2009

Same Author and Same Data Dependence in Meta-Analysis

In-Soo Shin

FLORIDA STATE UNIVERSITY

COLLEGE OF EDUCATION

SAME AUTHOR AND SAME DATA DEPENDENCE IN META-ANALYSIS

By

IN-SOO SHIN

A Dissertation submitted to the
Department of Educational Psychology and Learning Systems
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Summer Semester, 2009

The members of the committee approved the dissertation of In-Soo Shin defended on May 1, 2009.

Betsy Jane Becker
Professor Directing Dissertation

Fred Huffer
Outside Committee Member

Akihito Kamata
Committee Member

Yanyun Yang
Committee Member

Approved:

Akihito Kamata, Chair, Department of Educational Psychology & Learning Systems

Marcy Driscoll, Dean, College of Education

The Graduate School has verified and approved the above-named committee members.

I dedicate this to my late father (Yong-Ho Shin),
and my late mother in law (Kwang-Ja Yeo)

ACKNOWLEDGEMENTS

When I came to FSU in 2002 for the government officer-training program, my goal was to learn educational methodology and enjoy my life for the first time in my life before going back to Korea two years later. However, with Dr. Kamata's help, I dared to start my doctoral program. He helped me find a graduate assistant job in the summer of 2004 and a full-time job at the Florida Department of Education as a psychometrician in the summer of 2006. His kindness and help were what made me start to my doctoral program, and he always helped me get the tuition waiver that I needed for my doctoral program.

Dr. Becker came to FSU from MSU in the fall of 2004. I asked her permission to attend her SynRG meeting, and she gave me tutoring on meta-analysis every Friday. Her passion for teaching and expertise in meta-analysis has left a deep impression on me. I started my doctoral program with Dr. Kamata's help, and I started my dissertation with Dr. Becker's guidance. As an advisor, faculty member, and researcher, she is a great role model. To me, she always tackles the issues of impossible missions. However, whenever she tackles the 'mission impossible', she makes it a possible one. She is a magician to me. Since the fall of 2006, I have followed her guidance. She is my light-house in the dark when I have lost my way. She is a friend who encourages me, an advisor who gives me outlines to study, and a researcher who finds a methodology when I am stuck in statistical issues. She is a queen and emperess of editing, while I am a beggar of writing and the APA style. I should acknowledge that I could not finish this dissertation without the help and guidance of Dr. Becker.

I attended Dr. Huffer's Stat classes in 2005 and 2006, and he is always nice to his students. When I asked for help with my simulation in the summer of 2008, he was so generous and kind to help me with the R code for my study. Frankly speaking, the same data and same author situation are really difficult issues when generating data. However, his expertise made it possible to simulate this data, and whenever I asked for his help, he always spent almost all afternoon working with me, and he did this from the summer of 2008 to the spring of 2009.

Dr. Yang is always positive and gave a good insight on statistical issues for my dissertation. She is an expert on SEM, and her lectures made me understand the statistical issues more thoroughly. Before Dr. Yang came to FSU in 2007, I took most of my statistical courses from Dr. Tate. I express my thanks to both of them for their wisdom and guidance.

I also want to express my thanks to my FDOE friends (Florida Department of Education). Salih is the leader of our psychometric team, and he understands and helps me in the horrible situation of having to work during the day while studying at night and on weekends. Hiro is a really responsible guy who helped me a lot during difficult psychometric issues. He is a hero to our entire team. Eddy is a nice guy who is always smiling and welcoming while he works. He is always available to help me whenever I ask for his assistance. Harlon is super-man to the team. He really works hard and is a multi-tasking guy. Thank you, my FDOE friends and team! I have worked at the FDOE since the fall of 2006, and I can't imagine my life without their help. I also express my thanks to Eileen, Betty, Jackie (Director), Alana (TDC secretary), Vince (TDC director), Kris and Victoria (Bureau chiefs), and to Cornelia (Administrator). It was a wonderful experience for me, and I learned a lot about psychometrics. I can now finish my dissertation with the help and understanding I attained from all of my fellow FDOE staff members.

Before my work at the FDOE, my parents-in-law helped me financially. My mother-in-law visited us three times (in the winter of 2002, the summer of 2004, and the spring of 2005). She, like Dr. Kamata, encouraged me to begin my doctoral program, and she gave my family the financial help to do it. Without her and my father-in-law's help, I could not have started my doctoral program and would not be where I am today. My mother-in-law died of cancer in October of 2007. She is missed.

My father also died in April of 2007. I ended up visiting Korea three times in 2007. That was the hardest time in my life. At that time, I tried to give up my studies and go back to Korea because my mother's health was also suffering. I must admit that life was meaningless to me. However, I have a son, Hee-yoon, born in January of 2005, and a daughter, Da-won, born in February of 2001. Da-Won and Hee-yoon made me happy whenever I was tired with my studies. My children are my hope and source of energy. Their smiles make me forget my difficulties and hardships.

Three women in my life were instrumental in making me complete this study. The first one is my mother. She always sacrifices her life for us, and she cannot sleep well because she is always working for her family. Her efforts make me determined not to give up on my work and studies regardless of difficulties and obstacles in my life. She did not give verbal lessons to me, but her life is a lesson in and of itself for me. I just want to do something great, especially in the field of education, as she did something great for my family. I want to make her happy when I go back to Korea. The second woman who was instrumental in helping me is my wife. Her nickname was Angel to my college friends, and she is a real angel to me as a wife and friend. I am a foolish guy, but she always expresses her respect to the idiot guy that I am. Whenever I made mistakes, she understood and helped me restart my work and research. I always talk to my friends; only my wife can live with me. The third woman is Dr. Becker. She is the ideal of an advisor, since I cannot imagine a better one in the world. I am a really lucky guy to have met these three women in my life. Thank you, my mother, my wife, and Dr. Becker.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SAME AUTHOR AND SAME DATA DEPENDENCE IN META-ANALYSIS

ABSTRACT

When conducting meta-analysis, reviewers gather extensive sets of primary studies for meta-analysis. When we have two or more primary studies by the same author, or two more studies using the same data set, we have the issues we call ‘same author’ and ‘same data’ issues in meta-analysis. When a researcher conducts a meta-analysis, he or she first confronts ‘same author’ and ‘same data’ issues in the data gathering stage. These issues lead to between studies dependence in meta-analysis.

In this dissertation, methods of showing dependence are investigated, and the impact of ‘same author’ studies and ‘same data’ studies is investigated. The prevalence of these phenomena is outlined, and how meta-analysts have treated this issue until now is summarized. Also journal editors’ criteria are reviewed.

To show dependence of ‘same author’ studies and ‘same data’ studies, fixed-effects categorical analysis, homogeneity tests, and intra-class correlations are used. To measure the impact of ‘same author’ and ‘same data’ studies, sensitivity analysis and HLM analyses are conducted. Two example analyses are conducted using data sets from a class-size meta-analysis and ESL (English as a Second Language) meta-analysis. The former is an example of the ‘same data’ problem, and the latter is an example of the ‘same author’ problem. Finally, simulation studies are conducted to assess how each analysis technique works.

CHAPTER 1

INTRODUCTION

The reason we conduct meta-analysis is to find an overall conclusion about the direction and strength of a relationship between variables, because each primary study's result is limited, and sometimes studies are contradictory. Researchers want to find general findings without bias. Meta-analysis can be defined as a statistical method for synthesizing primary studies' findings. Meta-analysis cannot guarantee the truth, but can represent the present state of research results.

The popularity and impact of meta-analysis is increasing. A large number of meta-analyses have appeared in research journals in psychology and related areas (Hunter & Schmidt, 2000). We can see the impact of meta-analysis in the fact that textbooks summarizing knowledge within fields increasingly cite meta-analyses rather than a selection of primary studies (Hunter & Schmidt, 1996). As the popularity and the impact of meta-analysis increase, the analyses and interpretations of results should be more cautious.

Whenever a researcher conducts a meta-analysis, the meta-analyst gathers primary studies. When a meta-analyst gathers primary studies, they often encounter same-author studies on the same topic, and many studies using common public data sets like the Coleman et al. (1966) data set and STAR (Student Teacher Achievement Ratio) data (Achilles, 1994; Finn, Fulton, Zaharias & Nye,. 1989; Goldstein & Blatchford, 1998; Johnston et al., 1990; Mostellar, 1995; Nye et al., 1992). These studies lead to dependence in effect sizes.

This is different from the dependence discussed by Gleser and Olkin (1994) that arises due to multiple treatment studies and multiple endpoint studies. Those types of dependence are based on multiple effect sizes that use the same control group with different treatment groups, or repeated measures on the same samples within studies. However, my research issue is different because Gleser and Olkin's dependence is within study dependence and dependence due to repeated measures, but my issue concerns dependence between studies. For example, studies of class size and student achievement may have several effect sizes based on students assessed on different subjects. These several effect sizes have a kind of within study dependence. However, if an author writes several papers about the effect of class size on student achievement, or several papers using the STAR data set, we have the ‘same author’ and ‘same data’ issues: between study dependence. The ‘same author’ issue and ‘same data’ issue can have similar influences on effect-size estimates, so a researcher needs to consider dependence when estimating effect sizes from same-author studies and studies using the same data sets.

Meta-analysts have not paid much attention to same-author papers and papers using the same data. However, Rose and Stanley (2005) mentioned the same-author issue and papers using the same data, saying that; “Some estimates are highly dependent, being generated by the same data, methods, or authors” (p. 350). One author could have several primary papers on the same issue. Nowadays, information is increasing dramatically and research areas are narrowing and narrowing. So there are many more possibilities of finding same-author papers in the data gathering stage than ever before in the history of meta-analysis. This phenomenon will accelerate as time goes on.

In the process of meta-analysis, there are five stages (Cooper, 1998): Problem formulation, data collection (searching the literature), data evaluation (coding the literature), analysis and interpretation, and public presentation.

Here, I explain the importance of and reason for this research based on these five stages. Of the five stages, this research is particularly related to the data gathering, data evaluation, data analysis, and reporting stages, as described here:

- Data gathering stage: Meta-analysts will try to find as many as primary studies they can in the data gathering stage, and they often find several same-author studies because authors have a tendency to specialize on particular issues. Also, meta-analysts may find studies using common public data sets like the Coleman et al. (1966) data set and the Tennessee STAR data set.
- Data evaluation stage: Meta-analysts need to list all studies and should decide whether to include or exclude the same-author studies and studies using the same data set.

- Data analysis stage: If a researcher excludes all same-author studies and studies using the same data, it will cause loss of information. If a meta-analyst includes all of these studies, the researcher will encounter dependence. The meta-analyst then must decide how to address the dependence.
- Reporting results: Researchers want to report results without bias, so, same-author studies and studies using same data must be addressed in reports to ensure generalizable findings in meta-analysis.

Research questions

In this dissertation, two research questions are investigated for the 'same author' issue and 'same data' issue in meta-analysis:

First, "How can research synthesists show the dependence due to studies by the same author or many studies using the same data sets in meta-analysis?"

Second, "If there is dependence because of studies by the same author and studies using the same data set, how can meta-analysts measure the impact on the estimate of effect size?"

Purpose of this study

The 'same author' and 'same data' issues are related to the independence assumption in meta-analysis, to the question of which unit of analysis is appropriate, and to the generalizability of research findings. To study these issues, I used several approaches.

First, I examined how prevalent these phenomena are in the meta-analysis field. I reviewed existing meta-analyses to check how often studies by the same authors appear in real meta-analyses. I also assessed how many primary studies in the meta-analyses used common data sets, as in the above mentioned class-size review.

Second, I investigated how meta-analysts have treated these issues until now, and explored journal editors' attitudes about these issues. I also conducted a case study to understand the problems of these issues in depth, using two sample meta-analyses involving ‘same author’ studies and ‘same data’ studies: the former is a procrastination meta-analysis (Steel, 2007) and the latter is a class-size meta-analysis (Shin, 2008).

Third, given the existence of ‘same author’ studies and ‘same data’ studies in published meta-analyses, I proposed several statistical methods to show the dependence among results, including homogeneity tests, fixed-effects categorical analysis, and intra-class correlation. I also proposed sensitivity analysis and a Hierarchical Linear Model (HLM) approach to measure the impact of between studies dependence in estimating effect size.

Fourth, to evaluate the proposed analytical methods, empirical analyses are conducted using two example meta-analyses: a class-size meta-analysis (Shin, 2008) and meta-analysis on the teaching of English as a second language (ESL) (Ingrisone & Ingrisone, 2007).

Fifth, for the ‘same data’ issue, I generated hypothetical ‘same data’ sets, and investigated the effect and influence of ‘same data’ sets on estimating effect size under various meta-analysis scenarios. For the ‘same author’ issue, I proposed a possible analytical model based on the characteristics of ‘same author’ studies in meta-analysis, and also conducted a simulation study.

Until now, many meta-analysts have mentioned these issues only briefly when reporting on data gathering and data analysis. However, no one has furthered a systematic approach or proposed any guidelines for managing these issues thoroughly in meta-analysis.

