		
No Work Experience	8,836	74.76%
Had Work Experience	2,983	25.24%
Job Placement Assistance Service		
Not Received	7,347	62.20%
Received	4,472	37.80%
On-the-job Supports Service		
Not Received	11,076	93.70%
Received	743	6.30%
Rehabilitation Technology Service		
Not Received	9,610	81.30%
Received	2,209	18.70%
Total	11,819	100.00%

Note. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.

Table 5.3 Outcomes of the Usable Samples (*n*=11,819)

Outcome Measure	Frequency	Percentage
Rehabilitation Outcome		
Not Employment	5,201	44.01%
Employment	6,618	55.99%
Competitive Employment		
Not Competitive Employment	6,787	45.60%
Competitive Employment	6,429	54.40%
Weekly Earnings		
Below $100 Weekly Income	5,409	45.80%
$100-$200 Weekly Income	1,647	13.90%
$200-$300 Weekly Income	1,506	12.70%
Above $300 Weekly Income	3,257	27.60%
Total	11,819	100.00%

Note1. Median of Weekly Earnings = 148.0, Mean of Weekly Earnings = 224.5, and Standard Error of Mean (SEM) of Weekly Earnings = 3.0.
Note2. Weekly Earnings can also be deemed as an indicator of quality employment.

5.2 *Models and Variables Used for Simulations of ICC Analysis*

There are four multilevel modeling structures (Models 1-4; M1-M4) in the study to test the proposed methods of ICC estimation and inference. Furthermore, three disability-related covariates – significance of disability (dichotomous; W1), type of disability (nominal; W2), and previous work experience (dichotomous; W3) – are used as a "breaking" variable for subgroup analysis (i.e., separating the whole usable sample into different and mutually exclusive sub-samples) in the all four designated models (M1-M4). Three covariate sets are considered for statistical adjustment in the multilevel modeling procedure: (1) Covariate Set 1 (CVS1)

includes demographic information such as gender (dichotomous; X1), minority (dichotomous; X2), age (continuous; X3), social security benefits (dichotomous; X4), and education background (ordinal or approximately continuous; X5); (2) Covariate Set 2 (CVS2) includes rehabilitation service information such as job placement assistance (dichotomous; X6), on-the-job supports (dichotomous; X7), and rehabilitation technology (dichotomous; X8); (3) Covariate Set 3 (CVS3) combines the previous covariate sets together (both CVS1 and CVS2) to account for all individual information in multilevel modeling. Two different outcome measures, competitive employment (dichotomous; Y1) and weekly earnings (continuous; Y2), are used in the evaluation of each proposed method of ICC calculations. Note that the Pearson's correlation structures between predictors, covariates and outcomes are shown in Tables 5.4 and 5.5, and that the associations between disability type and outcomes are described via one-way analysis of variance (ANOVA) in Table 5.6. For outcome measure Y1, except for X1 (p-value=0.41), all other predictors (X2-X8) and covariates (W1 & W3) are correlated with the outcome measure Y1 at the significance level of 0.05 (see Table 5.4). For outcome measure Y2, all predictors (X1-X8) and covariates (W1 & W3) are correlated with the outcome measure Y2 at the significance level of 0.05 (see Table 5.5). For the association of W2 (Type of Disability) with both outcome measures Y1 & Y2, it demonstrates in Table 5.6 that disability type is a significant factor in explaining total variation of both outcome measures, and that the measure of strength of association (i.e., *F*-statistic in ANOVA along with Eta-squared as an ICC effect size measure) is significant at the alpha level of 0.05.

In all, it suggests those predictors (X1-X8) and covariates (W1-W3) have prospective associations with key outcome variables (Y1-Y2), and that this statistical evidence may provide supportive information linked to favorable and promising ICC calculations in the study.

Table 5.4 Correlation Structure of All Predictors and Outcome Y1 in Hierarchical Analysis

	Y1	X1	X2	X3	X4	X5	X6	X7	X8	W1	W3
Y1	1.00	0.01	-0.09	0.18	-0.16	0.16	0.07	0.07	0.31	-0.21	0.31
X1	0.01	1.00	-0.01	-0.04	-0.02	-0.08	0.02	0.03	-0.05	0.03	-0.06
X2	-0.09	-0.01	1.00	0.01	0.10	-0.05	-0.01	-0.06	-0.23	0.13	-0.19
X3	0.18	-0.04	0.01	1.00	0.01	0.51	-0.16	-0.13	0.40	-0.26	0.36
X4	-0.16	-0.02	0.10	0.01	1.00	0.00	0.10	0.12	-0.15	0.18	-0.17
X5	0.16	-0.08	-0.05	0.51	0.00	1.00	-0.08	-0.10	0.29	-0.18	0.28
X6	0.07	0.02	-0.01	-0.16	0.10	-0.08	1.00	0.21	-0.25	0.17	-0.26
X7	0.07	0.03	-0.06	-0.13	0.12	-0.10	0.21	1.00	-0.10	0.07	-0.08
X8	0.31	-0.05	-0.23	0.40	-0.15	0.29	-0.25	-0.10	1.00	-0.38	0.56
W1	-0.21	0.03	0.13	-0.26	0.18	-0.18	0.17	0.07	-0.38	1.00	-0.39
W3	0.31	-0.06	-0.19	0.36	-0.17	0.28	-0.26	-0.08	0.56	-0.39	1.00

Note1. Y1=Competitive Employment; X1=Gender; X2=Minority; X3=Age; X4=Social Benefits; X5=Education; X6=Job Placement; X7=On-the-job Supports; X8=Rehabilitation Technology; W1=Significance of Disability; W3=Previous Work Experience.
Note2. Except for X1 (p-value=0.41), all other predictors (X2-X8) and covariates (W1 & W3) are correlated with the outcome measure Y1 at the significance level of 0.05.
Note3. W2 (Type of Disability) is not included, due to the categorical (nominal) measurement.

Table 5.5 Correlation Structure of All Predictors and Outcome Y2 in Hierarchical Analysis

	Y2	X1	X2	X3	X4	X5	X6	X7	X8	W1	W3
Y2	1.00	0.05	-0.14	0.32	-0.22	0.28	-0.16	-0.07	0.51	-0.34	0.47
X1	0.05	1.00	-0.01	-0.04	-0.02	-0.08	0.02	0.03	-0.05	0.03	-0.06
X2	-0.14	-0.01	1.00	0.01	0.10	-0.05	-0.01	-0.06	-0.23	0.13	-0.19
X3	0.32	-0.04	0.01	1.00	0.01	0.51	-0.16	-0.13	0.40	-0.26	0.36
X4	-0.22	-0.02	0.10	0.01	1.00	0.00	0.10	0.12	-0.15	0.18	-0.17
X5	0.28	-0.08	-0.05	0.51	0.00	1.00	-0.08	-0.10	0.29	-0.18	0.28
X6	-0.16	0.02	-0.01	-0.16	0.10	-0.08	1.00	0.21	-0.25	0.17	-0.26
X7	-0.07	0.03	-0.06	-0.13	0.12	-0.10	0.21	1.00	-0.10	0.07	-0.08
X8	0.51	-0.05	-0.23	0.40	-0.15	0.29	-0.25	-0.10	1.00	-0.38	0.56
W1	-0.34	0.03	0.13	-0.26	0.18	-0.18	0.17	0.07	-0.38	1.00	-0.39
W3	0.47	-0.06	-0.19	0.36	-0.17	0.28	-0.26	-0.08	0.56	-0.39	1.00

Note1. Y2=Weekly Earnings; X1=Gender; X2=Minority; X3=Age; X4=Social Benefits; X5=Education; X6=Job Placement; X7=On-the-job Supports; X8=Rehabilitation Technology; W1=Significance of Disability; W3=Previous Work Experience.
Note2. All predictors (X1-X8) and covariates (W1 & W3) are correlated with the outcome measure Y2 at the significance level of 0.05.
Note3. W2 (Type of Disability) is not included, due to the categorical (nominal) measurement.

Table 5.6 Summary of Mean Differences in the Outcomes between Type of Disability

Type of Disability (W2)	Competitive Employment Outcome (Y1)	Quality of Employment Outcome (Y2)
VI	0.62	250.15
HI	0.86	578.54
PI	0.49	199.74
LD	0.48	140.62
ADHD	0.47	135.19
ID	0.48	103.63
TBI	0.48	180.94
ASD	0.52	123.64
MI	0.46	138.34
SA	0.50	173.38
Overall Mean (Standard Error)	0.54 (SE=0.01)	224.48 (SE=3.02)
F-value (p-value)	118.36 (*p*-value < 0.01)	421.52 (*p*-value < 0.01)
Eta-squared (or ICC)	0.08	0.24

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. F-value is based on One-way Analysis of Variance (ANOVA).
Note3. Eta-squared ($\eta^2 = SS_{Between}/SS_{Total}$) *is a measure of strength of association in ANOVA, and it can be computed as between-group sum of squares divided by total sum of squares, which is another form of effect-size measure of intraclass correlation coefficients (ICC). See more detail in Section 2.1.*

There are two types of multilevel modeling structures in the simulation study. The first one is unconditional model (Model 1, or M1) with no covariates adjusted; and the second one is conditional model (Model 2-4, or M2-4) with an adjustment of covariates (i.e., M2|CVS1, M3|CVS2, and M4|CVS3). Note that CVS1 is a pre-specified covariate set 1 about demographic information in Model 2 (M2), CVS2 is about rehabilitation service information in

Model 3 (M3), and CVS3 is about all individual information linking both CVS1 and CVS2 in Model 4 (M4).

The statistical model specification for both unconditional model (M1) and conditional model (M2-M4) is described as following:

(1) Unconditional Model (Model 1; M1):

In the two-level multilevel design structure (i.e., individual subjects are on the level 1, and office units are on the level 2), the unconditional model with no covariates-adjusted is shown in the system of Equation 15:

$$\begin{array}{ll} (Level\ 1) & Y_{ij} = \beta_{0i} + \varepsilon_{ij} \\ (Level\ 2) & \beta_{0i} = \mu_{00} + \xi_i \\ (Overall) & Y_{ij} = \mu_{00} + \xi_i + \varepsilon_{ij} \end{array} \quad (15)$$

, where Y_{ij} represents an outcome measure for the j-th individual subject (at the level 1; $j = 1, \dots, n_i$;) in the i-th office unit (at the level 2; $i = 1, \dots, m$; $\sum_{i=1}^{m} n_i = N$), μ_{00} is a grand mean outcome that can be estimated as $\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{ij}/N$, ε_{ij} is a random error term (or individual variation) at the level 1 (i.e., $\varepsilon_{ij} \sim NID(0, \sigma_W^2)$) corresponding to the j-th person in the i-th group, ξ_i is a random effect (i.e., $\xi_i \sim NID(0, \sigma_B^2)$) associated with the i-th office (or cluster variation at the level 2), the within-cluster (i.e., between-person) variance component is given by $Var(\varepsilon_{ij}) = \sigma_W^2$, the between-cluster variance component is given by $Var(\xi_i) = \sigma_B^2$, and the

random error terms at the level 1 and level 2 are assumed to be $Cov(\{\xi_i\},\{\varepsilon_{ij}\}) = 0$.

(2) Conditional Model (Models 2-4; M2-M4)

When a pre-specified covariate set (i.e., CSV1-CSV3) is added in the previous unconditional model (M1), the *conditional model with a covariate set-adjusted*, where the covariate set $\boldsymbol{CSV} = [X_{1ij}, \dots, X_{kij}]$ is to be centered at the group mean on each level, can be described in the system of Equation 16:

$$\begin{aligned} (Level\ 1) \quad & Y_{ij} = \beta_{0i} + \beta_{1i}\left(X_{1ij} - \bar{X}_{1i\cdot}\right) + \cdots + \beta_{ki}\left(X_{kij} - \bar{X}_{ki\cdot}\right) + \varepsilon_{ij} \\ (Level\ 2) \quad & \beta_{0i} = \mu_{00} + \mu_{01}\,\bar{X}_{1i\cdot} + \cdots + \mu_{0k}\,\bar{X}_{ki\cdot} + \xi_i \\ (Overall) \quad & Y_{ij} = \mu_{00} + \sum_{l=1}^{k} \mu_{0l}\,\bar{X}_{li\cdot} + \sum_{l=1}^{k} \beta_{li}\left(X_{lij} - \bar{X}_{li\cdot}\right) + \xi_i + \varepsilon_{ij} \end{aligned} \tag{16}$$

, where the conditional model with covariate mean adjustment uses group-mean centering for reducing correlation between groups (Paccagnella, 2006; Raudenbush & Bryk, 2002), Level 1 is for the j-th person ($j = 1, \dots, n_i$) and Level 2 is for the i-th group ($i = 1, \dots, m$), X_{lij} is the l-th covariate for the j-th individual subject in the i-th office, $\bar{X}_{li\cdot}$ is group mean of the l-th covariate for the i-th group, ξ_i is a random effect of the i-th office (a random residual at the level 2), ε_{ij} is an individual error term for the j-th person (a random residual at the level 1), β_{li} is the l-th covariate's slope for the i-th group (assuming each of slopes are varied across offices), μ_{00} is grand mean, μ_{0l} is the slope regressed on the grand mean for the l-th covariate adjusted by group mean, and independence is assumed between errors at levels 1 and 2.

5.3 *ICC Estimation Method and Its Inferential Statistics*

The proposed intraclass correlation (ICC) estimator via Analysis of Variance (ANOVA), shown below in Equation 17, is suitable for either a balanced (equal size over groups) or unbalance design (unequal size across groups):

$$\rho_{Intraclass} = {(MSA - MSW)}\Big/{[MSA + (n_0 - 1)MSW]} \quad (17)$$

, where *MSA* is Mean Squares Among Groups in the ANOVA, *MSW* is Mean Squares Within Groups in the ANOVA, n_i is the i-th group size, $n_0 = [N - \sum_{i=1}^{k} n_i^2/N]/(k-1)$, and N is the total sample size, i.e., $N = \sum_{i=1}^{k} n_i$. Note that computational information pertinent to the ANOVA for the ICC estimator (in Equation 17) is specified below in great detail.

Suppose Y_{ij} is decomposed by analysis of variance (ANOVA) for the intraclass correlation (ICC) estimator, where Y_{ij} is an outcome measure for the j-th person ($j = 1, \dots, n_i$) in the i-th group ($i = 1, \dots, k$). The source of overall variation (or sum of squares, SS) is defined by $SST = SSA + SSW$, where the among-group variation $SSA = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$, the within-group variation $SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2$, and the total variation $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2$. The mean squares source (MS) in ANOVA can be obtained through the formula (i.e., regression toward the mean or the average of variation) $MS = SS/DF$, that

is, $MSA = SSA/DF(Among\ Groups)$ and $MSW = SSW/DF(Within\ Groups)$, where $MSA = \frac{SSA}{DF(Among\ Groups)} = \frac{SSA}{(k-1)} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2/(k-1)$ is the mean variation among groups, $MSW = \frac{SSW}{DF(Within\ Groups)} = \frac{SSW}{k(n_0-1)} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\cdot}\right)^2/[k(n_0-1)]$ is the mean variation within groups (or the mean squared error), $DF(Among\ Groups)$ is $k-1$ (for k = the number of groups) representing the between-group degrees of freedom, the within-group degrees of freedom $DF(Within\ Groups)$ is $k \times (n_0 - 1)$ (for n_0 = the average number of within-group subjects = weighted mean group size). Note that the original idea of analysis of variable (ANOVA) for ICC estimation can be referred to Table 2.1 (Donner & Koval, 1980a).

Furthermore, the variance of the ICC estimate can be obtained by

$$Var(\rho_{Intraclass}) = {2N(1-\hat{\rho})^2}\Big/{N\sum_{i=1}^{k}[n_i(n_i-1)V_iW_i^{-2}] - \hat{\rho}^2[\sum_{i=1}^{k}n_i(n_i-1)W_i^{-1}]} \quad (18)$$

, where the sampling weights are $V_i = 1 + (n_i - 1)\hat{\rho}^2$ and $W_i = 1 + (n_i - 1)\hat{\rho}$, the total sample size is $N = \sum_{i=1}^{k} n_i$, and $\hat{\rho}$ is the ICC estimate as $\rho_{Intraclass}$.

Thus, the standard error of the ICC estimate is $SE(\rho_{Intraclass}) = \sqrt{Var(\rho_{Intraclass})}$.

The proposed testing statistic of the ICC estimate ($\rho_{Intraclass}$) can be written that

$$F_0 = \frac{MSA}{MSW} \quad (19)$$

, where the test statistic F_0 follows an F distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = k \times (n_0 - 1)$, for hypothesis testing H_0: $\rho_{Intraclass} = 0$ versus H_a: $\rho_{Intraclass} \neq 0$.

Given the sampling F distribution for the ICC estimate, the $100(1-\alpha)\%$ confidence interval on the intraclass correlation can be obtained by

$$P\left(\frac{F_0 - F_{1-\frac{\alpha}{2},df_1,df_2}}{F_0 + (n_0 - 1) \times F_{1-\frac{\alpha}{2},df_1,df_2}} \leq \rho_{Intraclass} \leq \frac{F_0 - F_{\frac{\alpha}{2},df_1,df_2}}{F_0 + (n_0 - 1) \times F_{\frac{\alpha}{2},df_1,df_2}}\right) = 1 - \alpha \quad (20)$$

, where this $100(1-\alpha)\%$ confidence limit for the ICC ($\rho_{Intraclass}$) represents the degree of total variability accounted for by between-group variation in multilevel design. It is noteworthy that the interval estimate on $\rho_{Intraclass}$ may not be very accurate and precise for a small sample size (i.e., small n_0 or N) or low reliability in measurements (i.e., large MSW or $\hat{\sigma}_W^2$). Also, it should be pointed out that the lower confidence limit on $\rho_{Intraclass}$ could be negative (especially when small sample size or large measurement error occurs in hierarchical modeling), but since $\rho_{Intraclass}$ normally should not be negative anyway by its mathematical definition (i.e., $0 \leq \rho_{Intraclass} \leq 1$), it is customary to replace the negative lower bound with "zero" for a post hoc adjustment.

For statistical planning in multilevel design, the proposed auxiliary statistics are used to help understand minimum detectable effect size with respect to desired power and required

sample size. Three types of measures linked with the intraclass correlation (ICC) estimator are:

(i) Design effect (DE), or variance inflation factor (VIF), is written by

$$DE = VIF = [1 + (n_0 - 1)\hat{\rho}] \tag{21}$$

, where $\hat{\rho}$ is the ICC estimate ($\rho_{Intraclass}$) which provides a statistical measure of homogeneity within groups (i.e., if within-group subjects are homogeneous perfectly $\sigma_W^2 \to 0$, then $\hat{\rho} \to 1$ and hence $DE \to n_0$). In general, grouping creates more variation than simple random sampling by a factor of DE (or $VIF > 1$), due to the major part of group-to-group variability plus the minor portion of within-group variation (i.e., samples in different groups vary more than those in the same group).

(ii) The unconditional intraclass correlation coefficient is given by

$$\rho = {\sigma_B^2}\Big/{(\sigma_W^2 + \sigma_B^2)} = {\sigma_B^2}\Big/{\sigma_T^2} \tag{22}$$

, where the unconditional total variance is $\sigma_T^2 = \sigma_W^2 + \sigma_B^2$, σ_W^2 and σ_B^2 represent error variances corresponding to the within- and between-group variation, respectively, in the unconditional model with no covariates adjusted in multilevel design.

In hierarchical models with covariates for statistical adjustment, the conditional intraclass correlation coefficient is defined as

$$\rho_A = {\sigma_{AB}^2}\Big/{(\sigma_{AW}^2 + \sigma_{AB}^2)} = {\sigma_{AB}^2}\Big/{\sigma_{AT}^2} \tag{23}$$

, where the covariate-adjusted total variance is $\sigma_{AT}^2 = \sigma_{AW}^2 + \sigma_{AB}^2$, σ_{AW}^2 and σ_{AB}^2 represent the variance components, adjusted by covariates, corresponding to the within- and between-group variation, respectively, in the conditional multilevel model.

(iii) For evaluating the relative efficiency of measures of homogeneity and heterogeneity in multilevel design, two statistical ancillary quantities, based on random variations of both unconditional and conditional hierarchical models, are given by

$$\eta_B^2 = {\sigma_{AB}^2}\Big/{\sigma_B^2} \tag{24}$$

and

$$\eta_W^2 = {\sigma_{AW}^2}\Big/{\sigma_W^2} \tag{25}$$

, where η_B^2 indicates the proportion of between-group variance remaining (after given by covariate adjustment) in multilevel design, and η_W^2 indicates the proportion of within-group variance remaining (after given by covariate adjustment) in multilevel design.

Both η_B^2 and η_W^2 measures show efficacy and effectiveness of covariate adjustment for between-group and within-group random variation in multilevel design and modeling.

The other two opposite measures (like a pseudo R-squared) for random variation by covariate adjustment in hierarchical modeling, are written by

$$R_B^2 = 1 - \eta_B^2 \tag{26}$$

and

$$R_W^2 = 1 - \eta_W^2 \tag{27}$$

, where R_B^2 and R_W^2 are defined as the proportion of between-group and within-group, respectively, variation explained by covariates adjusted in hierarchical design. Note that both R_B^2 and R_W^2 can also show efficacy of covariate adjustment in multilevel design.

5.4 *Results of ICC Estimates and Inferential Statistics*

There are two outcome measures in ICC's simulation studies: (1) One is a binary measure for *competitive employment* (Y1); (2) The other one is a continuous measure for weekly earned income or quality employment (Y2). Further, there are four different multilevel models for ICC calculations: (1) *Unconditional Model* (M1) is of no covariate adjustment; (2)

Conditional Model (M2) is fitted with covariate adjustment by the demographic predictors (Covariate Set1); (3) *Conditional Model* (M3) is fitted with covariate adjustment by the rehabilitation service predictors (Covariate Set2); (4) *Conditional Model* (M3) is fitted with covariate adjustment by both the demographic and service predictors (Covariate Set3). In addition, three breaking variables are considered for subset analysis of ICC estimation and inference using usable samples (n=11,819) in multilevel design: (1) Previous Work Experience – binary measure (i.e., yes or no); (2) Significance Disability – binary measure (i.e., yes or no); (3) Disability Type – nominal measure with 10 different disability categories (i.e., VI, HI, PI, LD, ADHD, ID, TBI, ASD, MI, and SA). In this section, the main results of the study are presented in the following Tables 5.7-5.16.

5.4.1 *Competitive Employment Outcome Measure*

The competitive employment (Y1) is fitted as a dichotomous outcome measure in the 2-level hierarchical generalized linear modeling (HGLM) framework, where individual subjects are on the level 1 and office units are on the level 2. The main results of the unconditional model M1 (Model 1) are shown in Table 5.7; the conditional model M2 (Model 2) in Table 5.8; the conditional model M3 (Model 3) in Table 5.9; the conditional model M4 (Model 4) in Table 5.10; Table 5.11 provides all the auxiliary information of ICC estimates such as design effect (DE), unconditional and conditional ICC's, and relative efficiency measures of η_B^2, η_W^2, R_B^2 and R_B^2; and Table 5.12 shows ICC evaluation results based on the bootstrap sampling procedure (the number of bootstrap repetitions=100).

The ICC estimates (including standard error, p-value, 95% confidence interval) for competitive outcome measure (Y1) under unconditional (Model 1) and conditional (Models 2-4) multilevel modeling structure, are summarized as follows.

For competitive employment (Y1 under Model 1; refer to Tables 5.7), the average (unadjusted) intraclass correlation is about 0.01 (SE=0.003, p<0.01, 95% CI = [0.01,0.02]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (unadjusted) ICC of 0.01 (SE=0.004, p<0.01, 95% CI = [0.01, 0.02]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (unadjusted) ICC of 0.02 (SE=0.009, p<0.01, 95% CI = [0.01, 0.05]). Breaking down by disability types, it finds that autism spectrum disorder (ASD) has the highest (unadjusted) ICC of 0.06 (SE=0.03, p<0.01, 95% CI = [0.00, 0.15]), followed by learning disability (LD; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.07]), hearing impairments (HI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the unadjusted ICC estimates in the following disabilities – visual impairments (VI, ICC=0.07, SE=0.10, p=0.25), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.54), intellectual disability (ID; ICC=0.02, SE=0.02, p=0.06), traumatic brain injury (TBI; ICC=0.02, SE=0.05, p=0.36), and substance abuse (SA; ICC=0.00, SE=0.01, p=0.48).

For competitive employment (Y1 under Model 2; refer to Tables 5.8), the average (adjusted by demographic information) intraclass correlation is about 0.01 (SE=0.003, p<0.01, 95% CI = [0.01,0.02]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (adjusted by demographic information) ICC of 0.01

(SE=0.004, p<0.01, 95% CI = [0.01, 0.03]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by demographic information) ICC of 0.02 (SE=0.01, p<0.01, 95% CI = [0.01, 0.05]). Breaking down by disability types, it finds that autism spectrum disorder (ASD) has the highest (adjusted) ICC of 0.06 (SE=0.03, p<0.01, 95% CI = [0.00, 0.15]), followed by learning disability (LD; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), hearing impairments (HI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the adjusted ICC estimates in the following disability types – visual impairments (VI, ICC=0.07, SE=0.10, p=0.26), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.54), intellectual disability (ID; ICC=0.02, SE=0.02, p=0.05), traumatic brain injury (TBI; ICC=0.02, SE=0.05, p=0.37), and substance abuse (SA; ICC=0.00, SE=0.01, p=0.48).

For competitive employment (Y1 under Model 3; refer to Tables 5.9), the average (adjusted by rehabilitation services information) intraclass correlation is about 0.01 (SE=0.003, p<0.01, 95% CI = [0.01,0.02]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (adjusted by rehabilitation services information) ICC of 0.01 (SE=0.005, p<0.01, 95% CI = [0.01, 0.03]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by rehabilitation services information) ICC of 0.02 (SE=0.01, p<0.01, 95% CI = [0.01, 0.05]). Breaking down by disability types, it finds that autism spectrum disorder (ASD) has the highest (adjusted) ICC of 0.08 (SE=0.04, p<0.01, 95% CI = [0.02, 0.17]), followed by learning disability (LD; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.02, 0.07]), intellectual disability (ID;

ICC=0.03, SE=0.02, p=0.02, 95% CI = [0.02, 0.09]), hearing impairments (HI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the adjusted ICC estimates in the following disability types – visual impairments (VI, ICC=0.09, SE=0.10, p=0.20), attention deficit hyperactivity disorder (ADHD; ICC=0.01, SE=0.02, p=0.29), traumatic brain injury (TBI; ICC=0.02, SE=0.05, p=0.35), and substance abuse (SA; ICC=0.00, SE=0.01, p=0.47).

For competitive employment (Y1 under Model 4; refer to Tables 5.10), the average (adjusted by both demographics and rehabilitation services) intraclass correlation is about 0.01 (SE=0.003, p<0.01, 95% CI = [0.01,0.02]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (adjusted by both demographics and rehabilitation services) ICC of 0.01 (SE=0.005, p<0.01, 95% CI = [0.01, 0.03]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by both demographics and rehabilitation services) ICC of 0.02 (SE=0.01, p<0.01, 95% CI = [0.01, 0.05]). Breaking down by disability types, it finds that autism spectrum disorder (ASD) has the highest (adjusted) ICC of 0.06 (SE=0.03, p<0.01, 95% CI = [0.01, 0.15]), followed by learning disability (LD; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), hearing impairments (HI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the adjusted ICC estimates in the following disability types – visual impairments (VI, ICC=0.07, SE=0.10, p=0.26), attention deficit hyperactivity disorder (ADHD;

ICC=0.00, SE=0.01, p=0.54), intellectual disability (ID; ICC=0.02, SE=0.02, p=0.05), traumatic brain injury (TBI; ICC=0.02, SE=0.05, p=0.37), and substance abuse (SA; ICC=0.00, SE=0.01, p=0.48).

For auxiliary information of ICC Estimates for Outcome Measure Y1 (see Tables 5.11), the unconditional model (Model 1; unconditional ICC=0.01 and design effect DE=4.44) is used as a baseline for measuring relative efficiency of between-group variance (η_B^2 and R_B^2) and within-group variance (η_W^2 and R_W^2) for ICC estimates. The conditional model (Model 2; conditional ICC=0.01 and design effect DE=4.59) with a covariate set of demographic information has a decrease of 3.05% of within-group variation and 0.00% of change in between-group variation, in comparison with the unconditional model (Model 1). The conditional model (Model 3; conditional ICC=0.01 and design effect DE=4.83) with a covariate set of rehabilitation service information has a decrease of 8.06% of within-group variation and an increase of 4.17% in between-group variation, in comparison with the unconditional model (Model 1). The conditional model (Model 4; conditional ICC=0.01 and design effect DE=4.59) with a covariate set of both demographic and rehabilitation service information has a decrease of 3.38% of within-group variation and no change (0.00%) in between-group variation, in comparison with the unconditional model (Model 1).