CHAPTER II

LITERATURE REVIEW

In the literature review, I have studied the following issues for the 'same author' and 'same data' studies. First, how prevalent are these phenomena? Second, what are the present ways of dealing with this issue? Third, what are journal editors' criteria for accepting studies by the same author or using the same data? Journal editors' decisions lead to the appearance of 'same author' studies and 'same data' studies in meta-analysis, because if journal editors will not accept studies on the same topic by the same author and studies using the same data, meta-analysts will not find these studies when doing meta-analyses. Fourth, I presented two case studies to illustrate the 'same author' and 'same data' issues: a meta-analysis with many studies by one author (a synthesis of procrastination studies by Steel, 2007), and a meta-analysis in which many studies use the same public data sets (a synthesis of class-size studies by Shin, 2008).

Prevalence of and way of dealing with 'same author' and 'same data' papers

1. Same author issue

I investigated a collection of existing meta-analyses to check how often multiple studies by the 'same author' appear in real meta-analyses. First, I reviewed articles in two journals (*Psychological Bulletin* and *Review of Educational Research*) from 2004 to March 2008. I found 39 and 13 meta-analysis articles among 212 and 71 total articles in *Psychological Bulletin* and *Review of Educational Research,* respectively; they are in Appendix A. These two journals are the main journals in education and psychology for meta-analysis. In both journals during this time frame, 18% of the articles were meta-analyses. Surprisingly, all 52 meta-analyses had more than two 'same author' papers. This means that the 'same author' issue happened in every meta-analysis article I examined.

In Appendix B, I show the frequency of the same-author issue in meta-analysis in detail using the 2006 and 2007 issues of the same two journals. In these two journals, most meta-analyses distinguished two kinds of references: one kind is cited papers, and the other papers used in the meta-analysis. I checked the frequencies of ‘same author’ papers based on a review of reference lists. Appendix B presents the frequencies of ‘same author’ papers in those meta-analyses. In Appendix B, the number of ‘same author’ papers is between 2 and 26 papers by one author. The largest numbers of papers by one author are 26 papers in Steel (2007) and 16 papers in Bar-Haim et al. (2007), respectively.

Bar-Haim et al. (2007) examined the boundary conditions of threat-related attentional biases in anxiety. This study had 172 primary studies. I counted the number of ‘same author’ papers in the reference list: 8 authors each wrote a pair of papers on this topic (16 papers), 5 authors wrote 3 papers (15 papers), 2 authors wrote 5 papers (10 papers), 2 authors wrote 6 papers (12 papers), 1 author wrote 8 papers, and finally, 1 author wrote 16 papers on this topic. So, the total count of ‘same author’ papers was 77 papers. This means that 45% of all papers (77 papers among 172 papers) were in sets of papers with the same author. The largest number of papers by one author was 16 papers. However, the meta-analysts did not mention the same author effect at all.

Similarly, Steel (2007) examined possible causes and effects of procrastination based on 691 correlations. This study had 216 primary studies: 17 authors wrote 2 papers (34 papers), 4 authors wrote 3 papers (12 papers), 3 authors wrote 4 papers (12 papers), 1 author wrote 7 papers, 1 author wrote 9 papers, 1 author wrote 12 papers and finally, 1 author wrote 26 papers. Appendix B shows that 52% of the papers (112 among 216 primary studies) were among sets by the same author. One author, J.R. Ferrari, had 26 first authored papers. I investigated Ferrari’s 26 papers to understand the characteristics of papers by the ‘same author’ in this meta-analysis.

When I investigated the meta-analysis articles in these two major journals (*Psychological Bulletin, Review of Educational Research*), I did not find much mention of the ‘same author’ issue in real meta-analyses. Only two articles briefly referenced this issue. However, Bateman and Jones (2003) described the ‘same author’ situation in detail.

First, Malle (2006) mentioned non-independence issues, stating “Effect sizes from samples collected in the same setting and by the same researchers will on average be correlated and may therefore inflate the effect size averages.” (p. 913). Here, ‘same researchers' mean 'same authors' and the effect sizes from the same researchers and similar settings are nested. To estimate the possible inflation effect in the same settings and same-researcher studies, Malle (2006) computed a per-article effect size average, but the effect-size value was virtually identical to the average based on the 173 individual studies/samples.

As Malle (2006) said, effect sizes from samples collected in the same setting and by the same researchers are nested and correlated, so meta-analysts need to check if there is dependence or not. Malle checked if there was possible inflation or not, by computing a per-article effect size average. However, I do not agree with Malle's computation method. Malle computed one average effect size per article to check for possible inflation of the overall effect-size estimate, but Malle did not distinguish 'same author' articles from 'different authors' articles. In this research, I have distinguished 'same author' papers and 'same data' papers from the 'different author' papers and 'different data' papers to investigate the possible dependence.

Second, consider the situation with two versions of the same paper: one is published, the other is unpublished. Nesbit and Adesope (2006) chose to include the published journal paper rather than an unpublished one, stating "When a study was reported in more than one source (e.g., dissertation and journal article), the version published in a journal article was used for coding" (p. 442). However, Weisz, McCarty and Valeri (2006) chose the unpublished paper if one was published and the other unpublished from the same data set. This is a little bit different from the 'same author' issue because this is just two versions of the same paper. 'Same author' papers do not mean the same paper, but represent different papers on the same topic by the same author.

Third, Bateman and Jones (2003) described the exact situation of the present 'same author' issue. Bateman and Jones (2003) described various meta-analysis models of woodland recreation benefit estimates, comparing a traditional meta-analysis model with multi-level model (MLM) techniques. They said, "Our conventional models suggest that studies carried out by certain authors are associated with unusually large residuals within our meta-analysis. However, the MLM approach explicitly incorporates the hierarchical nature of meta-analysis data, with estimates nested within study sites and authors." (p. 235). Bateman and Jones (2003) explained the benefit of the MLM approach in meta-analysis: “The MLM approach allows the researcher to explicitly incorporate potential nested structures within the data, and allows the researcher to relax strong and commonly adopted assumptions regarding the independence of estimates with respect to the numerous natural hierarchies within which they reside.” (p. 237). Their MLM is a useful approach for examining the clustering of estimates within authors. They also explained the problem of the traditional approach in meta-analysis: “A potential limitation of the application of conventional regression techniques in meta-analysis occurs if the observations being modeled possess an inherent hierarchy” (p. 247). Aside from the nested data structure issue, the traditional approach will poorly estimate parameters and standard errors: “Problems with standard error estimation arise due to the presence of intra-unit correlation” (Bateman & Jones, 2003, p. 248). If the intra-unit correlation is large, it will lead to underestimated standard errors and more easily rejected null hypotheses in the traditional approach (Bateman & Jones, 2003). Bateman and Jones described the same author issue very well, and their multi-level approach is one reasonable way to deal with the ‘same author’ issue in meta-analysis, compared to conventional approaches. Compared to the prevalence of the ‘same author’ problem in meta-analysis, meta-analysts have not much paid attention to the issue, but Bateman and Jones (2003) explained well the same author issue and laid out an appropriate multi-level approach.

2. Same data issue

I next reviewed how reviewers have treated the ‘same data’ issue in meta-analysis articles by investigating meta-analyses in two journals: *Review of Educational Research* and *Psychological Bulletin*. Compared to the ‘same author’ issue, the same data issue has many more references. However, these references cover concerns that are a little different from the present research issue. Meta-analysis articles mainly refer to ‘same sample’ issues, not ‘same data’ issues. The objective of the present research is to deal with the similarities and study level dependence that arise from repeatedly using same public data sets, but meta-analysis articles mainly are concerned with the dependence of effect–size estimates when the exact same sample appears in several studies. Most meta-analysts are concerned with repeated measures, and replication of samples, but my research concerns similarities or dependence based on nested study structures in meta-analysis. There are several examples of nested study structure dependences in meta-analysis: multi-site studies, same-laboratory studies, studies using the same public data set, and ‘same author’ studies.

Kalaian (2003) defined a multisite study as research that studies the effectiveness of similar, or variations of the same interventions across multiple similar or distinct sites. These sites can include multiple clusters of individuals, classrooms, schools and other similar settings. Gurevitch and Hedges (1999) presented a same-laboratory study example as follows, “when several effect-size estimates are computed from the same laboratory, there may be dependence if common materials, procedures, etc., in a laboratory make the outcomes of separate experiments obtained from the same laboratory less variable than those obtained from different laboratories” (p. 1147). Multi-site studies and same-laboratory studies are similar to same author studies because the data structure is nested in meta-analysis and a multi-level approach is suitable for meta-analysis.

In education and other research areas, we have many large public access data sets. Thus, many researchers use these data sets for their own studies. However, I categorized the same sample issues mentioned in most meta-analyses as being indirectly related to the ‘same data’ topic, because meta-analysts did not mention the ‘same data’ issue directly, and both issues involve the issue of dependence in estimating effect-size. From meta-analyst’ treatment of the same sample issue, I have drawn some ideas about how we should treat the ‘same data’ issue. These are described in the following subsections.

1) Choosing one study among nested studies, and excluding other papers using the same sample or same data.

When meta-analysts gather articles, they can encounter papers with samples that overlap each other, papers using repeated measures, and longitudinal study papers having repeated measures taken at different times. Tolin and Foa (2006) included the larger study when the samples of two studies overlapped. Webb and Sheeran (2006) excluded two studies because the studies both used a subset of data from a larger study also included in the review. Kuncel et al. (2005) also chose the largest and most complete study and excluded smaller overlapping studies in all cases with overlapping studies.

However, if analysts choose one study, it will cause some loss of information in meta-analysis. Frattaroli (2006) excluded 6 papers from a meta-analysis because the papers did not present new data, and used data reported in earlier sources. Else-Quest, Hyde, Goldsmith, and Van Hulle (2006) dropped 16 studies and 567 effect sizes using duplicate samples. Glasman and Albarracin (2006) also simply eliminated the statistically dependent within-subject measures in longitudinal papers. Exclusion of papers using the same sample can be an option in meta-analysis. However, the problem with excluding papers so as not to violate the independence assumption is loss of information. Furthermore, the same sample issue is directly related to violation of the independence assumption, but the same data issue is indirectly related to the violation of the independence assumption. If meta-analysts include papers using the same public data sets, they can use the same data sets between studies, but it is not as clear as the repeated use of the exact same sample within a study. With large public data sets, many researchers use the same data set, but they may use different parts of the data set based on their research question: For example, in using the STAR data set, the grade levels of students were not the same: Mostellar (1995) analyzed only first-grade data, and Achilles (1994) reported K-3 results, but the sample size and the results for the first grade sample were not the same as those of Mostellar's (1995) first-grade data.

It is not always clear that they have used exactly the same samples, but studies that have used the 'same data' set often have similarities which are different from the exact dependence existing for papers using the exact same samples.

2) Average or weighted average for multiple outcomes.

When meta-analysts have multiple effect sizes in a single study, making an average is also an option in synthesizing effect sizes in meta-analysis. Multiple outcomes are different from the 'same data' issue in that multiple outcomes are a within study issue, but the 'same data' issue is a between studies issue. Nestbit and Adesope (2006) coded the weighted average effect across multiple treatment groups when data from multiple experimental or comparison treatments were reported.

3) Shifting unit of analysis in a meta-analysis.

Meta-analysis authors are concerned with the independence assumption, unit of analysis and generalizability issues. Effect sizes are assumed to be independent, and considerable dependence among data points can threaten the validity of meta-analytic findings (Malle, 2006). To achieve statistical independence among effects, Sirin (2005) suggested three alternative approaches for choosing the unit of analysis in meta-analysis: each study as the unit of analysis, each effect size as the unit of analysis, or "shifting unit of analysis". Using study as the unit of analysis will lead to loss of information when a study has several effect sizes in it, while using the effect size as a unit of analysis will leave dependence within a study. The "Shifting unit of analysis" is a kind of compromise between study and effect size in the choice of unit of analysis.

When researchers estimate a total effect size, they effectively use the study as a unit, however, when researchers estimate effect sizes for sub-groups, they can use the effect size as a unit. When meta-analysts study the effect of class-size on student achievement, there are reading, math, and science achievement outcomes for the same samples. When the analyst measures effect size for each subject area, the unit of analysis is the effect size for each subject. However, when the meta-analyst measures the overall effect size, the researcher will use the study as the unit of analysis for the independence assumption thus averaging across the subject areas. Shifting units of analysis is a good strategy because no information is lost and the independence assumption is not violated. However, shifting the unit of analysis is a reasonable analysis method when multiple measures or multiple subgroups exist within a study. As another application of shifting units of analysis, reviewers may choose the author or the data set as the unit of analysis in meta-analysis.

4) Sensitivity analysis.

Other authors have proposed alternative methods related to same sample and data issues. Glasman and Albarracin (2006) suggested sensitivity analysis to analyze dependent data in meta analysis: "19 studies were based on longitudinal measures completed by the same group of participants. However, because the inclusion of the longitudinal reports violates statistical independence assumptions, we present report results that both include and exclude the dependent conditions." (p. 783).