For evaluation of bootstrapping ICC estimates (bootstrap repetition of 100 times) for outcome measure Y1 in the different resampling scenarios of the number of groups and subjects (see Table 5.12), it provides important information of sampling schemes in multilevel structure (based on Model 4 with the full set of covariates of demographics and rehabilitation services). For the low level of cluster samples (i.e., number of groups=5), the mean bias is about 0.0068, MSE is about 0.0004, the proportion of successful hits is about 34%. For the medium level of

cluster samples (i.e., number of groups=15), the mean bias is about 0.0049, MSE is about 0.0002, the proportion of successful hits is about 66%. For the high level of cluster samples (i.e., number of groups=25), the mean bias is about 0.0047, MSE is about 0.0001, the proportion of successful hits is about 68%. On the other hand, For the low level of subject samples (i.e., number of subjects=50), the mean bias is about 0.0062, MSE is about 0.0003, the proportion of successful hits is about 41%. For the medium level of subject samples (i.e., number of subjects=100), the mean bias is about 0.0053, MSE is about 0.0002, the proportion of successful hits is about 59%. For the high level of subject samples (i.e., number of subjects=150), the mean bias is about 0.0047, MSE is about 0.0001, the proportion of successful hits is about 70%. Overall, the sampling scheme with the high level of group samples (i.e., 25) and high level of subject samples (i.e., 150) achieve the best outcome (i.e., lowest bias & MSE, and highest successful hits); the sampling scheme with moderate cluster and subject samples (i.e., number of groups=15 and number of subjects=100) can provide the average performance of ICC estimation; the sampling scheme with the low level of group samples (i.e., 5) or the level of group subject samples (i.e., 50) is more likely to result in poor performance of ICC estimates in hierarchical generalized linear modeling structure.

Table 5.7 ICC Estimates of Unconditional Model M1 for Outcome Measure Y1

Model 1	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0097	0.0031	0.00	0.0053	0.0187
Work Experience								
No	8,821	33	266	0.0119	0.0038	0.00	0.0064	0.0232
Yes	2,998	33	90	0.0101	0.0053	0.00	0.0026	0.0254
Significance Disability								
No	1,233	33	36	0.0297	0.0145	0.00	0.0093	0.0675
Yes	10,586	33	319	0.0107	0.0034	0.00	0.0058	0.0208
Disability Type								
VI	87	29	3	0.0732	0.1008	0.25	-0.1241	0.3253
HI	1,989	32	61	0.0201	0.0093	0.00	0.007	0.0459
PI	2,154	33	65	0.0187	0.0084	0.00	0.0067	0.0429
LD	2,276	33	68	0.0286	0.0105	0.00	0.0134	0.0585
ADHD	443	33	13	-0.0032	0.0149	0.54	-0.0303	0.0495
ID	652	33	19	0.0223	0.0173	0.06	-0.0041	0.0727
TBI	132	27	5	0.0212	0.0513	0.36	-0.0823	0.1919
ASD	436	33	13	0.0641	0.0329	0.00	0.0141	0.1505
MI	3,073	33	92	0.0175	0.0070	0.00	0.0075	0.0376
SA	577	31	18	-0.0006	0.0085	0.48	-0.0208	0.0405

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.8 ICC Estimates of Conditional Model M2 for Outcome Measure Y1

Model 2	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0101	0.0032	0.00	0.0055	0.0194
Work Experience								
No	8,821	33	266	0.0119	0.0038	0.00	0.0064	0.0232
Yes	2,998	33	90	0.0119	0.0057	0.00	0.0038	0.0284
Significance Disability								
No	1,233	33	36	0.0356	0.0160	0.00	0.0131	0.0767
Yes	10,586	33	319	0.0109	0.0034	0.00	0.0059	0.0211
Disability Type								
VI	87	29	3	0.0671	0.0996	0.26	-0.1289	0.3190
HI	1,989	32	61	0.0236	0.0102	0.00	0.0093	0.0517
PI	2,154	33	65	0.0188	0.0084	0.00	0.0067	0.0430
LD	2,276	33	68	0.0289	0.0105	0.00	0.0136	0.0588
ADHD	443	33	13	-0.0032	0.0149	0.54	-0.0303	0.0495
ID	652	33	19	0.0234	0.0176	0.05	-0.0034	0.0743
TBI	132	27	5	0.0190	0.0501	0.37	-0.0837	0.1892
ASD	436	33	13	0.0640	0.0329	0.00	0.0140	0.1504
MI	3,073	33	92	0.0175	0.0070	0.00	0.0075	0.0376
SA	577	31	18	-0.0006	0.0085	0.48	-0.0208	0.0405

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., $p<0.01$).

Table 5.9 ICC Estimates of Conditional Model M3 for Outcome Measure Y1

Model 3	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0108	0.0033	0.00	0.0060	0.0206
Work Experience								
No	8,821	33	266	0.0130	0.0041	0.00	0.0071	0.0250
Yes	2,998	33	90	0.0124	0.0059	0.00	0.0041	0.0291
Significance Disability								
No	1,233	33	36	0.0356	0.0160	0.00	0.0131	0.0767
Yes	10,586	33	319	0.0119	0.0037	0.00	0.0066	0.0228
Disability Type								
VI	87	29	3	0.0905	0.1038	0.20	-0.1105	0.3430
HI	1,989	32	61	0.0215	0.0097	0.00	0.0079	0.0482
PI	2,154	33	65	0.0192	0.0085	0.00	0.0070	0.0437
LD	2,276	33	68	0.0340	0.0116	0.00	0.0170	0.0672
ADHD	443	33	13	0.0096	0.0189	0.29	-0.0219	0.0694
ID	652	33	19	0.0320	0.0199	0.02	0.0023	0.0877
TBI	132	27	5	0.0224	0.0519	0.35	-0.0815	0.1935
ASD	436	33	13	0.0782	0.0356	0.00	0.0238	0.1705
MI	3,073	33	92	0.0187	0.0073	0.00	0.0083	0.0396
SA	577	31	18	0.0001	0.0089	0.47	-0.0204	0.0416

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.10 ICC Estimates of Conditional Model M4 for Outcome Measure Y1

Model 4	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0101	0.0032	0.00	0.0055	0.0195
Work Experience								
No	8,821	33	266	0.0119	0.0038	0.00	0.0064	0.0232
Yes	2,998	33	90	0.0120	0.0058	0.00	0.0038	0.0286
Significance Disability								
No	1,233	33	36	0.0359	0.0161	0.00	0.0133	0.0771
Yes	10,586	33	319	0.0109	0.0034	0.00	0.0060	0.0211
Disability Type								
VI	87	29	3	0.0673	0.0996	0.26	-0.1287	0.3192
HI	1,989	32	61	0.0237	0.0102	0.00	0.0094	0.0519
PI	2,154	33	65	0.0188	0.0084	0.00	0.0067	0.0430
LD	2,276	33	68	0.0290	0.0106	0.00	0.0137	0.0591
ADHD	443	33	13	-0.0033	0.0148	0.54	-0.0304	0.0493
ID	652	33	19	0.0237	0.0177	0.05	-0.0032	0.0749
TBI	132	27	5	0.0191	0.0501	0.37	-0.0837	0.1893
ASD	436	33	13	0.0645	0.0330	0.00	0.0144	0.1511
MI	3,073	33	92	0.0175	0.0070	0.00	0.0075	0.0377
SA	577	31	18	-0.0006	0.0085	0.48	-0.0208	0.0405

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.11 Auxiliary Information of ICC Estimates for Outcome Measure Y1

Modeling Structure	ICC Estimate	Between Group Variance	Within Group Variance	Design Effect (DE)	η_B^2	η_W^2	R_B^2	R_W^2
Model 1 (M1)	0.0097	0.0024	0.2458	4.4436	NA	NA	NA	NA
Model 2 (M2)	0.0101	0.0024	0.2383	4.5856	1.0000	0.9695	0.0000	0.0305
Model 3 (M3)	0.0108	0.0025	0.2260	4.8341	1.0417	0.9194	-0.0417	0.0806
Model 4 (M4)	0.0101	0.0024	0.2375	4.5856	1.0000	0.9662	0.0000	0.0338

Note1. The ICC estimate for M1 represents the unconditional ICC quantity, while the ICC's for M2-M4 show the conditional ICC quantity.
Note2. Relative efficiency measures for ICC estimates between unconditional and conditional models (M1 versus M2-M4) are η_B^2, η_W^2, R_B^2 and R_B^2.

Table 5.12 Evaluation of Bootstrap ICC Estimates for Outcome Measure Y1

Number of Group	Within Group Size	Bias	MSE	Hits
5	50	0.0078	0.0005	0.19
5	100	0.0069	0.0004	0.34
5	150	0.0056	0.0003	0.50
15	50	0.0054	0.0003	0.50
15	100	0.0048	0.0002	0.70
15	150	0.0045	0.0001	0.78
25	50	0.0053	0.0002	0.54
25	100	0.0042	0.0001	0.73
25	150	0.0033	0.0000	0.82

Note1. Bias is defined as the mean difference between Bootstrap ICC and True ICC.
Note2. MSE is the mean squared error difference between Bootstrap ICC estimates.
Note3. Hits shows the proportion of Bootstrap ICC estimates successfully lying within the 95% confidence interval of True ICC.

5.4.2 *Earnings or Quality Employment Outcome Measure*

The weekly earned income, or quality employment, (Y2) is fitted as a continuous outcome measure in the 2-level hierarchical linear modeling (HLM) framework, where individual subjects are on the level 1 and office units are on the level 2. The main results of the unconditional model M1 (Model 1) are shown in Table 5.13; the conditional model M2 (Model 2) in Table 5.14; the conditional model M3 (Model 3) in Table 5.15; the conditional model M4 (Model 4) in Table 5.16; and Table 5.17 provides all the auxiliary information of ICC estimates such as design effect (DE), unconditional and conditional ICC's, and relative efficiency measures of η_B^2, η_W^2, R_B^2 and R_B^2; and Table 5.18 shows ICC evaluation results based on the bootstrap sampling procedure (the number of bootstrap repetitions=100).

The ICC estimates (including standard error, p-value, 95% confidence interval) for quality of employment outcome measure (Y2) under unconditional (Model 1) and conditional (Models 2-4) multilevel modeling structure, are summarized as follows.

For quality employment (Y2 under Model 1; refer to Tables 5.13), the average (unadjusted) intraclass correlation is about 0.02 (SE=0.01, p<0.01, 95% CI = [0.01,0.04]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (unadjusted) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.05]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (unadjusted) ICC of 0.05 (SE=0.01, p<0.01, 95% CI = [0.03, 0.09]). Breaking down by disability types, it finds that learning disability (LD) has the highest (unadjusted) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.07]), followed by substance abuse (SA; ICC=0.03,

SE=0.02, p=0.04, 95% CI = [0.00, 0.09]), hearing impairments (HI; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the ICC estimates in the following disability types – visual impairments (VI, ICC=0.00, SE=0.08, p=0.50), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.87), intellectual disability (ID; ICC=0.02, SE=0.02, p=0.05), traumatic brain injury (TBI; ICC=-0.08, SE=0.03, p=0.88), and autism spectrum disorder (ASD; ICC=0.02, SE=0.02, p=0.13).

For quality employment (Y2 under Model 2; refer to Tables 5.14), the average (adjusted by demographic information) intraclass correlation is about 0.02 (SE=0.01, p<0.01, 95% CI = [0.01,0.04]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (adjusted by demographic information) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.05]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by demographic information) ICC of 0.05 (SE=0.01, p<0.01, 95% CI = [0.03, 0.09]). Breaking down by disability types, it finds that learning disability (LD) has the highest (adjusted by demographic information) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.07]), followed by hearing impairments (HI; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), substance abuse (SA; ICC=0.03, SE=0.02, p=0.04, 95% CI = [0.00, 0.09]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the ICC estimates in the following disability types – visual impairments (VI, ICC=0.00, SE=0.08, p=0.49), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.87), intellectual disability (ID;

ICC=0.02, SE=0.02, p=0.05), traumatic brain injury (TBI; ICC=-0.08, SE=0.03, p=0.87),and autism spectrum disorder (ASD; ICC=0.02, SE=0.02, p=0.13).

For quality employment (Y2 under Model 3; refer to Tables 5.15), the average (adjusted by rehabilitation services) intraclass correlation is about 0.02 (SE=0.01, p<0.01, 95% CI = [0.01,0.04]). Given by work experience (binary coding of yes or no) for partitioning subset samples, both show the average (adjusted by rehabilitation services) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.05]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by rehabilitation services) ICC of 0.05 (SE=0.01, p<0.01, 95% CI = [0.03, 0.09]). Breaking down by disability types, it finds that learning disability (LD) has the highest (adjusted by rehabilitation services) ICC of 0.04 (SE=0.01, p<0.01, 95% CI = [0.02, 0.07]), followed by substance abuse (SA; ICC=0.03, SE=0.02, p=0.04, 95% CI = [0.00, 0.09]), hearing impairments (HI; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), intellectual disability (ID; ICC=0.03, SE=0.02, p=0.03, 95% CI = [0.00, 0.08]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the ICC estimates in the following disability types – visual impairments (VI, ICC=0.00, SE=0.08, p=0.52), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.81), traumatic brain injury (TBI; ICC=-0.08, SE=0.03, p=0.86),and autism spectrum disorder (ASD; ICC=0.02, SE=0.02, p=0.12).