Sensitivity analysis can be a good option for the ‘same data’ issue, for example, because reviewers can report both including the STAR studies and excluding the STAR studies when there are several papers from the STAR data sets.

Goldberg, Prause, Lucas-Thompson, and Himsel (2008) examined the relationship between children’s achievement and mother’s employment. Many studies used the National Longitudinal Survey of Youth (NLSY) to study this issue because the NLSY is widely accessible. The authors included several NLSY papers, “Despite the common source, analyses with the NLSY data have led to varying conclusions, reflecting differences in the subsets of the NLSY sample selected for analysis, measures of maternal employment, child outcome, and the choice of control variable” (Goldberg et al., 2008, p. 80). Goldberg et al. (2008) is an exact example of the same data issue. As the authors said, ‘same data’ studies can give various conclusions because of different analytical choices including various outcomes, constructs, and study characteristics. Goldberg et al. (2008) analyzed 19 studies from the NLSY among a total of 68 studies. The authors considered the dependence of NLSY studies, “The potential nonindependence of the NLSY samples and their overrepresentation in the findings for formal tests of achievement and intellectual functioning prompted us to devise procedures for handling these studies. To represent the range offered by the NLSY studies, the studies with the most negative and the most positive effects were designated as NLSY-low and NLSY-high, respectively” (p. 89). The authors ran their analysis once with the NLSY-low study and once with the NLSY-high study for the purpose of estimating the effect sizes for testing moderators. In this way, the authors analyzed NLSY studies that represent each “end” of the NLSY contribution to the effect size. These authors have another reasonable approach for the NLSY studies, “An additional strategy to give just due to these nationally representative studies was to use the full set of relevant NLSY studies as a moderator in analyses that directly contrasted the effect sizes from NLSY-based studies to those of non-NLSY-based studies” (Goldberg et al., 2008, p. 89). This strategy is good in two aspects: first, the authors used the full set of relevant NLSY studies as a moderator in analysis, second, they contrast the effect size from NLSY-based studies to that from non-NLSY-based studies.

Sohn (2000) employed a Monte-Carlo simulation to evaluate a situation in which several marketing studies had produced sets of more than one effect in military recruiting studies. She considered a situation similar to the ‘same data’ issue, “an estimated effect in one model could be correlated to the corresponding effect in the other model due to similar model specification or the data set partly shared, but their correlation is not known.” (p. 500). Her study is similar in that the data set can be partly shared, but the correlation is not known because the meta-analyst does not know how each data point is shared. The main purpose of Sohn’s study was to evaluate the impact of disregarding the potential correlation, and to ask whether such negligence led to biased estimates, potentially misleading the reader. Her study approach is a good reference for the ‘same data’ issue in that she has distinguished two variances: sampling error and random error variance. She analyzed the impact of the relative size of the variance component because the multi-level approach needs to distinguish level 1 (effect-size level) sampling error from level 2 (study level) random error when analyzing the ‘same data’ studies in meta analysis.

Journal editor’s criteria for accepting the ‘same author’ and ‘same data’ studies

Compared to the prevalence of ‘same data’ studies in meta-analysis, meta-analysts have not paid attention to this issue. In this section, journal editors’ criteria are investigated as the origin of ‘same author’ studies and ‘same data’ studies in meta-analysis.

Journal editors’ criteria affect the prevalence of ‘same author’ papers and ‘same data’ papers in meta-analysis. If journal editors did not accept papers by the ‘same author’ on a particular topic, there would be no ‘same author’ papers in meta-analysis. If journal editors decide not to accept papers using publicly accessible data sets, there will not be multiple papers using the ‘same data’ in meta-analysis.

Below, I have examined journal editors’ standards for accepting papers to understand how the ‘same author’ and ‘same data’ articles are published. This section describes the origin of this phenomenon, and how it relates to guidelines regarding publication of ‘same author’ studies and ‘same data’ studies. After first reviewing the general guidelines of acceptance criteria from several research domain, I describe the specific reasons or exceptions to rules related to ‘same author’ and ‘same data’.

Marketing journal editors have described three minimum criteria for publishing articles, as follows[1]: “First, the paper should make a contribution to the science and practice of marketing. Second, the paper should be based on sound evidence--literature review, theory and/or empirical research. Third, the paper should be valuable to marketing academicians and/or practitioners”. These criteria (or similar ones) will be applicable to other journals and research domains. General guidelines for accepting journal articles are described above, and below I have examined the specific reasons, or exceptional rule of accepting ‘same author’ papers and ‘same data’ papers in journal publication. Common sense would suggest that not every paper should be a replication by the same author on the particular topic, and a researcher should not use the ‘same data’ to replicate the same study repeatedly. However, it is clear that specific reasons and exceptions exist, because ‘same author’ papers are prevalent and because so many papers using the ‘same data’ set appear in various areas. The specific reasons are next described.

Editors have mentioned the “best benefit” and/or benefit of audience. The ‘International Committee of Medical Journal Editors’ (ICMJE)[2] has discussed duplicate submissions, noting that “Editors of different journals may decide to simultaneously or jointly publish an article if they believe that doing so would be in the best interest of the public’s health.” The best interest of the public could be one reason for ‘same author’ papers, ‘same data’ papers and duplicate publications. For this reason, meta-analysts need to decide if two papers are exactly the same or not, and whether to include all of them in the meta-analysis.

Second, there are editorial guidelines on redundant publications for ‘same author’ papers. ICMJE also gives some guidelines for such redundant publications. After an author publishes a preliminary report, the author can publish a complete report. If an author presents a paper at an academic conference, the author can publish the paper in a journal. Meta-analysts usually search for unpublished papers (conference papers/dissertations) to include along with published papers, to reduce publication bias.

[1] “Journal of Marketing” website: http://www.marketingjournals.org/jm/ama_edpolicy.php

[2] International Committee of Medical Journal Editors: http://www.icmje.org/

This is one way an author can have two reports on the same topic: an unpublished one (conference paper/dissertation) and a published one. In this case, meta-analysts need to choose one source for which to estimate effect sizes if the two reports are exactly the same.

Third, editors believe in acceptable secondary publications for papers using the 'same data'. There is publication tips[3] in the American Educational Research Association (AERA) website. The situation of using the 'same data' is discussed in the tips; A question addressed there is, "Can I use the same data for another journal article if the emphasis of that article is different from the original publication?" The answer is "You may refer to the same data, but you should not describe it to the depth of the original piece. Typically, but not always, this occurs when a second article also addresses a different audience. You will be using the data in a different way, depending on the audience and therefore it is ethical and permissible".

Even though a researcher has used the 'same data' set in two reports, both may be published if the audience and uses of data are different. Clearly if an author can use the 'same data', different authors can also. ICMJE also refers to acceptable secondary publications, giving examples like publishing in a different country, publishing for a different group of readers, and publishing with the previous notice of secondary publication to the readers in the article.

Fourth, editors consider competing manuscripts based on the same database. This is directly related to the present research interest in the 'same data' set issue. On its website ICMJE states, "Editors sometimes receive manuscripts from separate research groups that have analyzed the same data set, e.g., from a public database. The manuscripts may differ in their analytic methods, conclusions, or both. Each manuscript should be considered separately. Where interpretations of the same data are very similar, it is reasonable but not necessary for editors to give preference to the manuscript that was received earlier. However, editorial consideration of multiple submissions may be justified in this circumstance, and there may even be a good reason for publishing more

[3] Publishing Educational Research Guidelines and Tips

https://www.aera.net/uploadedFiles/Journals_and_Publications/Journals/pubtip.pdf than one manuscript because different analytical approaches may be complementary and equally valid.”

Based on the paragraph above, we can have many papers using the ‘same data’ if they are different in analytical methods, conclusions or both. Researchers can publish articles using the ‘same data’ set if the researcher uses advanced methods, or takes a different perspective in the conclusions.

If ‘same author’ papers and papers using the same data set exist, meta-analysts will have their own inclusion and exclusion criteria for ‘same author’ papers and ‘same data’ papers for generalizability of findings. The journal editors’ acceptance criteria will have implications for meta-analysts in the data gathering and data analysis stage. In the following section, the characteristics of ‘same author’ studies and ‘same data’ studies will be investigated to better understand this issue.

Two case studies: ‘Same author’ studies and ‘same data’ set studies

In this section I describe two case studies of meta-analyses with many ‘same author’ studies and ‘same data’ studies. I have investigated these to understand the characteristics of these issues in detail. One example is class-size and student achievement, as a case of a meta-analysis faced with the ‘same data’ issue. The other is a review of the literature on the nature of procrastination, as an example of the ‘same author’ issue.

Study 1: ‘Same data’ case study (Tennessee STAR class-size project)

The relationship of class-size and student achievement has a long history and is still an interesting topic for educational policy makers. The assumption of this research is that smaller class-sizes will increase student achievement. When I searched for primary studies, I found 123 primary studies for this issue shown in Appendix C. There are many same author studies, and also same data studies involving data from the Tennessee STAR project. For example, Achilles was the first author of 14 studies and Nye was the first author of 5 class size studies. Among the 123 papers in Appendix C, 26 papers were related to the STAR project. After I set some inclusion criteria and coded all studies, 16 studies were left for analysis as attached in Appendix D. Many studies had no data at all, and some studies had data but were not sufficient for analysis. For example, in some studies the author had reported effect size without reporting sample size. So, most of the 123 studies could not be included in the analysis, and only 16 studies were included and used for analysis.

Among the 16 studies, 8 studies had used STAR data. This description of STAR papers will give us some implications. First, differences between the 'same data' issue and 'same sample' issues (replication of samples) within studies are clarified. Second, a full description should give us a better understanding of 'same data' studies, including several reasons why meta-analysts might wish to analyze such papers and not exclude them from meta-analysis. Furthermore, this case study will give us the characteristics of similarities or possible dependence when using the 'same data' studies in meta-analysis. Finally, this description of STAR studies as a case study of 'same data' papers will have the implications for how to analyze 'same data' studies when conducting meta-analysis.

There are three phases of the Tennessee project: the STAR, the 'Lasting Benefit Study', and the 'Project Challenge'. The STAR studies began 1985 and finished in 1989. Seventy nine schools participated and the number of students in small classes was 13-17, and the number of students in large classes was 22-25. There were 108 small classes and 101 regular classes. Later, the 'Lasting Benefit Study' began as a follow up study in 1989. The experimental group students who had been in small classes during STAR returned to regular sized classes in grade 4, 5 and 6 and beyond. Researchers investigated the benefit of the earlier small class experience in grades K to 3 as a follow up study. Third, 'Project Challenge' also began in 1989, and investigated 17 economically poor school districts from among 139 school districts. The sample for 'Project Challenge' is K-3 students in small classes. Based on these three phases of the Tennessee project, there are a lot of class-size papers. Each study used data from the Tennessee class-size project, but their samples were not the same.

First, the experimental periods and samples drawn during each period were not exactly the same. For example, Mostellar (1995) reported first year results, Goldstein and Blatchford (1998) reported results from the first four years (1985-1989), and Finn, Fulton, Zaharias, and Nye (1989) reported follow-up study results. Actually, the STAR project studied K-3 students for four years: 1985-1989. However, Finn and Achilles (1999) reported that several operational complexities affected the composition of their sample because some students participated from first grade without attending the Kindergarten experiment. Some students moved away from their original project schools, so the participants in different experimental periods were not the same.

Second, the grade levels of students included in these studies were not the same. Mostellar (1995) analyzed only first-grade student data, while the full STAR project examines grades K-3 from 1985 to 1989. Achilles (1994) reported K-3 results, but the number of students and the results in his first grade sample were not same as those of Mostellar (1995).

Third, the Tennessee project had three phases as described above and several original principal investigators worked on the Tennessee Project including C.M. Achilles, H.P. Bain, F. Bellot, J. Folger, J. Johnson, and E. Word. They continued to reanalyze the STAR database to answer new questions, through all three phases of the STAR project. This is one reason for the existence of many published papers on the Tennessee project. Also the data set of STAR is now open to public access; this also makes it possible for a variety of researchers to produce many STAR papers.

In conclusion, I cannot say that these researchers used the same sample and that their focus has the same. So, it would be very difficult to choose only one paper for inclusion in a meta-analysis in this case, and much information would be lost if the meta-analyst does not use all of the papers. Based on this case study, it is clear that in some cases reviewers should not simply exclude papers using the 'same data'. However, meta-analysts should consider and treat the papers using the 'same data' set as dependent because the samples are often much more similar than samples from different data sets in terms of experimental conditions and experiment procedures. Papers using the 'same data' set appeared similar and related, but it was not clear if the exact same sample was repeatedly used. Reviewers need to pay attention to such similarities when doing data evaluation and data analysis. Similarities among papers using the 'same data' set may be similar to the nested structures in the hierarchical linear modeling, such as students being nested in the same classroom.

Study 2: 'Same author' studies (Steel 2007 meta-analysis)

The main question of this case study is follows: “What is the relatedness of same author papers in meta-analysis?”

I investigated Ferrari’s 33 papers among the 216 primary studies in Steel’s (2007) meta-analysis because Ferrari was the author who had the most papers as a first author in this meta-analysis. I have attached Appendix F; a review of these 33 ‘same author’ papers from this meta-analysis. The 33 papers include Ferrari’s dissertation and journal articles based on it. To investigate the characteristics of the ‘same author’ studies, I retrieved 25 of these 33 papers. Below I describe the characteristics of these 25 papers. Most of these studies used similar participants and had similar characteristics.

First, students who were in introductory psychology classes participated in the experiments in 20 of the 25 papers.

Second, the age of students was 18-22 in 22 papers.

Third, the incentive for participation was extra class credit in 15 papers. One paper gave money as an incentive, and for two papers participants were unpaid volunteers.

Fourth, in 10 papers, the location of the experiments was a university in the Midwest of the U.S.

Fifth, the instrument was also often the same. Twelve papers used the AIP (Adult Inventory of Procrastination), and fourteen papers used the DP (Decisional Procrastination) questionnaire as the instrument of measurement.

However, Steel (2007) did not mention anything about the same-author issue, nor were any moderator variables investigated, like age of sample, incentive, and measurement instrument used when synthesizing these studies’ results.

CHAPTER III

METHODOLOGY

In this chapter, methods of showing dependence are proposed, and analysis methods for showing the impact of 'same author' studies and 'same data' studies are suggested. Homogeneity testing, fixed effect categorical analysis, and an intraclass correlation approach are described to show dependence of 'same author' studies and 'same data' studies. After showing dependence, sensitivity analysis and a hierarchical linear modeling (HLM) approach are described as ways of measuring the impact of 'same author' studies and 'same data' studies in meta-analysis.

Homogeneity test

To synthesize research findings, reviewers check whether the finding shares a common population effect size or not. The homogeneity test has an important role in synthesizing research findings when meta-analysts determine the overall effect size. The homogeneity test is investigated here as a tool to show the relatedness of 'same author' studies and 'same data' studies. First, I present an overview of homogeneity testing in meta-analysis. Later, I describe a possible approach to showing dependence using the homogeneity test including quantification of heterogeneity.

1. Homogeneity test in meta-analysis

To check homogeneity, in general, meta-analysts use the *Q* statistic (Cooper and & Hedges, 1994):

$$Q = \sum_{i=1}^{k}[(T_i - \overline{T}_\bullet)^2 / v_i],$$

where T_i is the *i*th study's observed effect size,

$\overline{T}_\bullet$ is the weighted average effect size, and

v_i is the conditional variance of the *i*th standardized mean effect size.

The weighted average effect size for the *Q* calculation is

$$\overline{T}_\bullet = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i} .$$

The weights that minimize the variance of $\overline{T}_\bullet$ (the weighted average effect size) are inversely proportional to the conditional variance (the square root of the standard error) of each effect, specifically $w_i = \frac{1}{v_i}$.

Generally speaking, if *Q* is greater than the critical value of a chi-square at k-1 degrees of freedom, the observed variance in effect sizes is said to be significantly larger than what we would anticipate by chance if effect sizes in all studies shared a common population effect size (Hedges, 1994, p. 266). If Q is significant, it means the estimated effect sizes are heterogeneous, and the effect sizes do not share a common population effect size. However, Q depends on within-study sample size: If the sample sizes within study are very large, Q will be statistically significant even when the effect sizes are quite homogeneous; in such cases, meta-analysts may wish to pool effect-size estimates anyway because the Q is significant based on large within study sample size, not on heterogeneity (Shadish & Haddock, 1994, p. 266). This is a key characteristic of the STAR studies in the meta-analysis of class-size and student achievement. The STAR studies are based on a statewide experiment, so the sample size of this experiment is larger than those of other class-size studies. The overall Q for class-size studies may be significant because the within-study sample sizes in each report on the STAR study are very large.

2. Homogeneity test as an index of relatedness of 'same author' studies and 'same data' studies

The main question in research synthesis is whether there is methodological, contextual, or substantive variation in primary studies related to variation in estimated effect size parameters (Hedges, 1994, p. 286). Meta-analysts can sort their effect sizes into independent groups according to whether they are based on the 'same author' and 'same data', or not. The variance of effect sizes can be partitioned into two parts: between studies variance and within study variance. In the traditional analysis of variance, to test heterogeneity, ratios of sums of squared deviations from group means are used to test for systematic sources of variance. Because different sources of variation can partition the sums of squares, between groups and within-groups' sources can be separately computed from the total variation about the grand mean (Hedges, 1994, p. 289). "The total heterogeneity statistic $Q = Q_T$ (Weighted total sum of squares about the grand mean) is partitioned into a between-groups-of-studies part Q_{BET} (the weighted sum of squares of group means about the grand mean) and a within-groups-of-studies part Q_W (the total of the weighted sum of squares of the individual effect estimates about the respective group means)" (Hedges, 1994, p. 289). Meta-analysts should distinguish three Q statistics in homogeneity testing: overall Q, Q_{BET} and Q_W .

The overall Q tests if the full set of studies share a common effect size, and Q_{BET} tests if there is variation between groups, "Q_{BET} is just the weighted sum of squares of group mean effect sizes about the grand mean effect size to test the hypothesis that there is no variation across group mean effect size" (Hedges, 1994, p. 289). If meta-analysts make groups of 'same author' and 'same data' studies, the Q_{BET} will indicate whether there are differences between 'same author' studies and different author studies, or between 'same data' studies and different data studies. If Q_{BET} is significant, it is an indication that variation is due to the 'same author' variable, or the 'same data' variable.

The Q_W tests whether results are homogeneous within sets of studies, "The hypothesis of Q_W is that there is no variation among population effect sizes within groups of studies. Although Q_W provides an overall test of within-group variability in effects, it is actually the sum of p separate (and independent) within-group heterogeneity statistics, one for each of the p groups of effects. These individual within-group statistics are often useful in determining which groups are the major sources of within-group heterogeneity and which groups have relatively homogeneous effects." (Hedges, 1994, p. 290). In my framework, the group of 'same author' or 'same data' studies should be tested for homogeneity. Thus, the 'same author' and 'same data' factors are tested to see if they explain variation, and the meta-analyst then determines whether most of the total heterogeneity is between groups and relatively little remains within groups. If 'same author' and 'same data' factors can not explain variation or the heterogeneity between studies, it indirectly means that 'same author' and 'same data' studies have little dependence between studies. However, as an exceptional case, reviewers may have small Q_{BET} values and also a small Q_W , which may indicate the relatedness of same author (or same data) studies: If average effects are roughly equal but same-author or same-data studies are more homogeneous, Q_{BET} would be small but Q_W might also be small for same author/data studies, but larger for the other studies. This is the case for the ESL meta-analysis used to illustrate the 'same author' issue in chapter V.

Our interest in variance is based on the assumption that 'same author' studies and 'same data' studies will be more homogenous than different-author and different-data studies. Rose and Stanley (2005) mentioned the 'same author' and 'same data' issue, saying that; "Some estimates are highly dependent, being generated by the same data, methods, or authors" (p. 350). I investigated the characteristics of 'same author' papers in Appendix F. Most of the 'same author' studies used similar participants and used similar incentives and instruments. This suggested that the 'same author' studies would be more homogenous than different author (or different data) studies.

Researchers do not pay as much attention to the variance in other statistical analyses as is paid in meta-analysis, because the variance is considered as a nuisance parameter in standard statistical analysis. So it is included in the model, but is not interpreted (Hox, 2002). However, in meta-analysis, it is important to decide whether the estimated effect sizes are homogeneous or not. Outcomes from studies by the 'same author' can be similar because that author may employ similar sampling methods, use similar experimental manipulations, or measure the outcome with similar instruments.

If we do not know how consistent the estimates of effect size are, we cannot decide how generalizable the results of the meta-analysis may be (Higgins, Thompson, Deeks, & Altman 2003). The homogeneity test measures the consistency of findings, but it can also be used as an alternative way or indirect way to show the relatedness of 'same author' studies and 'same data' studies. This use of the homogeneity test would be a little bit different from the conventional use of the homogeneity test.

3. Quantification of heterogeneity

For quantification of heterogeneity, Higgins and Thompson (2002) proposed a simple, universal statistic which represents heterogeneity in meta-analysis. Their index makes it possible to account for how much heterogeneity is based on study-level covariates, or particularly influential studies. So, this quantification could be used to determine the impact of 'same author' studies and 'same data' studies in meta-analysis. Higgins and Thompson (2002) proposed H and I^2 to quantify heterogeneity in meta-analysis. They can be defined as follows, "H may be interpreted approximately as the ratio of confidence interval widths for single summary estimates from random effects and fixed effect meta-analyses. I^2 describes the percentage of variability in point estimates that is due to heterogeneity rather than sampling error" (p. 1553).

The authors also defined $I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}$, where τ^2 is the between-study variance, and σ^2 is sampling error.

It is very similar in concept to the intraclass correlation. Higgins et al. (2003) also proposed a simple method to calculate I^2: "I^2 can be readily calculated from basic results obtained from a typical meta-analysis as I^2 = 100 %*($Q - df$)/Q, where Q is Cochran's heterogeneity statistic and *df* the degrees of freedom" (p. 558). Higgins et al. (2003) proposed to interpret the index of I^2 as follows "A naïve categorization of values for I^2 would not be appropriate for all circumstances, although we would tentatively assign adjectives of low, moderate, and high to I^2 values of 25%, 50%, and 75%." (p. 559).

In this research, *H*, and I^2 are used as indirect indexes of dependency among 'same author' studies and 'same data' studies in addition to the *Q* statistics. Below I have examined the results of *H,* I^2, and *Q* statistics as indexes of consistency or dependency estimates for 'same author' studies and 'same data' studies. If the study results are more similar because of dependence, meta-analysts would expect to find higher *H* and lower I^2 values for the same author/data studies.

Fixed-effects categorical analysis

To show the similarities and dependence of same author/data studies, we first examine homogeneity across all primary studies. If the homogeneity test is not significant, meta-analysts may not investigate the characteristics of primary studies to find sources of heterogeneity. However, if the results are heterogeneous, analysts will pay attention to the characteristics of studies based on particular grouping variables to explain the heterogeneity. Rosenthal and DiMatteo (2001, p. 67) explained, "Nonindependence may be a problem if the same research lab contributes a number of studies and this fact is ignored. It is possible and often valuable to block by laboratory or researcher and examine this as a moderator variable". Laboratory and researcher can be a grouping variable or moderator for nonindependence or heterogeneity in meta-analysis.

In this categorical fixed-effects model, the researcher will use a 'same author' or 'same data' variable as a between studies characteristic and will make groups of primary studies based on the categories of same author and same data variables. The grouping is based on putting each author's studies and each data set's studies in a group, because all 'same author' group studies will not be homogenous with each other. If a review has many authors with just one paper each, one could categorize these studies into two groups: a particular author's studies vs. all other authors' studies. This research focused on the particular author or data studies' relatedness, compared to all other studies.

For each group of effect sizes, the researcher will make a plot of effect sizes, a confidence interval and box plot to compare the characteristics of same author and same data studies with those of different author/data studies.

Using this graphical approach, meta-analysts can indirectly assess and show the similarities or relatedness of same author and same data studies. In the fixed-effects categorical analysis, meta-analysts expect 'same author' and 'same data' studies to be less variable than different author and different data studies.

The fixed effect categorical analysis approach is an easy way to get a quick grasp of the big picture that shows the relatedness of 'same author' studies and 'same data' studies.

Intraclass correlation

The intraclass correlation (ICC) is described as a third option to show dependence of 'same author' studies and 'same data' studies. For example, "an ICC of .10 indicates that 10% of the variance in individual level responses can be explained by the group-level properties of the data" (Bliese & Halverson, 1998a, p. 159). "The strength of the ICC is that it allows determination of how much of the total variability is due to group membership. Also ICC values are not affected by group size" (Castro, 2002, p. 73). The ICC thus provides an estimate of the group-level properties of the data that are not based on either group size or the number of groups in the sample. Group means the same author vs. different author grouping in this study.

1. Correlation and ICC

'Same author' studies and 'same data' studies have a relatedness and similarities. Typical correlations can not measure this kind of relatedness directly because there are no matched samples, or pairs of scores. 'Same author' studies and 'same data' studies are groups of studies, so their characteristics are similar to a nested data situation. So, to measure this kind of relationship, the intraclass correlation is considered as an alternative method to measure the dependency in 'same author' and 'same data' studies.