For quality employment (Y2 under Model 4; refer to Tables 5.16), the average (adjusted by both demographics and rehabilitation services) intraclass correlation is about 0.02 (SE=0.01, p<0.01, 95% CI = [0.01,0.04]). Given by work experience (binary coding of yes or no) for

partitioning subset samples, both show the average (adjusted by both demographics and rehabilitation services) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.05]). By significance disability (binary coding of yes or no) for subset analyses, both show the average (adjusted by both demographics and rehabilitation services) ICC of 0.05 (SE=0.01, p<0.01, 95% CI = [0.03, 0.09]). Breaking down by disability types, it finds that learning disability (LD) has the highest (adjusted by both demographics and rehabilitation services) ICC of 0.03 (SE=0.01, p<0.01, 95% CI = [0.02, 0.07]), followed by substance abuse (SA; ICC=0.03, SE=0.02, p=0.04, 95% CI = [0.00, 0.09]), hearing impairments (HI; ICC=0.03, SE=0.01, p<0.01, 95% CI = [0.01, 0.06]), physical impairments (PI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.05]), and mental illness (MI; ICC=0.02, SE=0.01, p<0.01, 95% CI = [0.01, 0.04]). Also noted that at the significance level of 0.05, it shows non-significance for the ICC estimates in the following disability types – visual impairments (VI, ICC=0.00, SE=0.08, p=0.49), attention deficit hyperactivity disorder (ADHD; ICC=0.00, SE=0.01, p=0.87), intellectual disability (ID; ICC=0.02, SE=0.02, p=0.05), traumatic brain injury (TBI; ICC=-0.08, SE=0.03, p=0.87),and autism spectrum disorder (ASD; ICC=0.02, SE=0.02, p=0.12).

For auxiliary information of ICC Estimates for Outcome Measure Y2 (see Tables 5.17), the unconditional model (Model 1; unconditional ICC=0.02 and design effect DE=8.49) is used as a baseline for measuring relative efficiency of between-group variance (η_B^2 and R_B^2) and within-group variance (η_W^2 and R_W^2) for ICC estimates. The conditional model (Model 2; conditional ICC=0.02 and design effect DE=9.38) with a covariate set of demographic information has a decrease of 9.75% of within-group variation and an increase of 1.27% of between-group variation, in comparison with the unconditional model (Model 1). The conditional model (Model 3; conditional ICC=0.02 and design effect DE=8.70) with a covariate

set of rehabilitation service information has a decrease of 2.47% of within-group variation and an increase of 0.32% in between-group variation, in comparison with the unconditional model (Model 1). The conditional model (Model 4; conditional ICC=0.02 and design effect DE=9.41) with a covariate set of both demographic and rehabilitation service information has a decrease of 10.02 % of within-group variation and an increase of 1.31% of between-group variation, in comparison with the unconditional model (Model 1).

For evaluation of bootstrapping ICC estimates (bootstrap repetition of 100 times) for outcome measure Y2 in the different resampling scenarios of the number of groups and subjects (see Table 5.18), it provides important information of sampling schemes in multilevel structure (based on Model 4 with the full set of covariates of demographics and rehabilitation services). For the low level of cluster samples (i.e., number of groups=5), the mean bias is about 0.0164, MSE is about 0.0009, the proportion of successful hits is about 34%. For the medium level of cluster samples (i.e., number of groups=15), the mean bias is about 0.0152, MSE is about 0.0004, the proportion of successful hits is about 55%. For the high level of cluster samples (i.e., number of groups=25), the mean bias is about 0.0149, MSE is about 0.0003, the proportion of successful hits is about 64%. On the other hand, For the low level of subject samples (i.e., number of subjects=50), the mean bias is about 0.0160, MSE is about 0.0007, the proportion of successful hits is about 40%. For the medium level of subject samples (i.e., number of subjects=100), the mean bias is about 0.0154, MSE is about 0.0004, the proportion of successful hits is about 54%. For the high level of subject samples (i.e., number of subjects=150), the mean bias is about 0.0148, MSE is about 0.0004, the proportion of successful hits is about 66%. Overall, the sampling scheme with the high level of group samples (i.e., 25) and high level of subject samples (i.e., 150) achieve the best outcome (i.e.,

lowest bias & MSE, and highest successful hits); the sampling scheme with moderate cluster or subject samples (i.e., number of groups=15 or number of subjects=100) can provide the average performance of ICC estimates in multilevel structure; the sampling scheme with the low level of group samples (i.e., 5) or the level of group subject samples (i.e., 50) is more likely to result in poor performance of ICC estimates in hierarchical linear modeling structure.

Table 5.13 ICC Estimates of Unconditional Model M1 for Outcome Measure Y2

Model 1	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0211	0.0054	0.00	0.0127	0.0381
Work Experience								
No	8,821	33	266	0.0134	0.0042	0.00	0.0073	0.0257
Yes	2,998	33	90	0.0408	0.0118	0.00	0.0227	0.0758
Significance Disability								
No	1,233	33	36	0.0797	0.0237	0.00	0.0422	0.1434
Yes	10,586	33	319	0.0171	0.0048	0.00	0.0100	0.0316
Disability Type								
VI	87	29	3	-0.0044	0.0798	0.50	-0.1832	0.2422
HI	1,989	32	61	0.0273	0.0110	0.00	0.0117	0.0577
PI	2,154	33	65	0.0223	0.0092	0.00	0.0090	0.0488
LD	2,276	33	68	0.0342	0.0116	0.00	0.0171	0.0676
ADHD	443	33	13	-0.0219	0.0081	0.87	-0.0425	0.0198
ID	652	33	19	0.0233	0.0176	0.05	-0.0034	0.0743
TBI	132	27	5	-0.0773	0.0281	0.88	-0.1447	0.0604
ASD	436	33	13	0.0226	0.0225	0.13	-0.0138	0.0899
MI	3,073	33	92	0.0190	0.0074	0.00	0.0085	0.0401
SA	577	31	18	0.0283	0.0207	0.04	-0.0026	0.0853

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., $p<0.01$).

Table 5.14 ICC Estimates of Conditional Model M2 for Outcome Measure Y2

Model 2	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0236	0.0059	0.00	0.0143	0.0423
Work Experience								
No	8,821	33	266	0.0135	0.0042	0.00	0.0074	0.0259
Yes	2,998	33	90	0.0457	0.0126	0.00	0.0260	0.0838
Significance Disability								
No	1,233	33	36	0.0869	0.0246	0.00	0.0471	0.1540
Yes	10,586	33	319	0.0183	0.0050	0.00	0.0108	0.0336
Disability Type								
VI	87	29	3	-0.0016	0.0808	0.49	-0.1811	0.2453
HI	1,989	32	61	0.0293	0.0115	0.00	0.0130	0.0610
PI	2,154	33	65	0.0227	0.0093	0.00	0.0093	0.0495
LD	2,276	33	68	0.0346	0.0117	0.00	0.0174	0.0682
ADHD	443	33	13	-0.0218	0.0081	0.87	-0.0425	0.0198
ID	652	33	19	0.0238	0.0177	0.05	-0.0031	0.0750
TBI	132	27	5	-0.0750	0.0287	0.87	-0.1433	0.0636
ASD	436	33	13	0.0233	0.0226	0.13	-0.0134	0.0908
MI	3,073	33	92	0.0193	0.0074	0.00	0.0087	0.0405
SA	577	31	18	0.0283	0.0207	0.04	-0.0026	0.0853

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.15 ICC Estimates of Conditional Model M3 for Outcome Measure Y2

Model 3	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0217	0.0055	0.00	0.0131	0.0391
Work Experience								
No	8,821	33	266	0.0136	0.0042	0.00	0.0074	0.0260
Yes	2,998	33	90	0.0429	0.0121	0.00	0.0241	0.0793
Significance Disability								
No	1,233	33	36	0.0855	0.0244	0.00	0.0462	0.1520
Yes	10,586	33	319	0.0175	0.0049	0.00	0.0103	0.0324
Disability Type								
VI	87	29	3	-0.0099	0.0777	0.52	-0.1872	0.2361
HI	1,989	32	61	0.0273	0.0111	0.00	0.0117	0.0577
PI	2,154	33	65	0.0223	0.0092	0.00	0.0091	0.0488
LD	2,276	33	68	0.0359	0.0120	0.00	0.0182	0.0703
ADHD	443	33	13	-0.0171	0.0100	0.81	-0.0394	0.0274
ID	652	33	19	0.0275	0.0187	0.03	-0.0007	0.0807
TBI	132	27	5	-0.0727	0.0292	0.86	-0.1419	0.0669
ASD	436	33	13	0.0242	0.0229	0.12	-0.0128	0.0922
MI	3,073	33	92	0.0193	0.0074	0.00	0.0087	0.0405
SA	577	31	18	0.0282	0.0207	0.04	-0.0027	0.0851

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.16 ICC Estimates of Conditional Model M4 for Outcome Measure Y2

Model 4	Total Sample Size N	Number of Groups	Within Group Size n_0	ICC Estimate	SE of ICC Estimate	p-value	Lower Bound of ICC	Upper Bound of ICC
Overall Sample	11,819	33	356	0.0237	0.0059	0.00	0.0144	0.0424
Work Experience								
No	8,821	33	266	0.0135	0.0042	0.00	0.0074	0.0259
Yes	2,998	33	90	0.0458	0.0126	0.00	0.0261	0.0840
Significance Disability								
No	1,233	33	36	0.0872	0.0246	0.00	0.0473	0.1544
Yes	10,586	33	319	0.0184	0.0050	0.00	0.0108	0.0337
Disability Type								
VI	87	29	3	-0.0015	0.0808	0.49	-0.1810	0.2454
HI	1,989	32	61	0.0293	0.0115	0.00	0.0130	0.0610
PI	2,154	33	65	0.0227	0.0093	0.00	0.0093	0.0495
LD	2,276	33	68	0.0348	0.0117	0.00	0.0175	0.0684
ADHD	443	33	13	-0.0216	0.0082	0.87	-0.0424	0.0201
ID	652	33	19	0.0240	0.0178	0.05	-0.0030	0.0753
TBI	132	27	5	-0.0754	0.0286	0.87	-0.1436	0.0630
ASD	436	33	13	0.0235	0.0227	0.12	-0.0132	0.0912
MI	3,073	33	92	0.0193	0.0074	0.00	0.0087	0.0406
SA	577	31	18	0.0283	0.0207	0.04	-0.0026	0.0853

Note1. VI=Visual Impairments or Blindness; HI=Hearing Impairments or Deafness; PI=Physical Impairments; LD=Learning Disabilities; ADHD= Attention Deficit Hyperactivity Disorder; ID=Intellectual Disability; TBI= Traumatic Brain Injury; ASD=Autism Spectrum Disorder; MI=Mental Illness; SA=Substance Abuse.
Note2. P-value=0.00 indicates that the level of significance is below 0.01 (i.e., p<0.01).

Table 5.17 Auxiliary Information of ICC Estimates for Outcome Measure Y2

Modeling Structure	ICC Estimate	Between Group Variance	Within Group Variance	Design Effect (DE)	η_B^2	η_W^2	R_B^2	R_W^2
Model 1 (M1)	0.0211	2,275.62	105,264.82	8.4907	NA	NA	NA	NA
Model 2 (M2)	0.0236	2,304.53	95,000.22	9.3782	1.0127	0.9025	-0.0127	0.0975
Model 3 (M3)	0.0217	2,282.94	102,665.47	8.7037	1.0032	0.9753	-0.0032	0.0247
Model 4 (M4)	0.0237	2,305.32	94,718.63	9.4137	1.0131	0.8998	-0.0131	0.1002

Note1. The ICC estimate for M1 represents the unconditional ICC quantity, while the ICC's for M2-M4 show the conditional ICC quantity.
Note2. Relative efficiency measures for ICC estimates between unconditional and conditional models (M1 versus M2-M4) are η_B^2, η_W^2, R_B^2 and R_B^2.

Table 5.18 Evaluation of Bootstrap ICC Estimates for Outcome Measure Y2

Number of Group	Within Group Size	Bias	MSE	Hits
5	50	0.0175	0.0013	0.20
5	100	0.0162	0.0007	0.37
5	150	0.0156	0.0006	0.45
15	50	0.0154	0.0005	0.47
15	100	0.0153	0.0003	0.51
15	150	0.0148	0.0003	0.66
25	50	0.0152	0.0004	0.52
25	100	0.0147	0.0003	0.75
25	150	0.0139	0.0002	0.86

Note1. Bias is defined as the mean difference between Bootstrap ICC and True ICC.
Note2. MSE is the mean squared error difference between Bootstrap ICC estimates.
Note3. Hits shows the proportion of Bootstrap ICC estimates successfully lying within the 95% confidence interval of True ICC.

CHAPTER 6

CONCLUSION & DISCUSSION

6.1 Summary of the Results

The proposed method for ICC estimation and inference is based on the real-world data set of RSA-911, where the usable samples are those individuals with disabilities served in the Michigan Rehabilitation Services Programs in FY 2015 (n=11,819). To address the research questions of the study, the two-level multilevel modeling approach to the cluster-randomized design data structure, is used to fit the data simulations, where individual subjects are at the level 1 (i.e., the average within cluster size is 356 per unit) and rehabilitation offices are at the level 2 (i.e., there are 33 of vocational rehabilitation office structures statewide in Michigan).

There are two types of multilevel modeling in data simulations: (1) unconditional model (Model 1); and (2) conditional models (Models 2-4). To evaluate which multilevel modeling structures match better with which sampling schemes, a bootstrap resampling procedure is adopted in data simulation and analysis, to compare the ICC estimates between population (Research Question 1) and subsample (Research Question 2) models, in terms of statistical properties of accuracy and precision on ICC estimation and inference (Research Question 3).

(a) Research Question 1 for Outcome Measure Y1 (see Tables 5.7-5.10)

For overall sample of competitive employment, the ICC estimate on average is about 0.01 (SE=0.003, $p<0.01$). Given by work experience (i.e., no work experience, in particular), the ICC estimate is inflated slightly (i.e., 0.002) but so is the standard error

(i.e., 0.002), comparing with the overall sample. Given by disability significance (i.e., no disability significance, in particular), the ICC estimate is inflated more (i.e., 0.02) and so is the standard error (i.e., 0.01), comparing with the overall sample. Given by disability type, the ICC estimate is inflated most (i.e., 0.05) for ASD, followed by LD (i.e., 0.02), HI (i.e., 0.01), PI (i.e., 0.01), and MI (i.e., 0.01). Also note that VI has the highest ICC (i.e., about 0.07), but the estimate is not significant at the level of 0.05, due to small sample size (i.e., total sample size is 87 across 29 office units).

(b) Research Question 1 for Outcome Measure Y2 (see Tables 5.13-5.16)

For overall sample of quality employment, the ICC estimate on average is about 0.02 (SE=0.005, $p<0.01$). Given by work experience (i.e., having work experience, in particular), the ICC estimate is inflated to some extent (i.e., 0.02) and so is the standard error (i.e., 0.006), comparing with the overall sample. Given by disability significance (i.e., no disability significance, in particular), the ICC estimate is inflated much (i.e., 0.06) and so is the standard error (i.e., 0.02), comparing with the overall sample. Given by disability type, the ICC estimate for LD is inflated most (i.e., 0.01) followed by SA (i.e., 0.01), HI (i.e., 0.01), and PI (i.e., 0.001). Also note that the ICC estimate for MI is relatively lower than the overall sample by about 0.002.