The intraclass correlation is based on variance partitioning like the homogeneity test, but gives a different perspective on the 'same author' and 'same data' issues. Donner (1986) explained that the intraclass correlation coefficient has a long history of application in several different fields of research: epidemiological research for familial resemblance, psychology for reliability theory, genetics for the heritability of selected traits, and sensitivity analysis for effectiveness of experimental treatments. These fields seem to have parallels with the 'same author' studies and 'same data' studies. The situation suggests a need to find some relatedness among groups or samples within groups, compared to variation between groups. Murray and Blistein (2003) conducted research on "Group-randomized trials (GRTs)". GRTs have a hierarchical or nested data structure with members nested within groups, much like studies nested in the 'same author' or using the 'same data'. The research of Hannan et al. (1994) shows that community trials involve the assignment of intact social groups to study conditions and are getting popular in epidemiological research. Such studies have used the intraclass correlation coefficient for measuring the effect of treatment differences between groups. The similarities are in the intact groups, and in the nested and hierarchical data structures to which the intraclass correlation is applied.

However, the reason why researchers in other fields measure the intraclass correlation is different from the ‘same author’ and ‘same data’ issues. In the above research fields using the intraclass correlation, they want to measure the treatment effect without inflating type I error due to the dependence caused by the nested data structure. Wampold and Serlin (2000) described three reasons why researchers have ignored nested factors in treatment studies: analysis is simpler when data are treated as independent, an inflated type I error rate may lead to an inflated treatment effect, and analysts have emphasized effect sizes rather than statistical significance, so researchers have ignored the dependence while analysts pay attention to effect size. Ignoring the ICC can cause inflated type I error rates and inflated effect sizes (Wampold & Serlin, 2000). In the ‘same author’ studies and ‘same data’ studies, the ICC can show the dependence within ‘same author’ studies and within ‘same data’ studies, which ordinary correlations can not show. ICCs are used to evaluate the group level properties of data, or the ratio of between group variance to total variance (Castro, 2002). The ICC is useful for measuring the degree of dependence of observations within a group and the ICC can be defined using the decomposition of the variance in a random effects model (Commenges & Jacqmin, 1994). The most fundamental interpretation of an ICC is that it is a measure of the proportion of variance that is attributable to study characteristics like the ‘same author’ factor and ‘same data’ factor (Shrout & Fleiss, 1979). However, for ‘same author’ studies and ‘same data’ studies, the ICC can be used to check if there is more relatedness within the ‘same author’ studies and ‘same data’ studies rather than for different author studies and different data studies. The reviewer will categorize the primary studies into two sets based on the ‘same author’ factor and ‘same data’ factor. In these groups of ‘same author’ studies and ‘same data’ studies, the ICC is used as an index to check whether the ‘same author’ and ‘same data’ studies variable explains any of the variance in the meta analysis.

This is very similar to Q_{BET} , but the ICC should indicate the amount of variation explained by the group-level properties of data.

Shrout and Fleiss (1979) introduced six different forms of the ICC for estimating rater reliability. Shrout and Fleiss (1979) consider the ICC as a reliability index, but the concept is similar to that used in the HLM context, "Many of the reliability indices available can be viewed as versions of the intraclass correlation, typically a ratio of the variance of interest over the sums of the variance of interest plus error" (p. 420).

Table 1: Comparison of two ICC approaches

	Rater reliability (Shrout and Fleiss 1979)	HLM (Raudenbush and Bryk 2002)
1. Definition of ICC	ICC = $\frac{MS_b - MS_w}{MS_b + (N_g - 1) * MS_w}$, Where MS$_b$ is the between-group mean square, MS$_w$ is the within-group mean square, and N$_g$ is the group size (Castro 2002). Form of variance ratio: $\rho = \sigma_T^2 / (\sigma_T^2 + \sigma_W^2)$	$\rho = \tau_{00} / (\sigma^2 + \tau_{00})$
2. Statistical Model of ICC	One way ANOVA $x_{ij} = \mu + b_j + w_{ij}$ μ is the overall population mean of the ratings b_{ij} is the difference from μ of the jth target's true score w_{ij} is residual component	The one-way ANOVA The level-1 or student-level model is $Y_{ij} = \beta_{0j} + r_{ij}$, At level 2 or the school level, $\beta_{0j} = \gamma_{00} + \mu_{0j}$ $Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}$
3. Unequal sample size consideration	$N_g = \frac{1}{k}[\sum_{i=1}^{k} N_i - \frac{\sum_{i=1}^{k} N_i^2}{\sum_{i=1}^{k} N_i}]$	

When I reviewed the literature on the use of the ICC, I found that the ICC was often used as an inter-rater reliability measure (including coder reliability in meta-analysis), and also used in HLM. These two ICCs have different formulas, but are almost the same in Table 1. The one-way random-effects model will be used to estimate ICC because this represents a nested design, with unordered observations nested within groups (McGraw & Wong 1996).

In the general calculation of ICC as a reliability estimate, equal sample sizes across groups are often assumed. However, in meta-analysis, different sample sizes across groups are more typical. Bliese and Halverson (1998b) note that articles discussing the calculation of the ICC rarely address the issue of unequal group sizes. Bliese and Halverson (1998b) presented a formula for unequal sample size, noting "Blalock (1972) presents a formula from Haggard (1958) recommending that N_g be calculated as follows:

$$N_g = \frac{1}{k}[\sum_{i=1}^{k} N_i - \frac{\sum_{i=1}^{k} N_i^2}{\sum_{i=1}^{k} N_i}]$$

where N_i represents the number of cases in each group, and k represents the number of groups" (Bliese & Halverson, 1998b, p. 168)

The ICC in the HLM approach also supposes an one-way ANOVA model with a random effect in HLM. The definition of ICC is $\rho = \tau_{00} /(\sigma^2 + \tau_{00})$. "This coefficient measures the proportion of variance in the outcome that is between groups (i.e., the level-2 units). It is also sometimes called the cluster effect. It applies only to random-intercept models (i.e., τ_{11} = 0)" (Raudenbush & Bryk, 2002, p. 36). The one-way ANOVA model with random effects usually presents preliminary information about partitioning variance in the outcome into within and between groups portions. The model of one-way ANOVA in HLM can be applied directly to meta-analysis (Raudenbush & Bryk, 2002, p. 69-70), noting "The level-1 is $Y_{ij} = \beta_{0j} + r_{ij}$, assuming r_{ij}~ independently N $(0, \sigma^2)$ for i =1, to n_j effect size in study j, and j = 1, … J studies. σ^2 is the level-1 variance".

Notice that this model characterizes the effect size in primary studies in meta-analysis with just an intercept, β_{0j}, which in this case is the mean effect size across studies. At level 2 or the study level, each author's mean effect size, β_{0j}, is represented as a function of the grand mean, γ_{00}, plus a random error, μ_{0j}:

$\beta_{0j} = \gamma_{00} + \mu_{0j}$, assuming μ_{0j} is distributed independently N (0, τ_{00}). τ_{00} is the second level variance. This yields a combined model, also often referred to as a mixed model, with fixed effect, γ_{00}, plus two random errors, μ_{0j} and r_{ij}: $Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}$.

The formula for the ICC in rater reliability is almost same to the ICC in HLM approach. Using this formula, the ICC coefficient is the percent of variance explained by between group variables (Raudenbush & Bryk 2002, p. 69-70). The ICC will now be used to estimate the effect of 'same author' and 'same data' variables. I expect the between studies variance of the same author/data studies will generally be smaller than that of different author/data studies when examined in terms of the ICC.

Sensitivity analysis

The best way to deal with ‘same author’ and ‘same data’ issues remains open to debate and exploration. Sensitivity analysis is a way to discover the impact of particular problematic observations, like outliers in ordinary statistical analysis. In this research, ‘same author’ and ‘same data’ studies in meta-analysis may be problematic, so we may want to check the impact of removing them on the synthesis of research findings.

Rappaport (1967) said, “The ‘what if’ question may be viewed as an introduction to sensitivity analysis in the face of uncertainty. Uncertainty refers to situations for which probabilities of outcomes cannot even be predicted in probabilistic terms” (p. 441). In synthesizing research results, no one can predict the effect of ‘same author’ and ‘same data’ studies even though it is essentially related to the generalizability of synthesizing research results. In synthesizing research findings, if ‘same author’ papers and ‘same data’ papers will are very similar and dependent, and the impact of these papers is larger than that of other papers, it will be very difficult to safely generalize the results of the meta-analysis. In that case, reviewers need to distinguish the same author/data papers from the overall synthesis results by sensitivity analysis. In every stage of the meta-analysis, reviewers make decisions: which papers to gather, which papers to include in the analysis, which analysis results to report. The decision will have an impact on the results of estimating effect size like sensitivity analysis.

Greenhouse and Iyengar (1994) also mentioned, “Since at every step of a research synthesis decisions and judgments are made that can affect the conclusions and generalizability of the results, we believe that sensitivity analysis has an important role to play in research synthesis” (p. 397). They think that the findings of meta-analysis will convince a wider range of readers if meta-analysts assess the effect of these decisions and judgments. “At every step in a research synthesis decisions are made that can affect the conclusions and inferences drawn from the analysis. It is important to check how sensitive the conclusions are to the method of analysis or to changes in the data” (Greenhouse & Iyengar, 1994, p. 384). Each decision can affect the generalizability of the conclusions of the meta-analysis: ‘same author’ papers and ‘same data’ papers are dependent, if reviewers decide not to include all the same author/data papers, they will lose much information. If reviewers decide to include all the same author/data papers, it will be problematic because of the violation of independence assumption.

The present research will investigate the same author/data issues from the perspective of sensitivity analysis as follows: Effect size of all studies vs. effect size of all studies excepting the same author/data group studies

The sensitivity analysis results of removing the ‘same author’ and ‘same data’ studies will have an effect on the direction and amount of effect sizes for all studies. So, if the meta-analyst removes the same author/data studies that have a smaller effect size than the mean effect size of all studies, the overall summary measure will increase, and vice versa. Meta-analysts need to present the various results of sensitivity analysis to readers considering ‘same author’ and ‘same data’ factors. This is a way to deal with these issues for remaining open to debate and to explore research findings in case of having ‘same author’ and ‘same data’ studies in synthesizing research findings. This sensitivity analysis focuses on the effect size of all studies vs. the effect size of all studies except the same author/data studies, while the categorical analysis pays attention to the effect size of same author/data studies vs. different author/data studies without considering the direction and effect of each on the overall effect size.

Based on sensitivity analysis results, reviewers need to investigate why the results differ between the effect size of all studies vs. the effect size of all studies except same author/data papers.

HLM

In meta-analysis, the results of primary studies are inconsistent because of various reasons. Meta-analysts investigate the homogeneity, and if the results are not homogeneous, meta-analysts will check the study characteristics to explain the variation between studies. This research focuses on 'same author' and 'same data' factors because these factors can influence the variability in estimating effect size. The similarity of 'same author' studies and 'same data' studies is investigated by using HLM.

1. Advantages of meta-analysis using HLM

HLM provides a useful framework for addressing the problem of components of variability in meta-analysis (Raudenbush & Bryk, 1985). The hierarchical model will show the variability that result from different study characteristics. In 'same author' studies and 'same data' studies, analysts can misestimate standard errors because the analysts fail to take into account the dependence within 'same author' and 'same data' studies. "Hierarchical linear models can resolve this problem by incorporating into the statistical model a unique random effect for each within study and between study units" (Raudenbush and Bryk, 2002, p.100).

2. Nested structure in meta-analysis

The difference between the traditional linear model and hierarchical linear model is the data structure. The traditional model assumes that subjects respond independently, but the hierarchical linear model supposes a "nested" data structure. Subjects are nested in the organizations where the subjects belong, like classrooms and schools. When analysts ignore the nested structure of data, it leads to the problem of aggregation bias and misestimated results (Raudenbush, 1988). Raudenbush and Bryk (2002) explained the hierarchical structure of meta-analytic data (p. 206). In meta-analysis, analysts need a model to take into account variation at the subject level and study level because subjects are nested within studies. Meta-analysts need to learn about sources of variation among subjects and to sort out variation across studies.

3. Nested structure and homogeneity

In meta-analysis, to investigate the variance between studies is very important. Little variation means that the results of primary studies are very similar to each other. The nested data structure in HLM is very closely related to homogeneity issues in meta-analysis because nested samples are more similar to each other than people randomly sampled from the entire population. Nested samples have similar experiences which may lead to increased homogeneity over time (Osborn, 2000).

It is not satisfactory to treat 'same author' studies and 'same data' studies independently or to average results for the 'same author' studies and 'same data' studies. Meta-analysts need to disentangle a study effect and group of studies effects (same author/data factors) in estimating effect size.

4. Same author/data characteristics: Nested and unbalanced

To show the similarities and impact of 'same author' studies and 'same data' studies, the characteristics of 'same author' studies and 'same data' studies need to be clarified. 'Same author' studies and 'same data' studies have two characteristics: one is nested within 'same author' studies and 'same data' studies, and the other is unbalanced in the number of studies within 'same author' groups or 'same data' groups.

First, a nested data structure means that if one author writes several papers on a single issue or if different authors write many papers using common public data sets, these papers will be more similar than papers by different author and papers using different data sets. It is a very similar situation to that of students nested in a classroom, school and district.