(c) Research Question 2 for Outcome Measure Y1 (see Tables 5.11-5.12)

As for examination of bootstrapping ICC estimates (repetitions=100) for competitive employment in the different sampling scenarios, it provides important sampling design information about hierarchical modeling with the full set of covariates of individual characteristics and rehabilitation services. With an average cluster sample size (e.g., the

number of clusters is about 10-15), the mean bias is about 0.005, MSE is about 0.0002, the proportion of successful hits is about 70%. With an average level of subject sample size (e.g., the number of subjects is around100), the mean bias is about 0.0053, MSE is about 0.0002, the proportion of successful hits is close to 60%. That is, the within-cluster subject size also plays an auxiliary role in quality of ICC estimation and inference, while the between-cluster sample size determines overall quality of ICC estimates.

In general, with large cluster samples (e.g., cluster size is 15-25) and average within-cluster samples (e.g., within-cluster size is 100-150), the ICC estimation and inference can result in effective performance in terms of accuracy and precision; on the other side, with a smaller cluster size (e.g., 5 or below) or a smaller within-cluster sample size (e.g., 50 or below), the ICC estimate is susceptible to be less reliable and more biased in the hierarchical generalized linear modeling framework for a binary outcome measure.

(d) Research Question 2 for Outcome Measure Y2 (see Tables 5.17-5.18)

As for examination of bootstrapping ICC estimates (repetitions=100) for quality of employment in the different resampling scenarios, it provides crucial sampling design information about multilevel modeling with the full set of covariates of individual characteristics and rehabilitation services. With an average cluster sample size (e.g., the number of clusters is about 10-15), the mean bias is about 0.015, MSE is about 0.0003, the proportion of successful hits is about 55%. With an average level of subject sample size (e.g., the number of subjects is around 100), the mean bias is also about 0.015, MSE is about 0.0004, the proportion of successful hits is close to 55% as well. That is, the within-cluster size also plays a supplemental role in ICC estimation and inference, while the between-cluster size still can boost effective performance of ICC estimates.

In general, with large cluster samples (e.g., cluster size is 15-25+) and average within-cluster samples (e.g., within-cluster size is 100-150+), the ICC estimation and inference can result in effective performance in terms of accuracy and precision; on the other hand, with a smaller cluster size (e.g., 10 or less) or a smaller within-cluster sample size (e.g., 50 or less), the ICC estimate is prone to be less consistent and more biased in the hierarchical linear modeling framework for a continuous outcome measure.

(e) Research Question 3 for Outcome Measure Y1 (see Tables 5.11-5.12)

As for auxiliary statistics of the ICC estimates for competitive employment, the unadjusted ICC is about 0.01 (DE=4.44), while the adjusted ICC is also about 0.01 (DE=4.67). The unconditional model is used as a baseline to measure relative efficiency of between- and within-group variances for ICC estimates in conditional models. Among the three competing conditional models (Models 2-4), Model 3 (the one with a covariate set of service information) has the most decrease of 8.06% of within-group variation as well as a significant increase of 4.17% in between-group variation, comparing with the baseline model (Model 1). Note that both Model 2 (demographic model) and Model 4 (full model) have similar performance that result in a decrease of 3.05% of within-group variation and 0.00% of change in between-group variation, comparing with the baseline.

(f) Research Question 3 for Outcome Measure Y2 (see Tables 5.17-5.18)

As for auxiliary statistics of the ICC estimates for quality of employment, the unadjusted ICC is about 0.02 (DE=8.49), while the adjusted ICC is also about 0.02 (DE=9.17). The unconditional model is used as a baseline to measure relative efficiency of between- and within-group variances for ICC estimates in conditional models. Among the three

competing conditional models (Models 2-4), Model 4 (the one with the full covariate set of demographics and services) and Model 2 (the one with a covariate set of demographic information) has the most decrease of about 9.88% of within-group variation as well as a slight increase of 1.29% in between-group variation, comparing with the baseline model (Model 1). Note that Model 3 (service model) has relatively ineffective performance that result in a modest decrease of 2.47% of within-group variation and a tiny increase of 0.32% in between-group variation, comparing with the baseline model.

6.2 Implications

(a) Statistical perspectives on the ICC estimation and inference

The intraclass correlation coefficients (ICC) at experimental designs has been one of the oldest statistical measures since Sir RA Fisher invented it last century (Fisher, 1925a). Like Pearson's correlation coefficient, it has been used as one of the most popular and important tools in scientific inquiries including educational and social research. In a theoretical perspective, both correlation coefficient and intraclass correlation share mathematical similarities and features. For example, ICC can be used to measure the level of similarity or resemblance within a group of measurements (e.g., students in a classroom or school), and the general formula of intraclass correlation can be written by a very similar form of Pearson's product moment correlation coefficient. Fisher (1925a) also pointed out that the ICC can be geometrically equivalent to the overall Euclidean distance between the paired samples on the standardized scale (see Figures

2.2 and 2.3 as examples). In terms of effect size measures, both correlation and ICC can determine the effect size magnitude of a studied phenomenon of interest; in particular, the ICC can show the amount of total variance explained by between-group variation in an experimental design model (e.g., hierarchical linear models), and that it is another form of the squared correlation (R-squared) in analysis of variance models which accounts for the true proportion of outcome variance across different clusters.

One research gap in methodology for ICC estimation and inference is about the testing statistic and its related sampling distribution. This study aims to address that important issue by developing the mathematical foundations of the ICC estimator at a hierarchical design (e.g., cluster randomized trials). Donner & Koval (1980a) derived maximum likelihood estimator (MLE) of the intraclass correlation using variance component in analysis of variance (ANOVA) models. Since the traditional method (Fisher's approach) requires distributional assumptions (based on multivariate normal theory), it is analysis of variance (ANOVA) that provides an alternative estimator of intraclass correlation (by relaxing the multi-normal assumptions) via classical ANOVA. This study finds the estimator of intraclass correlation via the Fisher's definition, but further extends it to utilize relevant information in the ANOVA table by Donner's approach (i.e., utility of between- and within-group variation) for developing a general statistical framework for the ICC in the multilevel structure (i.e., a flexible approach to either a balanced design with equal group size or an unbalanced "natural" design with unequal group size). It is noteworthy that the approximate group size (or the average within-group size by Donner & Koval, 1980a) is a key in an unbalanced design case for computation of the proposed ICC estimator (see Figure 2.4 as an illustration).

As for statistical testing of the proposed ICC estimator ($\rho_{Intraclass}$), this study suggests the use of *F*-distribution (with $k-1$ and $N-k$ degrees of freedom) and *F*-testing statistic (based on ANOVA) for determining if the null or alternative hypothesis of the magnitude of effects is true at the chosen level of significance (i.e., $H_o: \rho_{Intraclass} = 0$ vs. $H_a: \rho_{Intraclass} > 0$). A significant *F*-testing statistic value implies that members of the same group tend to be more alike and similar with respect to the attribute or characteristic in question than those from different groups (i.e., if within-group subjects are perfectly homogeneous, or equivalently $\sigma_W^2 = 0$, then it implies $\rho_{Intraclass} = 1$).

As for a $100(1-\alpha)\%$ confidence interval on the ICC, this study provides the formulas for the corresponding interval for an ICC estimand (i.e., the true proportion of variance accounted for by a grouping factor of interest in a hierarchical design). Also, it is notable to be pointed out that the lower confidence limit on an ICC interval estimate could be negative using the proposed method, especially when a small sample size or large measurement error occurs in hierarchical modeling; but since ICC is normally non-negative in anyway by the mathematical definition, it is a common practice to replace the negative lower bound with "zero" for a post-hoc adjustment (Hays, 1994).

As for the variance of the proposed ICC estimator, this study uses the MLE approach (multivariate normality in a large sample theory) by Donner & Koval (1980a) to obtain the standard error of the ICC estimate. It is interesting to note that the MLE of ICC is statistically equivalent to the Pearson's product moment correlation (i.e., a quick shortcut solution for the ICC estimation) for a balanced design in hierarchical modeling; but for an unbalanced design, the MLE of ICC needs to be solved by a different approach – either numerical optimization via multivariate log-likelihood by

Donner & Koval (1908b) or using invariance property of MLEs by Karlin et al. (1981).

The proposed theoretical framework for ICC in the study is mainly inspired by Hedges' approach (Hedges & Hedberg, 2007) that uses ICC via hierarchical modeling to collect the clustering information of variance components in cluster randomized trials (CRT). Nowadays CRT have become more and more popular in education and social studies for some practical reasons that RCT (randomized control trial) is too expensive for the assignment of each individual subject, whereas CRT is more economical by dealing with an entire intact group at one time. Since ICC has been considered as an ancillary statistic to provide design effect (DE, or variance inflation factor, VIF) for statistical planning in multilevel design, ICC can play a key role in effectively quantifying the amount of inherent clustering effects for a CRT survey study (Hedges et al., 2012; Hedges & Hedberg, 2013). It is important to note that clustering design (CRT) has more total variation (i.e., cluster-to-cluster plus within-cluster variance) than simple random sampling (RCT) by a factor of DE (that is why it is also called VIF).

As for experimental design with a binary outcome (e.g., a dichotomous variable), the proposed ICC estimator in this study is derived by using the hierarchical generalized linear modeling framework (HGLM; Raudenbush & Bryk, 2002). It is conventional (and also mathematically convenient) to use a constant variance (i.e., $\pi^2/3$) as within-group variance based on the standard logistic distribution (location $= 0$ and shape $= 1$), whereas this strong assumption of "holding within-group variance as a constant" often is not met in real world, so the recommended modification strategy from the study is to introduce a more flexible estimation procedure by incorporating a data-driven within-group variance via HGLM for the proposed ICC estimation and inference.

Last but not least, the proposed ICC method is also connected with statistical planning in experimental design for sample size determination and power calculation, which is critical for researchers to conduct rigorous scientific investigations for detecting true effects at a desired effect size, statistical power, and significance level. Traditionally, the design and planning for sample and power calculations requires a classical restrictive assumption of simple random samples, which is not quite met for multilevel modeling. Hence, this study proposes a theoretical framework for the ICC estimator to circumvent such a shortcoming by taking into account heterogeneity in hierarchical structures of cluster samples (such as CRT). The proposed ICC estimation and inference is feasible via the use of between- and within-group variance in ANOVA of hierarchical linear modeling, and the testing statistic is based on F-distribution to serve a foundation for statistical inference of the ICC estimand in a multilevel design.

(b) Policy perspectives on the ICC estimation and inference

In behavior, educational, psychological and social research, cluster randomized design that assigns intact groups (e.g., classrooms or schools) to interventions, has been become more increasingly adopted in the era of evidence-based education and policy (Lingard, 2013). Since experimental design with such a cluster randomization is deemed as a hierarchical data structure (i.e., subjects nested within a cluster), statistical planning would require relevant information of ICC to account for clustering effects to achieve adequate power and collect sufficient sample. Through the real data set of

RSA-911 from U.S. Department of Education, this study provides a comprehensive analysis of ICC of employment outcomes (i.e., competitive employment and quality of employment measures) which are adjusted by covariates of interest (i.e., demographics and rehabilitation services) that can be used for statistical planning on CRT research (randomized trials or quasi experiments) in future education studies. In addition, this study also provides relative variance component information (i.e., between-group and within-group variation) that can be useful to understand which types of covariates should be involved in multilevel design for statistical planning and analysis.

In an era of evidence-based practice in rehabilitation counseling & education, researchers are more aware of incorporation of scientific inquiry for finding "best" ways to empower people with impairments through effective services (Chan et al., 2009). The recent legislation of The Workforce Innovation and Opportunity Act of 2014 (WIOA), state and federal VR agencies have to assist the target disability populations, to succeed in the today's jobs and prepare for tomorrow's labor markets in the global economy (WIOA Legislation, 2018). Thus, those rehabilitation counselors, educators, practitioners, and researchers all need to work together to adopt the new EBP paradigm to improve the quality of life for VR customers through rehabilitation services. Further, evidence-based best practices in rehabilitation counseling would significantly improve outcomes for people with disabilities by translating knowledge and making good decisions in VR (Leahy et al., 2009, 2010, 2014a, 2014b).

The use of EBP has become a new standard to conduct effective research and gather reliable data for improving practices and outcomes (Eignor, 2013). Rehabilitation counselors and practitioners can integrate best EBP research evidence with clinical

judgement expertise, to make better decisions that enhance the outcomes, so the EBP can provide a significant improvement of knowledge translation in practice (Kosciulek, 2010). So, not only does EBP provide the foundations incorporating scientific evidence as well as clinical judgement expertise, to make best decisions about interventions, services, or treatments for people with disabilities, but EBP also assists rehabilitation counselors to identify relevant "evidence" of literature, assess "available" information such as the RSA-911 data, and constitute "best available evidence" on rehabilitation services for people with disabilities. So, under the data-driven framework with RSA-911, this study provides the proposed method of ICC in multilevel data structure (i.e., individual subjects are on level 1 and rehabilitation office units are on level 2) that can help rehabilitation counseling researchers better understand the target population of people with disabilities when conducting CRT design and analysis for gathering relevant information of EBP by taking into account of the clustering effects via the ICC (w.r.t. the office units statewide) in the RSA-911 data using hierarchical linear models.

Hierarchical data structures are ubiquitous in education and social studies (Raudenbush & Bryk, 1992). In rehabilitation counseling & education, for example, clients are nested into field office structures, which are also nested into local districts, and local districts are nested into states, and states are nested into regions, and so on. So, it is important to take into account all these multilevel structures and related topological relationships by using the hierarchical modeling framework for design and analysis.

As for the origin of the RSA-911 data, Rehabilitation Services Administration Case Service Report (RSA-911for short) is the state vocational rehabilitation agencies collect and report summary data in a federally mandated format. The RSA-911 provides

researchers a good resource for gathering evidence of EBP (Schwanke & Smith, 2004). Through data mining and deep learning of the RSA-911 data, rehabilitation researchers can study complex issues to build EBP for people with disabilities (Pi & Thielsen, 2011), and they can also explore big data of RSA-911 to examine what and how factors (e.g., variables in the individual level or office level) affect VR outcomes in which type disability groups. Therefore, rehabilitation researchers can exploit the RSA-911 data to develop EBP (either by CRT design or quasi-experimental analysis), in particular, for conducting individual-level and employment-related interventions, finding effective strategies for VR outcome improvement, and best VR practices to achieve successful outcomes for individuals with disabilities (Fleming et al., 2013; Pi, 2006).