Second, an unbalanced data structure means that every study has an author and data set, but if a meta-analyst tries to group effects using a same author/data factor, the number of papers per author/data set is not likely to be balanced. Structured experiments or surveys with equal cluster sizes will have a balanced data structure. Sliwinski and Hall (1998) described the unbalanced situations that can occur in multi-site studies because there are unequal numbers of observations within each unit of analysis and within experiments. The multi-site studies are very similar to same author/data studies from the perspective of unbalanced data structure. A multi-site study often is used in research to study the effectiveness of similar interventions or variations of the same intervention across multiple similar or distinct sites (Kalaian, 2003).

Nested and unbalanced data structures are the main characteristics of same author/data studies in meta-analysis. These characteristics suggest the use of the HLM approach to investigate the relationship of the same author/data studies in meta-analysis

5. Dependence in meta-analysis

In meta-analysis, dependence is a very important topic in unbiased estimation of effect size. There are at least two kinds of dependences in meta-analysis. One is repeated measures' dependence and the other is the similarities or dependences based on nested study structure in meta-analysis. Two different labels of dependence will need different approaches in estimating effect sizes. Gurevitch and Hedges (1999) distinguish two fundamentally different dependences based on different sources of variation: one is within-study sampling error, the other is between-study variation.

The first dependence has several characteristics: replication of sample, correlated data, and within-study similarities. However, meta-analysts have rarely paid attention to dependences of between studies. Most meta-analyses have paid much attention to the variance component of primary studies in meta-analysis.

Gurevitch and Hedges (1999) distinguish between these two dependences in ecology research. The similarity of studies from the same laboratory is very close to the 'same author' studies. In studies from the same laboratories, the procedure, sampling and research method will be similar.

For the dependence based on the between studies variation, there are several similar situations including same laboratories, same site studies, same public data studies and same author studies. These studies will have less variability, and they will have non-zero intraclass correlations. The possible solution of this dependence is based on the nested study characteristics. Gurevitch and Hedges (1999) proposed a hierarchical linear modeling (HLM) approach for this dependence. The meta-analyst can estimate components of variance within and between studies to consider relatedness between studies. If the meta-analyst fails to disentangle the between study and within study effects, it leads to several problems in estimating effect sizes in meta-analysis. Gurevitch and

Hedges (1999) were concerned about the underestimation of the standard error of the mean effect, and liberal evaluation of the statistical significance of effects. Sliwinski and Hall (1988) also worried that a misleading R^2 could arise as a result of failing to disentangle between-study from within-experiment effects.

6. Function of HLM in meta-analysis

The purpose of same author/data meta-analysis using HLM is as follows: first is to estimate the variance of the effect size parameters in a random effect unconditional model. If there is heterogeneity in a random effect unconditional model, the second is to explain variation in the effect-size parameters in a fixed-effects conditional model incorporating the same author/data factor. A third goal is to estimate the residual variance of the effect-parameter in a random effects conditional model, and is to estimate how much variance is explained by the same author/data factor. Meta-analysis applications in HLM have several characteristics different from other HLM contexts (Raudenbush & Bryk, 2002): First, meta-analysis has no raw data. Meta-analysis usually uses summary statistics for analysis. Second, meta-analysis uses different outcome measures. However, the analysis can use different outcome measures by putting them onto a standardized scale such as by using the standardized mean difference effect size. Third, meta-analysis assumes each data point's sampling variance is known. Raudenbush and Bryk (2002) refer to this situation as the level-1 variance known problem: "The essential statistical features of meta-analysis applications that distinguish them from the others discussed in this book are two: Only summary data are available at level 1; and the sampling variance, V_j, of the level-1 parameter estimate, d_j, can be assumed known" (p. 217).

7. HLM analysis procedure in meta-analysis

Meta-analysis using HLM can be performed under the known variance model and HLM considers two analyses: unconditional and conditional models. Unconditional analysis is conducted first. The Unconditional analysis does not include any level-2 variables in the model. The unconditional analysis estimated the mean in the fixed-effects model and variance of the true effects in the random-effects model. If the results are homogeneous across studies, the next step is not performed. If the results show heterogeneity across studies, a conditional analysis is performed. Conditional analyses include level 2 variables. The conditional analysis estimate the effect size in the fixed-effects model and the residual variance of the effect sizes in the random-effects model. For this study, analysts run the unconditional model first, and if there is heterogeneity, analysts examine the ‘same author’ and ‘same data’ factors in the conditional model to explain the impact of these factors in meta-analysis.

Kalaian (2003) proposed a meta-analytic methodology as the application of the mixed-effects linear model via hierarchical linear modeling. A within-study model formulates each primary study effect size as a function of the true effect size and sampling error. “A between-studies model formulates the distribution of the true effect sizes from the within-study model as a function of study characteristics and random errors” (Raudenbush & Bryk, 2002, p. 209). Hierarchical linear modeling software needs information: the study identification numbers, study effect sizes, their variances, and coded study characteristics like the same author/data factors.

Within-study model

In the within-study model for the mixed-effects meta-analysis, the calculated study effect size, d_i, of study i, depends upon a population study effect size δ_i plus a random sampling error, e_i . Thus, the basic within-study model for study i can be represented as

$d_i = \delta_i + e_i$,   i = 1,2,….,J.

We assume e_i ~ N(0, V_i).

Between-studies model

In the between-study model, the population study effect size, δ_i , depends on study characteristics and a level-2 random error:

$$\delta_i = \gamma_0 + \sum_s \gamma_s W_{si} + u_i \ ,$$

where

W_{si} is a study characteristics predicting these effect sizes;

γ_s is regression coefficient; and

u_i is a level-2 random error for which we assume $u_i \sim$ N (0, τ).

8. Strength of HLM in meta-analysis: Intraclass correlation, and 'variance explained statistics'

Meta-analysis using HLM is a good way to estimate effect size in nested and unbalanced data structures like same author/data studies and it allows the meta-analysts to distinguish two variances (level-1 sampling error variance and level-2 random error) simultaneously.

When analysts use HLM approach for meta-analysis, the analysis will produce two pieces of the information: the intraclass correlation and 'variance explained' at level 2.

In the One-Way ANOVA model, the intraclass correlation represents the proportion of variance in effect sizes (d_j) between studies, via

$$\rho = \tau /(\tau + V\ )$$

The One-Way ANOVA model is an unconditional model which has no level 2 predictors, specifically

$$\delta_j = \gamma_0 + u_j\ .$$

When analysts take into account study characteristics like same author and same data factors, the analysts can use conditional model,

$$\delta_j = \gamma_0 + \gamma_1 W_{1j} + u_j$$

By comparing the τ estimates across the two models: the unconditional model and the conditional model, analysts can develop an index of 'variance explained' at level 2, specifically the proportion of 'variance explained' in δ_i is

$$= \frac{\tau\text{ (unconditional model) - }\tau\text{ (conditional model)}}{\tau\text{(unconditional model)}}\ .$$

Using the HLM approach for meta-analysis, we can get the intraclass correlation and 'variance explained' information. Intraclass correlations in the unconditional model indicate how much variance in effect size lies between studies, and the 'variance explained' indicates the percentage of variance accounted for by study characteristics like 'same author' and 'same data' factors.

The HLM approach efficiently performs two goals simultaneously compared to conventional meta-analysis, and gives a direct indication of the explanation of variability in the two models: the unconditional model and the conditional model.

CHAPTER IV

SIMULATION

Introduction

In this chapter, I describe how I generated data to represent the ‘same data’ and ‘same author’ situations. These data are used to examine the proposed methods and investigate several possible factors in meta-analysis such as the number of studies, the sizes of samples (study size), and magnitudes of effect size. For the ‘same data’ issue, raw scores are generated, and the effect of overlapping samples is investigated. For the ‘same author’ issue, effect sizes are generated, and the author effect is examined as the main study characteristic.

‘Same data’ issue

In order to examine the proposed analysis methods for ‘same data’ studies, raw data sets similar to the STAR data set are generated. A simulated data set should be similar to a real situation. Thus, the data structure and characteristics of the STAR data set are described briefly. The STAR studies began in 1985 and finished in 1989. A total of 79 schools participated and the number of students in small classes was 13-17, and the number of students in large classes was 22-25. There were 108 small classes and 101 regular classes in the 79 schools.

Data generation for ‘same data’ studies requires a two-stage process. At the first stage I generate population data sets, and at the second stage, I extract sub-samples and report effect sizes of sub-samples which form hypothetical primary studies.

First, I generate the population data sets considering the number of schools, number of small classes and large classes per school, and the number of students per small class and large class. I use a three level data generation model including random effects at each level. To generate the student test scores, the generated data include school effects, a class-size effect, and between student variance for the student test scores. School effects and class-size effects also include a variance for each effect in the data generating process.

The data generating model given here generates test scores for students in small classes. Specifically, $T_{jkl} = \mu + \alpha + \beta_l + \gamma_{kl} + \eta_{jkl}$, with the following components:

T_{jkl} : Test score for student *j* in a small class,

μ : Mean for a large class,

α : Mean difference between small classes and large classes,

β_l : School effect in school *l*, where we assume β_l are distributed independently as N (0, σ_β^2),

σ_β^2 : Between school variance,

γ_{kl} : Class effect in class *k* in school *l*, where we assume γ_{kl} are distributed independently as N (0, σ_γ^2),

σ_γ^2 : Between class variance,

η_{jkl} : Effect for student *j* in class *k* in school *l*, where we assume η_{jkl} are distributed independently N (0, σ_η^2), and

σ_η^2 : Between student variance.

To generate test scores for students in large classes, the mean difference $\alpha = 0$ and the other terms are the same as for small-class students' test scores.

After generating the population data of the student test scores, school indicators, and class-size indicators, at the second stage I extract random sub-samples from the generated population data set. Randomly making sub-samples is similar to generating a "hypothetical primary study" in the real 'same data' situation. A researcher would select a certain number of small classes and large classes from the larger data set, and get test scores for all of the students in these classes, for their own study. The main point of this simulation is to investigate the effect of taking overlapping nonindependent samples from the 'same data' set. Thus, the next step is to replicate making the random overlapping sub-samples, and to calculate an effect size based on each randomly extracted sub-sample. Using the studies generated via the above process, the methods proposed in chapter 3 are examined.

The simulation factors in this research are

a) The size of the large initial data set (N = 10,000, 50,000, or 100,000 cases),

b) The overlap ratio of the number of cases in all sub-sample pseudo studies (Σn_i) to the size of the initial data set (N) as an indicator of overlap: Ratios 0.04, 0.1, 0.2 and 0.5 indicate small degrees of overlap, the ratios 1 and 3 represent medium overlap and the ratio of 5 means much overlap among the sub-samples.

c) The magnitude of effect size (δ = 0.2, 0.5, and 0.8 defined below), and

d) The ratio of pseudo-study-size (n_i) to N, the size of the population data set, with values 0.02, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5.

The size of the large initial data set represents the size of a hypothetical public data set such as the STAR data in the 'same data' situation. The factors b) and d) are the main factors for estimating effect size in meta-analysis. The reason we consider the factor b) is to examine the effect of having overlapping samples in the 'same data' situation. The population effect size is calculated as $\delta = \alpha / \sigma_{score}$, where $\sigma^2_{score} = \sigma^2_{\beta} + \sigma^2_{\gamma} + \sigma^2_{\eta}$. In the effect-size calculation, the generated student test scores include the school effect, class-size effect, and within-class student effect. Thus, σ^2_{score} includes the variances of school, class, and student effects together and is the pooled variance, and α is the mean difference between small-class and large-class test scores.

Examination of proposed statistical methods

For evaluation of each condition of generated 'same data' studies, I summarize and make graphs for *H, Q,* the width of the confidence interval (CI), or the between studies variance based on the 1,000 replications. Then I compare these values for different simulation conditions.

I investigate the results of these generated data sets to determine if the analysis results are consistent with expectations, and with the results of the empirical analyses in chapter 5. We might expect that the more the sub-samples overlap, the more similar outcomes will be, according to *H, Q,* the width of the CI, and between studies variance. The simulation factors b) (overlap ratio) and d) (study-size ratio) are of primary interest in this simulation.

Results

This study aims to show the extent of dependence between studies, and focuses on the variability of the 'same author' studies and 'same data' studies. The measures of variability are related to the homogeneity test in meta-analysis. Several researchers have reported that measures of variability and homogeneity test values are dependent on sample size and number of studies in the meta-analysis (Higgins & Thompson, 2002; Kim, 2000; Rucker, Schwarzer, Carpenter, & Schumacher, 2008). I proposed several methods to show the between studies dependence. The results are expected to reflect which methods are dependent on the sample size or number of studies in this simulation.