In previous literature of multilevel modeling using RSA-911 data, Alsaman & Lee (2017) examined the cross-sectional inter-relationships between contextual factors (unemployment rates at the state level), individual factors (demographic background at the person level) , and employment outcomes (competitive employment of a binary measure) for the youth population with disabilities using the 2-level hierarchical generalized linear modeling (HGLM) framework. Chan et al. (2014) studied the impact of the economic recession on VR employment by controlling for the contextual factor of unemployment rate in each state, where the 2-level HGLM approach is applied. Pi (2006) used the 2-level HLM method with the micro- and macro-level factors related to VR outcomes, to study relationship between predictors across levels in the VR.

One knowledge gap in rehabilitation counseling research and literature for the ICC applications is about how to incorporate relevant ICC information into design and analysis using the RSA-911 data by taking into account the clustering effects via the

ICC and the related DE estimates using multilevel models. This study aims to address that important issue by examining the ICC values via HLM and HGLM. The proposed framework for ICC estimation and inference in the study is examined via the real-life data set of RSA-911, where the target samples of interest are people with disabilities in Michigan Rehabilitation Services in FY 2015 (n=11,819). To address the ICC-related research questions of the study, the two-level HLM and HGLM approach to the CRT (or clustering RCT) type of study design is used to conduct the simulations, where person subjects are on the level and cluster units are on the level 2. Results show that: (i) the overall ICC estimate for both outcome measures (competitive employment and quality employment) tends to be low (0.01 and 0.02, respectively), implying that the clustering effects of rehabilitation office structures cannot capture much total variation in the RSA-911 data; (ii) rehabilitation services play a bigger role than individual characteristics in accounting for total variation in the both employment outcome measures; (iii) previous work experience, significance of disability, and type of disability (i.e., covariates for subgroup analysis) can affect outcome measures, but also they show differences in the ICC estimates, which indicates that researchers should pay attention to those groups with a high ICC value when conducting a CRT design study; (iv) should a CRT experiment be conducted, the recommended minimum cluster samples are about 10-15 units, and person samples are about 100-150 subjects, for attaining sufficient quality sample in analysis. It is interesting to notice that the average (unadjusted) ICC estimates in the simulation study are comparable to those psychological mental health data in school-based intervention designs in which ICCs range from 0.01 to 0.05 (Murray & Short, 1995), although they are relatively lower than the standards of 0.05-0.15 based on education data in reading and

mathematics across Grades K-12 (Bloom et al., 1999, 2007; Hedges & Hedberg, 2007; Schochet, 2008). The low ICC is an indicator of small clustering effects in the multilevel design and analysis, but the effective sample size (i.e., a total sample size divided by design effect) is inflated to a certain degree, meaning the bottom line (minimum sample size) is risen to maintain high statistical power and low standard error given by the same model.

6.3 Limitations of the Study

There are four limitations in the study.

(1) Of the different types of effect magnitude measures for the correlation ratio (ρ), the intraclass correlation (ICC; $\rho_{Intraclass}$) is a parametric estimator in ANOVA via HLM to quantify the true proportion of total variance ($\sigma_Y^2 = \sigma_a^2 + \sigma_e^2$) accounted for in the outcome. Although the underlying ANOVA framework in HLM suggests the total variance consists of two independent variance components (i.e., both always be a positive real number) – group variance (σ_a^2) and error variance (σ_e^2), an unbiased estimate of group variance may be failed and found $\hat{\sigma}_a^2 = 0$, especially when MSE ($\hat{\sigma}_e^2$) is greater than or equal to MSA ($\hat{\sigma}_a^2$) (Hayes, 1994). As a consequence, the ICC estimate value is forced to become zero, which would be shown as a warning of "estimation failure" from the command for HLM or HGLM in statistical software (like *lmer* or *glmer* from the package of *lme* or *lme4* in R). In this case of estimation failure in HLM, model modification is suggested to remedy the situation that there is more within-group error variation than between-group variation, i.e., $\hat{\sigma}_a^2 \leq \hat{\sigma}_e^2$, in ANOVA via HLM (Raudenbush & Bryk, 2002).

(2) In the simulation using the RSA-911 data, there have other options to build a different multilevel design and analysis for ICC estimation and inference. In this study, a two-level hierarchical design structure (i.e., individuals are at the level 1, and offices at the level 2) is fitted by HLM and HGLM to find the unadjusted ICC (by the unconditional model without any covariates) and adjusted ICC (by the conditional model with covariates). On the other side, alternative modeling choice can be the latent variable modeling (LVM) approach to investigate the multilevel data of RSA-911. Austin & Lee (2014) built a structural equation model (SEM) of VR services via RSA-911, to study predictors of employment outcomes in VR for people with intellectual and co-occurring psychiatric disabilities. And Alsaman & Lee (2017) examine the relationships between contextual factors, individual factors, and employment outcomes of transition youth with disabilities in VR using the RSA-911 data in by the 2-level HGLM (individuals are on Level 1, and states are on Level 2). Since the current study does not use latent factors in the HLM and HGLM framework due to the limitation of HLM and HGLM modeling structure, the alternative LVM approach can provide a holistic modeling structure with latent constructs and manifest variables both at the same time to study latent factor structures of interest (Raykov & Marcoulides, 2006). In the VR context, SEM can also be used to examine important predictive associations between individual characteristics, rehabilitation services, and employment outcomes, while HLM is essentially to provide the overall "big picture" of ICC in multilevel design (such as CRT).

(3) In the simulation study using the RSA-911 data, it does not consider any interactions at the person level or the office level (e.g., demographic variables and service indicators at the level 1, or their group means at the level 2) due to statistical simplicity for simulations, but they may exist two-way interactions somewhat between those individual

characteristic and rehabilitation service variables. For example, age group (X3) can be related to education (X5), rehabilitation services (X6-X8) for both employment outcome measures (Y1 and Y2), according to the sample correlation structures of all predictors in hierarchical analysis (see Tables 5.4 and 5.5). With those important two-way interactions added into HLM and HGLM, the ICC estimation and inference can be influenced to some degree due to between- and within-group variation affected by new predictors (those important two-way interactions) in the HLM and HGLM model. Theoretically, after adding those significant predictors in an HLM or HGLM model, MSE (within-group variation) would be decreasing to some extent, and the new ICC could be increasing to a certain degree, comparing with the old ICC (based on the baseline model without newly added important two-way interactions).

(4) The ICC estimation would require a minimum total sample size (N), the number of groups (k), and within-group size (n_0). If one of the criteria (i.e., N, k, and n_0) is not met, it is very likely to obtain an invalid ICC estimate value (either the ICC estimate is a negative value or zero, or the lower bound of confidence interval is not positive at all). For example, the lower bound of ICC confidence interval (CI) for visual impairments (VI) on Y1 under Model 1 is not valid (see Table 5.7), due to the small total sample size ($N = 87$) and within group size ($n_0 = 3$); similarly, the lower bound of ICC confidence interval (CI) for visual impairments (VI) on Y2 under Model 1 is not valid either (see Table 5.13) and so is the ICC estimate negative, due to again the small total sample size ($N = 87$) and within group size ($n_0 = 3$). The threshold of sample size criteria for ICC estimation and inference would need future research to determine the minimum sample size for statistical analysis in HLM and HGLM. From the simulations, the rule of thumb is total sample size (N) greater than 600 and within group size (n_0) larger than 20, given by the number of groups about 30. In other

words, the quick formula is $N = n_0 \times k$, where N is total sample size, k is the number of groups, n_0 is within group size; and the simulation finding in the study (based on the RSA-911 data) suggests that the sample size criterion $N \geq 30 \times n_0$, or $n_0 \geq N/30$, would assure the ICC estimation and inference is more likely to get a valid and reliable result in the case of CRT (or cluster RCT) via the HLM and HGLM framework using the RSA-911 data.

6.4 Future Research

Future work should address the following five potential issues that have not been fully addressed in this study.

First, as for the traditional approach to ICC estimation, the practical method is based on a two-level multilevel structure (e.g., the person level is defined as Level 1, and the group level is defined as Level 2), where the ICC estimation is to utilize relevant information from the ANOVA table including the source of both between- and within-group variation in the HLM and HGLM framework. For more complex multilevel structure in CRT experiments (e.g., 3-level and 4-level hierarchical design), ICC estimation (using variance component decomposition in ANOVA via HLM) has been discussed (Hedges et al., 2012; Hedges & Hedberg, 2013), but ICC inference (hypothesis testing by confidence interval and p-value) has not been done yet for complex 3-level or 4-level multilevel models. For this development, one statistical challenge and difficulty is to find out an effective way to quantify standard error of ICC (based on the pooled weighted variance of ICC across different levels) in complex multilevel design via the HLM or HGLM framework, or to extend the 2-level

multilevel framework in the study to 3- or 4-level HLM or HGLM by using multiple comparison procedures for ANOVA (e.g., Bonferroni's correction method, and Benjamini-Hochberg procedure) to control for familywise Type I error rate or the overall false discovery rate (i.e., the probability of making one or more Type I errors or false discoveries when performing multiple hypotheses tests).

Second, complex data integration (or data fusion) has become an important issue in the big-data era with today's technology, and researchers may look into multiple sources of large-scale complex data sets (or data platforms) to conduct interdisciplinary studies. For example, it would be interesting to integrate the RSA-911 data with a set of covariates from Census data for a comprehensive research investigation about how the between- and within group variation sources are varied by the ICC estimates, in terms of statistical effectiveness perspectives for design and analysis, for statistical estimation and inference at each level of multilevel modeling across different data platforms. In such a way, multilevel design models are inherently nested at each level in different data platforms (note: data platform can be viewed as a "block" and treated as an additional level in the HLM and HGLM framework) Given by this complex design structure (multiple data platforms), it would be interesting to study how statistical planning can be conducted for power and sample size calculations, and what ICC estimates are varied (using sensitivity analysis) to a point in different platforms.

Third, covariate adjustment is an important technique in statistical modeling to take into account the confounder effects in a model (HLM or HGLM). In the complex multilevel design (i.e., more than two levels in hierarchal models), it would be interesting to understand how covariate adjustment (with or without subgroup analysis or stratification) affects adjusted ICC estimation and inference. In the study (as the case of 2-level hierarchical design),

simulations show covariate adjustment (with stratification by a "breaking" variable) can improve the ICC estimates to some extent, yet in some cases (especially for a small total sample size or within-group sample size) the ICC estimation and inference cannot work at all (i.e., estimation failure). Therefore, it would be important to find out how to develop the remedial strategy for statistical adjustment and stratification in complex multilevel design via HLM and HGLM, and what type of statistical centering or standardizing procedures can be used to modify (or "customize") covariate adjustment (e.g., group and grand centering or standardizing) at each level to make "adjusted" ICC estimation and inference more accurate and precise by accounting for the localized multilevel substructure adjusted by covariates.

Fourth, this study considers only one-year data (FY 2015) of RSA-911 for simulations to testify the proposed method of ICC estimation and inference. It would be interesting to study the statistical properties of ICC by extending the current framework to a complex multilevel structure such as longitudinal design across multiple years or cross-cohort design with multiple year data resources. In this type of complex multilevel modeling structure (e.g., longitudinal analysis in HLM and HGLM), the variance-and-covariance structure (i.e., a random component on the "time" factor in ANOVA) needs to be considered (e.g., compound symmetry for homogeneous data, and autoregressive structure for heterogeneous data) so as to take into account the correlation structure across different time periods or cohorts. In addition, it would be interesting to use multiple year data sets of RSA-911 to verify statistical performance of ICC estimation and inference in terms of consistency and efficiency.

Lastly, missing data analysis is a common issue in statistics. Although the listwise procedure (i.e., only include complete data, but exclude those subjects with any incomplete information) is a convenient way to deal with missing data, it would often lose much

statistical information and compromise statistical power in analysis (e.g., HLM or HGLM). Hence, it would be important to study how to cope with missing values (assuming missing at random) in a multilevel design data structure for ICC estimation and inference, and what remedial procedures (EM or multiple imputation for discrete or continuous variables) can be applied to improve the ICC estimation process via sensitivity analysis in HLM or HGLM. For the proposed method of ICC with a full complete data, the simulation results suggest the total sample size needs to be greater than 1,500 and within group sample size larger than 100 (over 15 groups). Nevertheless, the guidelines need to be adjusted for incomplete data case.

6.5 Conclusion

In conclusion, this study provides a comprehensive methodology for intraclass correlation (ICC) estimation and inference using the hierarchical mixed modeling framework. The proposed methodology for ICC estimation and inference incorporate the analysis of variance (ANOVA) approach to the development of the ICC estimator and its inferential statistic of the pivotal quantity of the ICC estimand for deriving the sampling distribution (F-distribution) to test ICC as well as construct confidence interval on ICC. The proposed statistical procedures for ICC estimation and inference can be easily used and applied in any large-scale or small-scale data sets, whereas small total sample size and small within group size and missing data are limitations can affect the results of ICC estimates to a certain degree in terms of precision and accuracy. More research study is needed to better understand the ICC in complex multilevel design structures.

APPENDICES

APPENDIX A: Definitions of the VR Variables in RSA-911

The following are the definitions of VR variables, according to the manual of RSA-911 (Policy Directive of RSA-PD-16-04 for Revision of RSA-PD-14-01; https://www2.ed.gov/policy/speced/guid/rsa/subregulatory/pd-16-04.pdf).

This appendix section includes three tables: (1) VR services are shown in Table A.1; (2) demographic backgrounds are listed in Table A.2; and (3) rehabilitation outcomes are given in Table A.3.