Table 2 present the characteristics of data generation for the 'same data' issue, for an initial data set of 10,000 cases. The number of studies is dependent on the overlap ratio, so as the overlap ratio increases, the number of studies increases automatically in Table 2. Study-size ratio is the ratio of study size (n_i) divided by the initial data set size *N*. Overlap ratio is the ratio of total sample size (Σn_i) divided by the initial data set size *N*.

Table 2: *Characteristics of ‘same data’ simulation data generation (Initial data set size N 10,000)*

Study ID	Initial data set size N	Sample size n_i	Study-size ratio n_i /N	Number of studies k	Total sample size Σn_i	Overlap ratio $\Sigma n_i/N$	Effect size
*A**4	10000	200	0.02	2	400	0.04	0.2
A10	10000	200	0.02	5	1000	0.1	0.2
A20	10000	200	0.02	10	2000	0.2	0.2
A50	10000	200	0.02	25	5000	0.5	0.2
A100	10000	200	0.02	50	10000	1	0.2
A300	10000	200	0.02	150	30000	3	0.2
A500	10000	200	0.02	250	50000	5	0.2
B10	10000	500	0.05	2	1000	0.1	0.2
B20	10000	500	0.05	4	2000	0.2	0.2
B50	10000	500	0.05	10	5000	0.5	0.2
B100	10000	500	0.05	20	10000	1	0.2
B300	10000	500	0.05	60	30000	3	0.2
B500	10000	500	0.05	100	50000	5	0.2
C20	10000	1000	0.1	2	2000	0.2	0.2
C50	10000	1000	0.1	5	5000	0.5	0.2
C100	10000	1000	0.1	10	10000	1	0.2
C300	10000	1000	0.1	30	30000	3	0.2
C500	10000	1000	0.1	50	50000	5	0.2
D60	10000	2000	0.2	3	6000	0.6	0.2
D100	10000	2000	0.2	5	10000	1	0.2
D300	10000	2000	0.2	15	30000	3	0.2
D500	10000	2000	0.2	25	50000	5	0.2
E60	10000	3000	0.3	2	6000	0.6	0.2
E90	10000	3000	0.3	3	9000	0.9	0.2
E300	10000	3000	0.3	10	30000	3	0.2
E500	10000	3000	0.3	17	51000	5	0.2
F80	10000	4000	0.4	2	8000	0.8	0.2
F120	10000	4000	0.4	3	12000	1.2	0.2
F320	10000	4000	0.4	8	32000	3.2	0.2
F480	10000	4000	0.4	12	48000	4.8	0.2
G100	10000	5000	0.5	2	10000	1	0.2
G300	10000	5000	0.5	6	30000	3	0.2
G500	10000	5000	0.5	10	50000	5	0.2

1) *In the study ID, A, B, C, D, E, F and G represent the study-size (n_i)

2) **In the study ID, 4, 10, 20, 50, 100, 300, and 500 represent the overlap ratio: $\Sigma n_i/N$*100%

Next I discuss results of the simulation. All figures follow the description of these results. Overlap ratio in Table 2 expressed as the percent of overlap ratio in Figures 1-6 (overlap ratio × 100).

H

A value of H = 100% indicates perfect homogeneity because this indicates the fixed-effects confidence interval width exactly matches with the random-effects confidence interval width. When many studies examine the ‘same data’ set, meta-analyses with more overlapping studies will likely be more homogenous. So, meta-analyses with more overlap among studies are expected to have higher H values compared to other meta-analyses.

Figure 1 shows the H values calculated for the ‘same data’ studies. Seven different overlap ratios ($\Sigma n_i/N$) represent 4%, 10%, 20%, 50%, 100%, 300% and 500% overlap for the ‘same data’ studies as described in Table 2. For the smaller study-size ratios (n_i/N) such as 2%, 5% and 10%, H does not show the expected trend, possibly because smaller studies from the ‘same data’ will not have much similarity. For the study-size ratio over 20%, Figure 1 shows that larger overlap ratios have higher H values as we expected. This indicates that the overlap ratio effect will not be large if the study-size ratio is less than 20%. In the ‘same data’ situation, if a researcher uses only small samples (e.g., the study-size ratio is less than 20%), the similarities among ‘same data’ studies will not be large.

Figure 1 also shows the effect of study-size ratio (n_i/N). The larger study-size ratios show larger H values as I expected.

Functions of *Q*

Based on functions of Q, the percent of significant Q tests and the averages of the Q statistic values are next summarized. Also the Birge ratio, a function of Q, is discussed.

To check homogeneity, in general, meta-analysts use the Q statistic (Cooper & Hedges, 1994):

$$Q = \sum_{i=1}^{k} [(T_i - \overline{T}_{\bullet})^2 / v_i],$$

where T_i is the observed effect size of *i*th study, $\overline{T}_{\bullet}$ is the weighted average effect size, and v_i is the conditional variance of the *i*th standardized mean effect size.

Generally speaking, if *Q* is greater than the critical value of a chi-square with *k*-1 degrees of freedom, "the observed variance in study effect sizes is significantly greater than what we would expect by chance if effect sizes in all studies shared a common population effect size" (Cooper & Hedges, 1994, p. 266). If *Q* is significant, this indicates the estimated effect sizes are heterogeneous, and the effect sizes do not share a common population effect size. The Birge ratio, *Q*/(*k*-1) will be large when effects are heterogeneous, and vice versa.

1) Birge ratio

Figure 2 shows that the Birge ratio values are inconsistent (decreasing, flat, or increasing) but fairly similar as the overlap ratio increases. However, Figure 2 shows that the Birge ratio values decrease as the study-size ratio increases. The Birge ratio values show a pattern similar to that of the between studies variance in Figure 6.

2) Percent of significant *Q* values

Figure 3 shows that, in general, the percent of *Q* values significant at the .05 level decreases as the overlap ratio increases. However, when the study-size ratio is small, specifically when the study-size ratio is 2%, the pattern is not exactly as we expected. Figure 3 shows that for all other study-size ratios, the percent of *Q* values significant at the .05 level decreases as the overlap ratio increases. Figure 3 also shows that the percent of significant *Q* values significant at the .05 level decreases as the study-size ratio increases. This is a pattern similar to that for *H* and the Birge ratio

I checked the confidence interval (CI) of the percent of significant *Q* values based on the formula, 0.05±1.96*SE*, with $SE = \sqrt{\frac{\alpha*(1-\alpha)}{N_{reps}}} = \sqrt{\frac{.05*(.95)}{1000}} = 0.007$. So, the CI goes from 0.036 to 0.064 if the null hypothesis of homogeneity for *k* independent effects were true. This indicates that the percent of significant *Q* value would be between 0.036 and 0.064. Figure 3 shows the number of significant *Q*s is within the 95% confidence interval even when study-size ratio (n_i/N) is 2%.

However, for the study-size ratios larger than 2%, Figure 3 shows that the percent of *Q* values significant at the .05 level is much smaller than the 95% confidence interval limit. This indicates that the 'same data' studies with large study-size ratios have large similarities.

3) Average *Q* values

Figure 4 needs a more complex interpretation, because according to statistical theory *Q* increases as the number of studies increases. However, Figure 4 shows that *Q* decreases as the study-size ratio increases.

The real issue here is whether the mean *Q* is lower than expected. Under H_0, the means should be at or near the approximate *df* in Figure 4. So for $k = 250$, $df = 249$, $k = 150$, $df = 149$ etc. The case where the means are really where they should be seems to be when study-size ratios are small such as 2%, 5%, and 10% in Figure 4. The average *Q* values show a pattern similar to that of the *H* in Figure 1.

Width of confidence intervals

Figure 5 shows the effect of overlapping studies on fixed-effects and random-effects confidence interval widths. Viechtbauer (2007) said, "it may be also be useful to report a confidence interval for the amount of heterogeneity, which not only indicates the precision of the heterogeneity estimate, but also communicates all the information contained in corresponding homogeneity tests" (p. 38). Figure 5 shows that the confidence interval widths of higher overlap studies are smaller than those of other meta-analyses. Figure 5 also shows the study-size ratio's effect on the width of the confidence interval. Both confidence intervals decrease as the study-size ratio increase, as expected.

Between-studies variance in effect sizes

The amount of variance among the effect-size estimates is investigated. When many studies examine the 'same data' set, meta-analyses with more overlap among studies will be more homogeneous. So, meta-analyses with more overlap among studies are expected to have smaller between studies variance, compared to other meta-analyses. Figure 6 shows that the between studies variance is fairly consistent as the overlap ratio increases. However, Figure 6 clearly shows that between studies variance decreases as study-size ratio increases. This is a pattern to similar to that for *H* and the Birge ratio.

Conclusion

Values of *H*, the Birge ratio, percent of significant *Q* values, the width of the CI, and between studies variance worked well to show dependence. The *Q* average values are dependent on the number of studies (*k*) because *Q* is a chi-square with degrees of freedom equal to *k*-1.

The first lesson in this simulation is that reviewers should pay attention to study-size ratio in addition to overlap ratio, if many ‘same data’ studies exist when doing meta-analysis. The study-size ratio is another factor that helps to determine the similarities among ‘same data’ studies in this simulation. As I discussed before, *H* does not show the expected trend when study-size ratio is 2%, 5% and 10%. The pattern of percent of significant *Q* values is not exactly as we expected when the study-size ratio is 2%. Based on the results for *H* and percent of significant *Q* values, the reviewer should investigate the study-size ratio first. If the study-size ratio is less than 10%, we can assume the similarities among studies will not much matter. However, if the study-size ratio is larger than 20%, reviewers should pay attention to dependence issues for the studies using ‘same data’ sets. This is the most important finding in this simulation.

The other lesson for the ‘same data’ issue is that the homogeneity test based on *Q* for meta-analysis should be cautiously interpreted because the homogeneity test can be dependent on study size, number of studies, and similarities between studies. Reviewers need to check *H,* the Birge ratio, the width of the CIs and between studies variance in addition to the homogeneity test (*Q*) in meta-analysis.

Overall, when reviewers investigate the dependence of ‘same data’ studies, reviewers should pay attention to study-size ratio first, in addition to overlap ratio, and can use *H*, the Birge ratio, the width of the CI and between studies variance to represent the degree of dependence.

Results for other initial data set sizes (50,000 and 100,000 cases) are attached in Appendix J. Their results show the same pattern as is reported for the initial data sets of 10,000 cases.

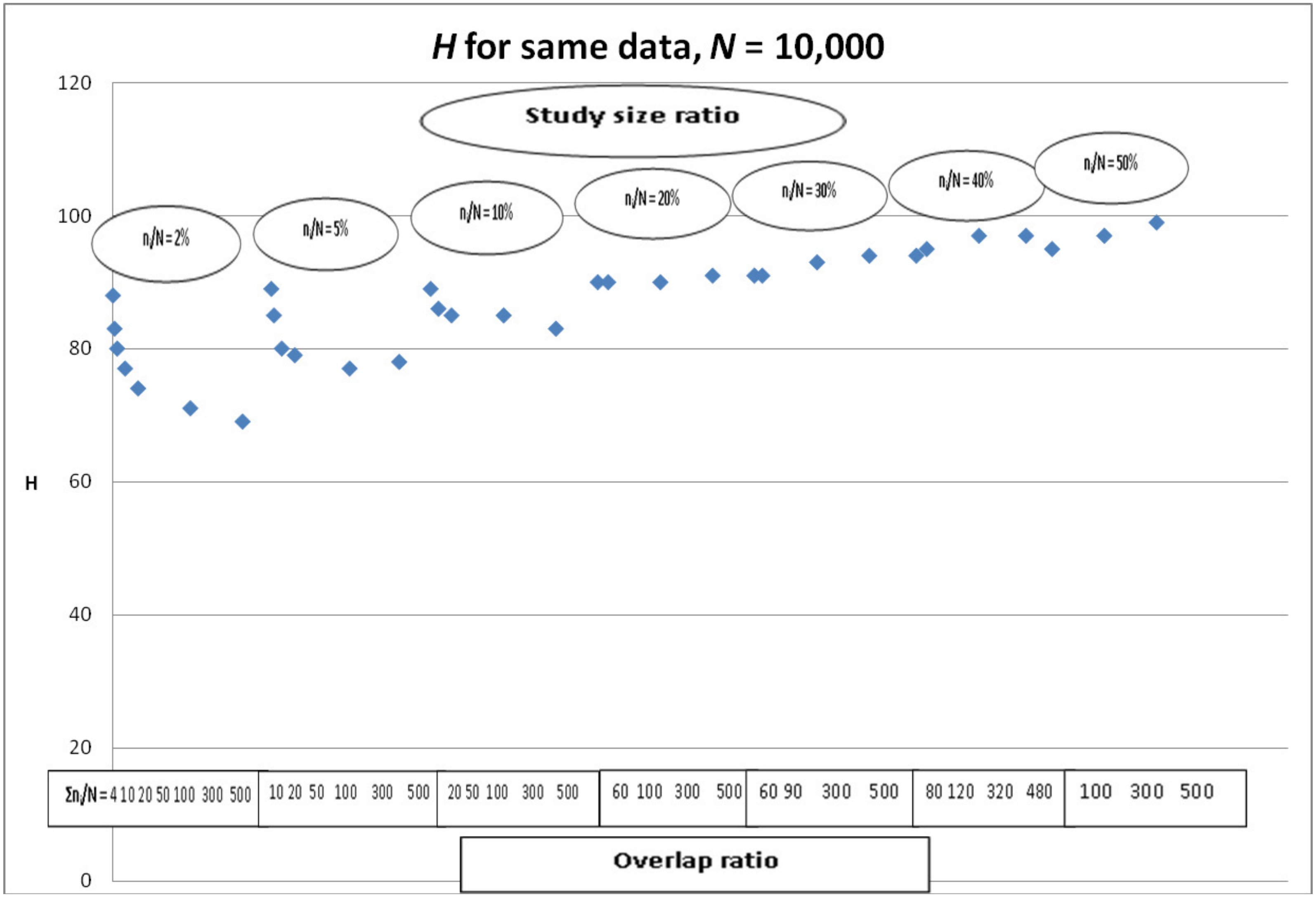


Figure 1: H for 'same data' issue for N = 10,000.