Table A.1. List of the Definitions of VR Service Variables Used in the Study

Rehabilitation Service	RSA Definition
Job Placement Assistance	This is a referral to a specific job resulting in setting up a job interview and obtaining a job on behalf of a customer (1=received; 0=not received)
On-the-Job Supports	Services such as job coaching, follow along services to assist a customer adjust to the job and become stable to enhance job retention (1=received; 0=not received)
Rehabilitation Technology	The application of rehabilitation engineering, assistive devices, technologies, or services, to meet the needs and address the barriers (1=received; 0=not received)

Table A.2. List of the Definitions of VR Demographic Variables Used in the Study

VR Demographics	RSA Definition
Age	Indicate age when he or she is applied for VR services (continuous measure)
Gender	Indicate an individual is male or female (1=male; 0=female)
Minority (Non-White)	Indicate an individual's race/ethnicity if s/he is minority (including Black, Native, Asian, Pacific Islander, and Hispanic) or not (White) (1=minority; 0=non-minority)
Social Security Benefits (Insurance Benefits)	Indicate if an individual receives Social Security Disability Insurance (SSDI) or Supplemental Security Income (SSI) (1=received; 0=not received)
Employment Status at Application (Previous Work Background)	Employment status of the individual at application (1=employment; 0=not employed)
Type of Disability	Individual's primary physical or mental impairment includes: blindness/visual impairment, deafness/hearing impairment, physical or orthopedic/neurological impairment, LD, ADHD, intellectual disability (ID), TBI, autism, mental illness (MI), substance abuse (SA) (categorical/qualitative measure)
Level of Education	Level of education the individual had attained includes: elementary/secondary education, special education, high school graduate or equivalency certificate (GED), college or above (categorical/ordinal measure)
Significance of Disability	Whether the individual was considered a person with a significant disability or a most significant disability during VR (1=yes; 0=no)

Table A.3. List of the Definitions of VR Outcome Variables Used in the Study

Rehabilitation Outcome	RSA Definition
Rehabilitation Outcome	Individual exited the VR program either with or without an employment outcome after receiving services (1=exited with an employment; 0=exited without an employment)
Competitive Employment	Employed either at or above minimum wage in integrated setting (1=yes; 0=no)
Weekly Earnings (or Quality of Employment)	The approximate amount of money earned in a typical week (continuous measure)

APPENDIX B: Descriptive Data Statistics

Table B.1 Descriptive Summary of Usable Sample by Office Level in Michigan (*n*=11,819)

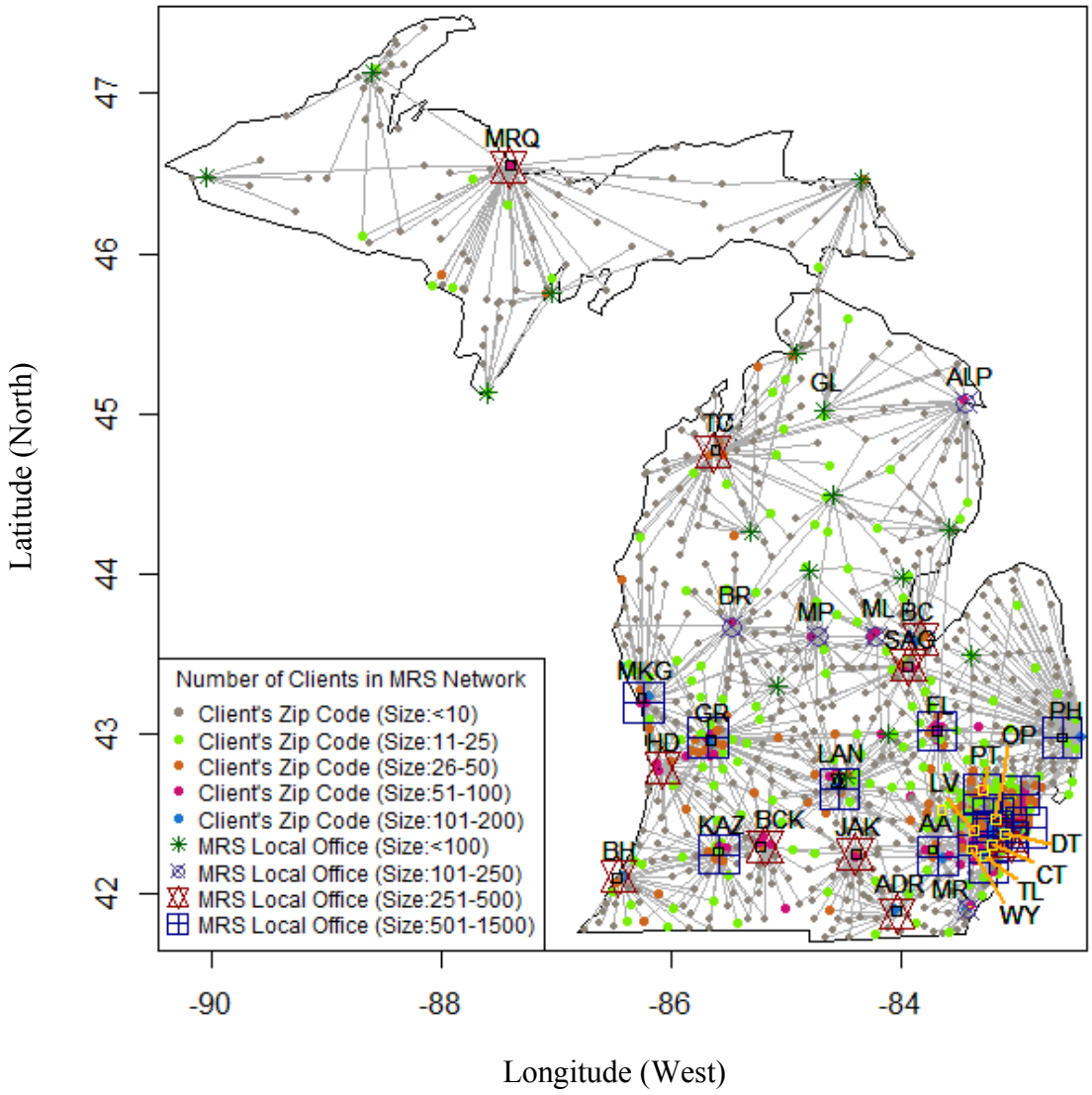
Office Unit	Frequency	Percentage
Adrian Unit	244	2.06%
Alpena Unit	160	1.35%
Ann Arbor Unit	484	4.10%
Battle Creek Unit	298	2.52%
Bay City Unit	281	2.38%
Benton Harbor Unit	289	2.45%
Big Rapids Unit	175	1.48%
Clinton Township Unit	732	6.19%
Detroit Fort Street Unit	320	2.71%
Detroit Grand River Unit	423	3.58%
Detroit Hamtramck Unit	463	3.92%
Detroit Mack Unit	332	2.81%
Detroit Porter Unit	421	3.56%
Flint Unit	418	3.54%
Gaylord Unit	174	1.47%
Grand Rapids Unit	764	6.46%
Holland Unit	335	2.83%
Jackson Unit	163	1.38%
Kalamazoo Unit	345	2.92%
Lansing Unit	631	5.34%
Livonia Unit	441	3.73%
Marquette Unit	405	3.43%
Midland Unit	125	1.06%
Monroe Unit	200	1.69%
Mt. Pleasant Unit	136	1.15%
Muskegon Unit	366	3.10%
Oak Park Unit	540	4.57%
Pontiac Unit	416	3.52%
Port Huron Unit	485	4.10%
Saginaw Unit	281	2.38%
Taylor Unit	213	1.80%
Traverse City Unit	377	3.19%
Wayne Unit	382	3.23%
Total	11,819	100.00%

Note. There are 33 offices located statewide in Michigan, serving the target population of people with disabilities of N=17,633 in FY 2015. Of the target samples, the usable sample size is n=11,819 for data analysis in the study and ICC calculations.

Table B.2. A Summary of the Geographic Information System of Office Units in Michigan

Latitude (N)	Longitude (W)	Abbreviation	MRS Unit
41.90	84.04	ADR	Adrian
45.06	83.43	ALP	Alpena
42.28	83.73	AA	Ann Arbor
42.30	85.23	BCK	Battle Creek
43.60	83.89	BC	Bay City
42.10	86.48	BH	Benton Harbor
43.70	85.48	BR	Big Rapids
42.31	83.21	CT	Clinton Township
42.38	83.10	DT	Detroit Fort Street
			Detroit Grand River
			Detroit Hamtramck
			Detroit Mack
			Detroit Porter
43.02	83.69	FL	Flint
45.03	84.67	GL	Gaylord
42.96	85.66	GR	Grand Rapids
42.78	86.10	HD	Holland
42.25	84.40	JAK	Jackson
42.27	85.59	KAZ	Kalamazoo
42.71	84.55	LAN	Lansing
42.40	83.37	LV	Livonia
46.55	87.41	MRQ	Marquette
43.62	84.23	ML	Midland
41.92	83.40	MR	Monroe
43.60	84.77	MP	Mt. Pleasant
43.23	86.26	MKG	Muskegon
42.47	83.18	OP	Oak Park
42.65	83.29	PT	Pontiac
42.98	82.60	PH	Port Huron
43.42	83.95	SAG	Saginaw
42.24	83.27	TL	Taylor
44.77	85.62	TC	Traverse City
42.28	83.39	WY	Wayne

Figure B.1 Spatial Network of Target Sample in Michigan by Hierarchical Structure



Note1. MRS represents the Michigan Rehabilitation Services Programs.
Note2. Client's Zip code indicates an individual's residence; each MRS local office is plotted on geometric graph according to the geographic information system (GIS) in Table B.2.

APPENDIX C: Glossary of Abbreviations

This glossary contains abbreviations, acronyms and some definition used in this study.

Table C.1 Glossary of Abbreviations

ANOVA	Analysis of Variance
ASD	Autism Spectrum Disorder
CSPD	Comprehensive System of Personnel Development
CTT	Classical Test Theory
EBP	Evidence Based Practice
ESRA	Education Sciences Reform Act
FY	Fiscal Year
GIS	Geographic Information System
HGLM	Hierarchical Generalized Linear Model
HLM	Hierarchical Linear Model
ICC	Intraclass Correlation Coefficient
ID	Intellectual Disability
IPE	Individualized Plan for Employment
LVM	Latent Variable Modeling
MI	Mental Illness
MLE	Maximum Likelihood Estimate
MRS	Michigan Rehabilitation Services
NCLB	No Child Left Behind
RCT	Randomized Control Trial
REML	Restrictive Maximum Likelihood
RSA	Rehabilitation Service Administration
SE	Standard Error
SEM	Standard Error Measurement
SEM	Structural Equation Model
TBI	Traumatic Brain Injury
VR	Vocational Rehabilitation
WIOA	Workforce Innovation and Opportunity

BIBLIOGRAPHY

BIBLIOGRAPHY

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences*. Upper Saddle River, N.J: Pearson Prentice Hall.

Alsaman, M. A., & Lee, C.-L. (2017). Employment Outcomes of Youth With Disabilities in Vocational Rehabilitation: A Multilevel Analysis of RSA-911 Data. *Rehabilitation Counseling Bulletin*, *60*(2), 98-107.

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*.

Anderson, T., & Shattuck, J. (2012). Design-based research: A decade of progress in education research? *Educational researcher*, *41*(1), 16-25.

Austin, B. S., & Leahy, M. J. (2015). Construction and validation of the clinical judgment skill inventory: Clinical judgment skill competencies that measure counselor debiasing techniques. *Rehabilitation Research, Policy, and Education*, *29*(1), 27.

Austin, B. S., & Lee, C.-L. (2014). A structural equation model of vocational rehabilitation services: Predictors of employment outcomes for clients with intellectual and co-occurring psychiatric disabilities. *Journal of Rehabilitation, 80(3),* 11-20.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The journal of the learning sciences*, *13*(1), 1-14.

Bartholomew, D. J. (1987). *Latent variable models and factors analysis*. Oxford University Press, Inc.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Bloom, H.S., Bos, J.M., & Lee, S.W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. Evaluation Review, 23(4), 445-469.

Bloom, H.S., Richburg-Hayes, L., & Black, A.R. (2007). Using Covariates to Improve Precision: Empirical Guidance for Studies that Randomize Schools to Measure the Impacts of Educational Interventions. Educational Evaluation and Policy Analysis, 29(1), 30-59.

Bolton, B. F., Bellini, J. L., & Brookings, J. B. (2000). Predicting client employment outcomes from personal history, functional limitations, and rehabilitation services. *Rehabilitation Counseling Bulletin*, 44(1), 10-21.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Australia: Thomson Learning.

Chan, F., Tarvydas, V., Blalock, K., Strauser, D., & Atkins, B. J. (2009). Unifying and elevating rehabilitation counseling through model-driven, diversity-sensitive evidence-based practice. *Rehabilitation Counseling Bulletin*, *52*(2), 114-119.

Chan, F., Bezyak, J., Ramirez, M. R., Chiu, C. Y., Sung, C., & Fujikawa, M. (2010). Concepts, Challenges, Barriers, and Opportunities Related to Evidence-Based Practice in Rehabilitation Counseling. *Rehabilitation Education*, *24*.

Chan, F., Wang, C. C., Fitzgerald, S., Muller, V., Ditchman, N., & Menz, F. (2016). Personal, environmental, and service-delivery determinants of employment quality for state vocational rehabilitation consumers: A multilevel analysis. *Journal of Vocational Rehabilitation*, 45(1), 5-18.

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational researcher*, *32*(1), 9-13.

Chan, F., Lee, G. K., Lee, E., Kubota, C., & Allen, C. A. (2007). Structural equation modeling in rehabilitation counseling research. *Rehabilitation Counseling Bulletin*, 57(1), 44-57.

Chan, J. Y., Wang, C. C., Ditchman, N., Kim, J. H., Pete, J., Chan, F., & Dries, B. (2014). State unemployment rates and vocational rehabilitation outcomes: A multilevel analysis. *Rehabilitation Counseling Bulletin*, *57*(4), 209-218.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Connelly, L. B. (2003). Balancing the number and size of sites: an economic approach to the optimal design of cluster samples. *Controlled clinical trials*, *24*(5), 544-559.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276-291.

Cox, D. R. (1971). The choice between alternative ancillary statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 251-255.

Ditchman, N. M., Miller, J. L., & Easton, A. B. (2018). Vocational Rehabilitation Service Patterns: An Application of Social Network Analysis to Examine Employment Outcomes of Transition-Age Individuals With Autism. *Rehabilitation Counseling Bulletin*, *61*(3), 143-153.

Donner, A., Birkett, N., & Buck, C. (1981). Randomization by cluster: sample size requirements and analysis. *American Journal of Epidemiology*, *114*(6), 906-914.

Donner, A., & Koval, J. J. (1980a). The estimation of intraclass correlation in the analysis of family data. *Biometrics*, 19-25.

Donner, A., & Koval, J. J. (1980b). The large sample variance of an intraclass correlation. *Biometrika*, *67*(3), 719-722.

Donner, A., & Koval, J. J. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics*, *46*(3), 271-277.

Dutta, A., Gervey, R., Chan, F., Chih-chin, C., & Ditchman, N. (2008). Vocational rehabilitation services and employment outcomes for people with disabilities: A united states study. *Journal of Occupational Rehabilitation*, 18(4), 326-334.

Efron, B., & Hinkley, D. (1978). Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information. *Biometrika, 65*(3), 457-482. doi:10.2307/2335893

Eignor, D. R. (2013). *The standards for educational and psychological testing*. American Psychological Association.

Ellis, P. D. (2009, September 7). *Thresholds for interpreting effect sizes* [Website log post on Hong Kong Polytechnic University]. Retrieved August 11, 2018, from http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html

ESRA Legislation - U.S. Department of Education. (May 2008). *Public Law Print: Education Sciences Reform Act*. Retrieved from https://ies.ed.gov/director/pdf/ESRAreauth.pdf

Fisher, R. A. (1915). Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika, 10*(4), 507-521. doi:10.2307/2331838
(Editorial). (1915). On the Distribution of the Standard Deviations of Small Samples: Appendix I. To Papers by "Student" and R. A. Fisher. *Biometrika, 10*(4), 522-529. doi:10.2307/2331839

Fisher, R. A. (1925a). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fisher, R. A. (1925b, July). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 22, No. 5, pp. 700-725). Cambridge University Press.

Fisher, R. A. (1942). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1958a). Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science*, *2*, 151-166.

Fisher, R. A. (1958b). Lung cancer and cigarettes. *Nature*, *182*(4628), 108.

Fleming, A. R., Del Valle, R., Kim, M., & Leahy, M. J. (2013). Best practice models of effective vocational rehabilitation service delivery in the public rehabilitation program: A review and synthesis of the empirical literature. *Rehabilitation Counseling Bulletin*, *56*(3), 146-159.

Flom, P. (2015, March 10). *What is adjusted correlation* [Website log post on Quora]. Retrieved August 8, 2018, from https://www.quora.com/What-is-adjusted-correlation

Givens, G. H., & Hoeting, J. A. (2012). *Computational statistics* (Vol. 710). John Wiley & Sons.

Hauck, W. W., Gilliss, C. L., Donner, A., & Gortner, S. (1991). Randomization by cluster. *Nursing research*, *40*(6), 356-358.

Hays, W. L. (1994). *Statistics*. Fort Worth: Harcourt Brace College Publishers.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlations in three-and four-level models. *Educational and Psychological Measurement*, *72*(6), 893-909.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation review*, *37*(6), 445-489.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945-960.

Karlin, S., Cameron, E. C., & Williams, P. T. (1981). Sibling and parent--offspring correlation estimation with variable family size. *Proceedings of the National Academy of Sciences*, *78*(5), 2664-2668.

Klar, N., & Donner, A. (2001). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in medicine*, *20*(24), 3729-3740.

Klar, N., & Donner, A. (2015). The impact of EF Lindquist's text on cluster randomisation. *Journal of the Royal Society of Medicine*, *108*(4), 142-144.

Kosciulek, J. F. (2010). Evidence-Based Rehabilitation Counseling Practice: A Pedagogical Imperative. *Rehabilitation Education*, *24*.

Kosciulek, J. F., & Merz, M. (2001). Structural analysis of the consumer-directed theory of empowerment, *Rehabilitation Counseling Bulletin*, 44(4), 209-216.

Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*. Boston: McGraw-Hill Irwin.

Leahy, M. J., Thielsen, V. A., Millington, M. J., Austin, B., & Fleming, A. (2009). Quality assurance and program evaluation: Terms, models, and applications. *Journal of Rehabilitation Administration*, *33*(2), 69.

Leahy, M. J., & Arokiasamy, C. V. (2010). Prologue: Evidence-based practice research and knowledge translation in rehabilitation counseling. *Rehabilitation Research, Policy, and Education*, *24*(3/4), 173.

Leahy, M. J., Chan, F., & Lui, J. (2014a). Evidence-based best practices in the public vocational rehabilitation program that lead to employment outcomes. *Journal of Vocational Rehabilitation*, *41*(2), 83-86.

Leahy, M. J., Chan, F., Lui, J., Rosenthal, D., Tansey, T., Wehman, P., Kundu, M., Dutta, A., Anderson, C. A., Del Valle, R., & Sherman, S. (2014b). An analysis of evidence-based best practices in the public vocational rehabilitation program: Gaps, future directions, and recommended steps to move forward. *Journal of Vocational Rehabilitation*, *41*(2), 147-163.

Lee, C.-L. (2014). *Linking paths between rehabilitation customer characteristics, services and outcomes by decision tree models* (Unpublished apprenticeship paper. Michigan State University. Department of Counseling, Educational Psychology, Special Education).

Lee, C.-L., Pi, S., & Thielsen, V. (2012). *Relationships of Customer Characteristics, Services and Outcomes Using a Data Mining Approach* (An unpublished internal report to Michigan Rehabilitation Services. Project Excellence, Program of Rehabilitation Counseling, Department of Counseling, Educational Psychology, Special Education, Michigan State University).

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59-66

Lingard, B. (2013). The impact of research on education policy in an era of evidence-based policy. *Critical Studies in Education*, *54*(2), 113-131.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). An Introduction to Statistical Concepts. New York: Routledge.

Kelly, K. (2018). *CEP932: Quantitative Methods in Education Research I [Spring 2018], class notes for bivariate measures of association [Pearson's product moment correlation coefficient]*. College of Education, Michigan State University, East Lansing, Michigan, USA.

Mayhew, S. (2015). *A dictionary of geography*. Oxford University Press.

Maas, C. J., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational statistics & data analysis*, 46(3), 427-440.

Menon, A., Korner-Bitensky, N., Kastner, M., McKibbon, K., & Straus, S. (2009). Strategies for rehabilitation professionals to move evidence-based knowledge into practice: a systematic review. *Journal of Rehabilitation Medicine*, *41*(13), 1024-1032.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics*. New York: McGraw-Hill.

Moore, C. L., Flowers, C. R., & Taylor, D. (2000). Vocation al rehabilitation services: Indicators of successful rehabilitation for persons with mental retardation. *Journal of Applied Rehabilitation Counseling*, 31(2), 36-40.

Moore, C. L. (2001). Disparities in closure success rates for African Americans with mental retardation: An *ex post-facto* research design. *Journal of Applied Rehabilitation Counseling*, 32(2*),* 31-36.

Moore, C. L., Feist-Price, S., & Alston, R. J. (2002a). Competitive employment and mental retardation: Interplay among gender, race, secondary psychiatric disability, and rehabilitation services. *Journal of Rehabilitation*, 68(1), 14-19.

Moore, C. L., Feist-Price, S., & Alston, R. J. (2002b). VR services for persons with severe/profound mental retardation: Does race matter? *Rehabilitation Counseling Bulletin*, 45(3), 162-167.

Moore, C. L., Harley, D. A., & Gamble, D. (2004). Ex-post-facto analysis of competitive employment outcomes for individuals with mental retardation: National perspective. *Mental Retardation*, 42(4), 253-262.

Murray, D.M. & Short, B. (1995). Intra-Class Correlation Among Measures Related to Alcohol Use by Young Adults: Estimates, Correlates, and Applications in Intervention Studies. Journal of Studies on Alcohol, 56(6), 681-694.

Muthén, L. K., & Muthén, B. O. (2012). Mplus user's guide*: Statistical analysis with latent variables* (Version *6*). *Los Angeles, CA: Muthén & Muthén*.

NCLB Legislation - U.S. Department of Education. (January 8, 2002). *Public Law Print: No Child Left Behind Act*. Retrieved from https://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional children*, *71*(2), 137-148.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*(1), 201-211.

O'Neill, J., Kang, H. J., Sánchez, J., Muller, V., Aldrich, H., Pfaller, J., & Chan, F. (2015). Effect of college or university training on earnings of people with disabilities: A case control study. *Journal of Vocational Rehabilitation*, 43(2), 93-102.

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Evaluation review*, *30*(1), 66-85.

Pearson, K., & Lee, A. (1903). On the Laws of Inheritance in Man: I. Inheritance of Physical Characters. *Biometrika, 2*(4), 357-462. doi:10.2307/2331507

Pearson, K. (1904). On the Laws of Inheritance in Man: II. On the Inheritance of the Mental and Moral Characters in Man, and Its Comparison with the Inheritance of the Physical Characters. *Biometrika, 3*(2/3), 131-190. doi:10.2307/2331479

Pearson, K. (1920). Notes on the History of Correlation. *Biometrika, 13*(1), 25-45. doi:10.2307/2331722

Pearson, K. (1923). On the Correction Necessary for the Correlation Ratio η. *Biometrika, 14*(3/4), 412-417. doi:10.2307/2331822

Pi, S. (2006). *Micro-and Macro-level Factors Related to Vocational Rehabilitation Outcomes* (Doctoral dissertation, Michigan State University. Department of Counseling, Educational Psychology, Special Education).

Pi, S., & Thielsen, V. (2011). RSA 911 Data Is a Gold Mine If You Have the Right Shovel, presented at *the 4th Summit on Vocational Rehabilitation Program Evaluation & Quality Assurance*. September 13th & 14th, 2011. Grand Hyatt Tampa Bay, Tampa, Florida, U.S.A. Retrieved from http://vocational-rehab.com/wp-content/uploads/2013/04/C802.0007.01.pdf

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, *11*(4), 621-637.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. New York, NY: Psychology Press, Tylor and Francis Group, LLC.

Raykov, T., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable analysis. *Structural Equation Modeling*, *17*(4), 629-641.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Raykov, T. (2011). Intraclass correlation coefficients in hierarchical designs: Evaluation using latent variable modeling. *Structural Equation Modeling*, *18*(1), 73-90.

Raykov, T., & Marcoulides, G. A. (2015a). Intraclass correlation coefficients in hierarchical design studies with discrete response variables: A note on a direct interval estimation procedure. *Educational and psychological measurement*, *75*(6), 1063-1070.

Raykov, T., & Marcoulides, G. A. (2015b). On examining the underlying normal variable assumption in latent variable models with categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 581-587.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (Vol. 1 Advanced quantitative techniques in the social sciences)*. CA: Sage.

Rehabilitation Services Administration Policy Directive (2013). RSA-PD-14-01. Washington, DC. Retrieved from https://www2.ed.gov/policy/speced/guid/rsa/subregulatory/pd-14-01.pdf

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*(2), 135-147.

Rizzo, M. L. (2007). *Statistical computing with R*. Chapman and Hall/CRC.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of social service Research*, *21*(4), 37-59.

Rosenthal, D. A., Dalton, J. A., & Gervey, R. (2007). Analyzing vocational outcomes of individuals with psychiatric disabilities who received state vocational rehabilitation services: A data mining approach. *International Journal of Social Psychiatry*, *53*(4), 357-368.

Ross, S. M. (2013). *Simulation*. Amsterdam: Academic Press.

Roussas, G. G. (2002). *A course in mathematical statistics*. San Diego: Academic.

Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International journal of epidemiology*, *44*(3), 1051-1067.

Schoen, B. (2010). *An examination of employment outcomes for individuals with spinal cord injury served by the state vocational rehabilitation services program between 2004 and 2008* (Doctoral dissertation, Michigan State University. Department of Counseling, Educational Psychology, Special Education).

Schoen, B. A., & Leahy, M. J. (2012). An Analysis of the Changing Demographics of Individuals with Spinal Cord Injury Who Received State Vocational Rehabilitation Services between 2004 and 2008. *Journal of Rehabilitation*, *78*(3).

Schonbrun, S. L., Sales, A. P., & Kampfe, C. M. (2007). RSA Services and Employment Outcome in Consumers with Traumatic Brain Injury. *Journal of Rehabilitation*, *73*(2).

Schneider, B. (Ed.). (2018). *Handbook of the Sociology of Education in the 21st Century*. Springer.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational design*. American Educational & Research Association.

Schochet, P. (2005). Statistical Power for Random Assignment Evaluations of Education Programs. Princeton, NJ: Mathematica Policy Research, Inc.

Schwanke, T., & Smith, R. O. (2004). Technical report–Vocational rehabilitation database analysis: RSA-911 case service report and database linking (Version 1.0). *Rehabilitation Research Design & Disability: University of Wisconsin-Milwaukee*.

Sink, T., Bua-Iam, P., Hampton, J. E., & Snuffer, D. W. (2014). Applying Location Theory in Vocational Rehabilitation. *Journal of Rehabilitation Administration*, *38*(2), 73-86.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher*, *31*(7), 15-21.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational researcher*, *37*(1), 5-14.

Shavelson, R. J., Phillips, D. C., Towne, L., & Feuer, M. J. (2003). On the science of education design studies. *Educational researcher*, *32*(1), 25-28.

Soper, H., Young, A., Cave, B., Lee, A., & Pearson, K. (1917). On the Distribution of the Correlation Coefficient in Small Samples. Appendix II to the Papers of "Student" and R. A. Fisher. *Biometrika, 11*(4), 328-413. doi:10.2307/2331830

Stapleton, J. H. (2009). *Linear statistical models* (Vol. 719). John Wiley & Sons.

Student. (1917). Tables for Estimating the Probability that the Mean of a Unique Sample of Observations Lies Between -∞ and Any Given Distance of the Mean of the Population from Which the Sample is Drawn. *Biometrika, 11*(4), 414-417. doi:10.2307/2331831

Sullivan, G. M. (2011). Getting off the "gold standard": randomized controlled trials and education research. *Journal of Graduate Medical Education*, *3*(3), 285-289.

Supporting Information for the RSA-911 Data. (n.d.). Retrieved November 18, 2018 from https://rsa.ed.gov/display.cfm?pageid=75

Tachibana, Y., Miyazaki, C., Mikami, M., Ota, E., Mori, R., Hwang, Y., Terasaka, A., Kobayashi, E., & Kamio, Y. (2018). Meta-analyses of individual versus group interventions for pre-school children with autism spectrum disorder (ASD). *PloS one*, *13*(5), e0196272. https://doi.org/10.1371/journal.pone.0196272

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Pearson Addison Wesley.

The What Works Clearinghouse (WWC). (n.d.). *Standards Handbook Version 4.0.* Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education*. Upper Saddle River, New Jersey: Pearson Education, Inc.

U.S. Department of Education. (September 16, 2016). *Guidance and Regulatory Information*. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf

WIOA Legislation - U.S. Department of Labor. (June 1, 2018). *Overview and Highlight: Workforce Innovation and Opportunity Act*. Retrieved from https://www.doleta.gov/WIOA/Overview.cfm