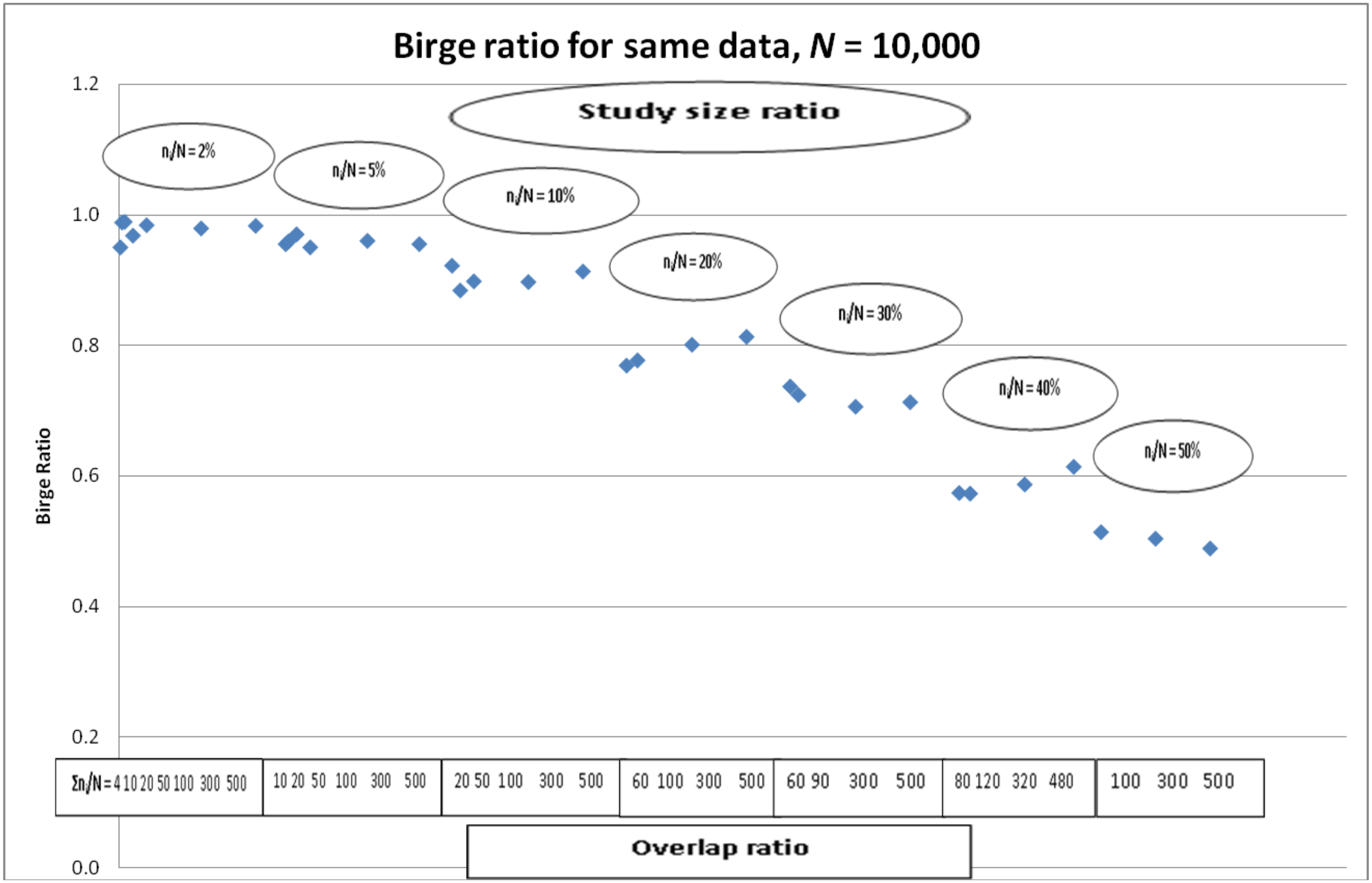


Figure 2: Birge ratio for ‘same data’ issue for $N = 10{,}000$.

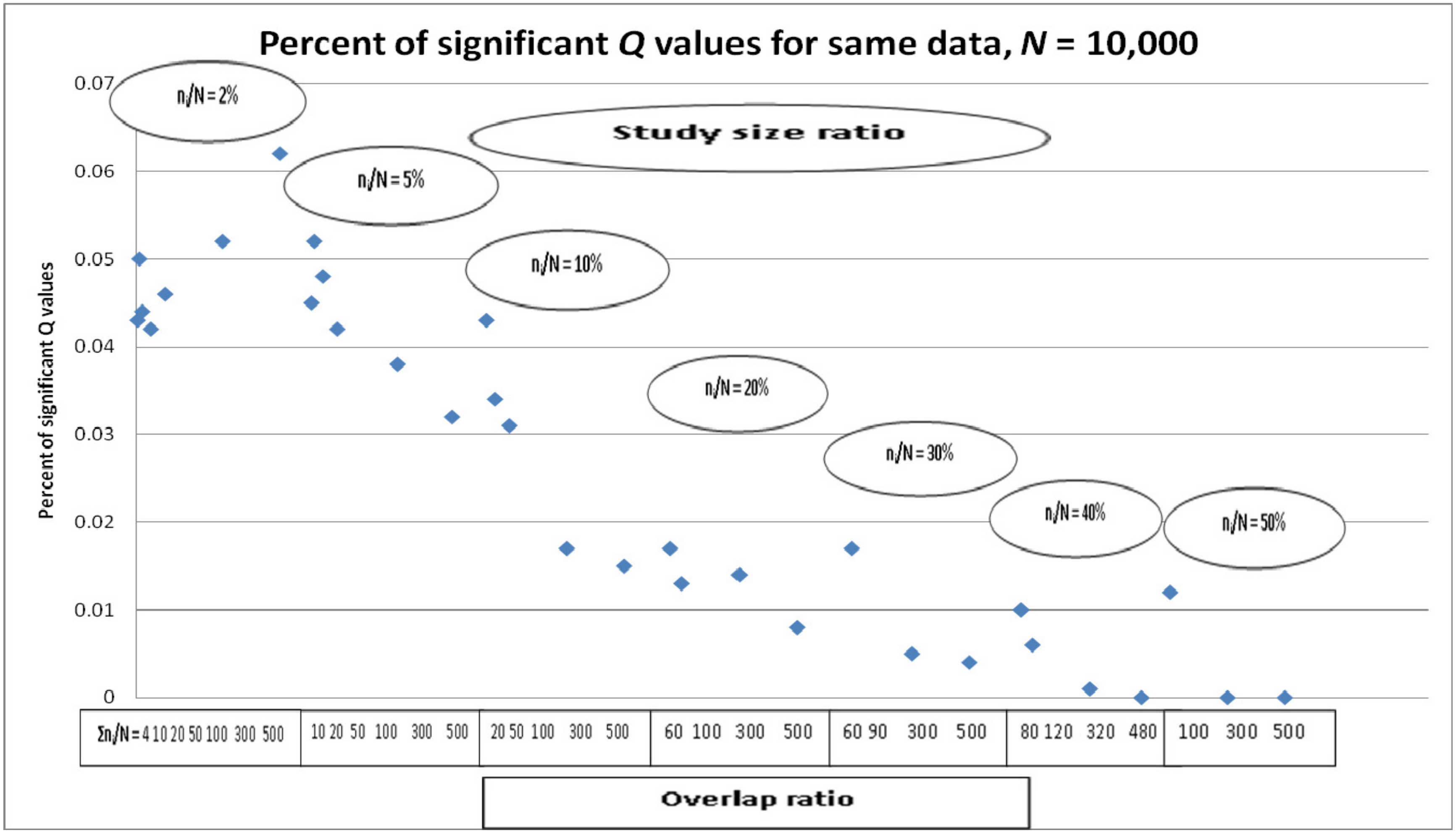


Figure 3: Percent of significant *Q* values for ‘same data’ issue for *N* = 10,000.
* In the Y-axis, 0.01 to 0.07 represent 1% to 7%.

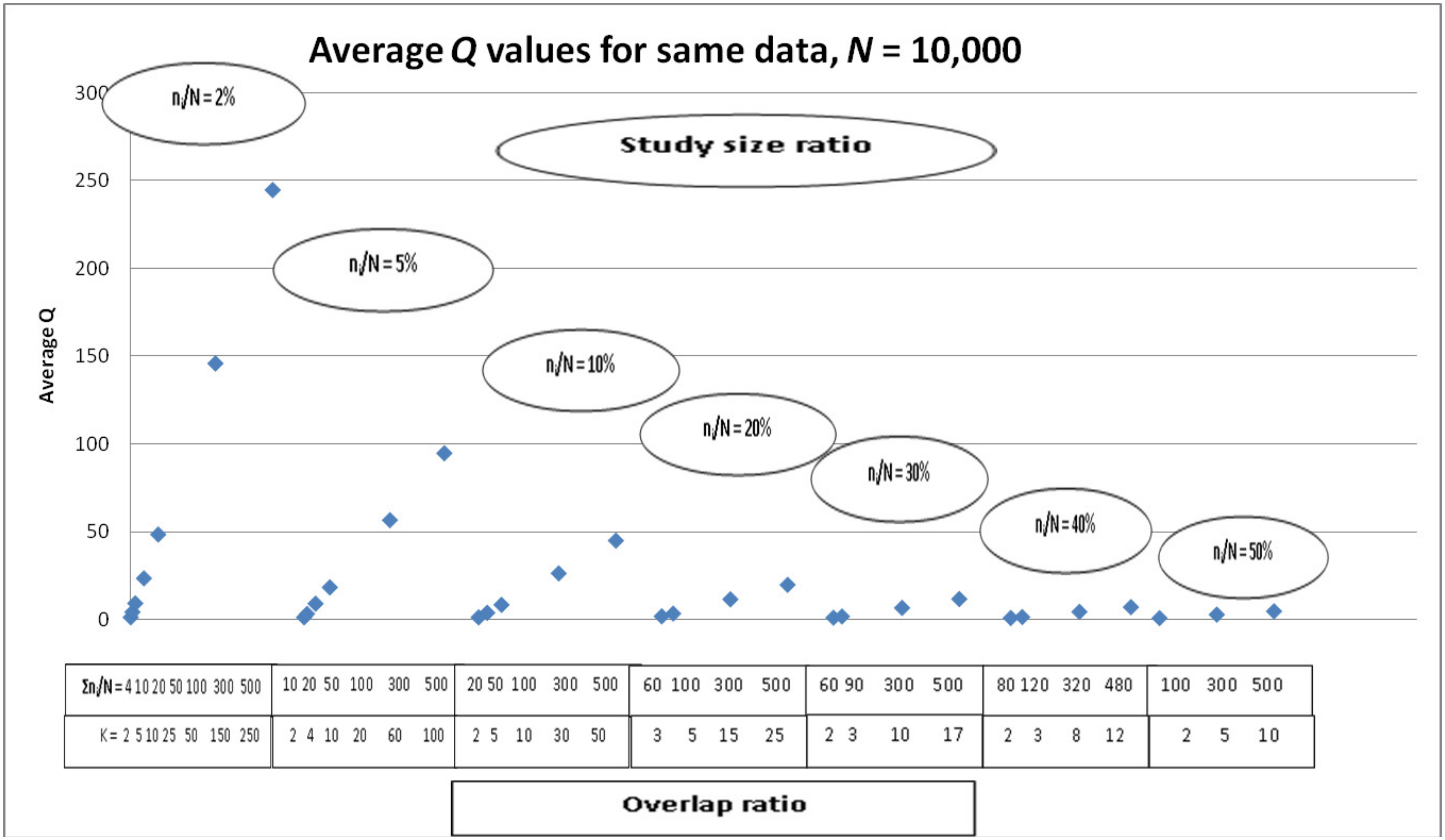


Figure 4: Average *Q* values for ‘same data’ issue for *N* = 10,000.

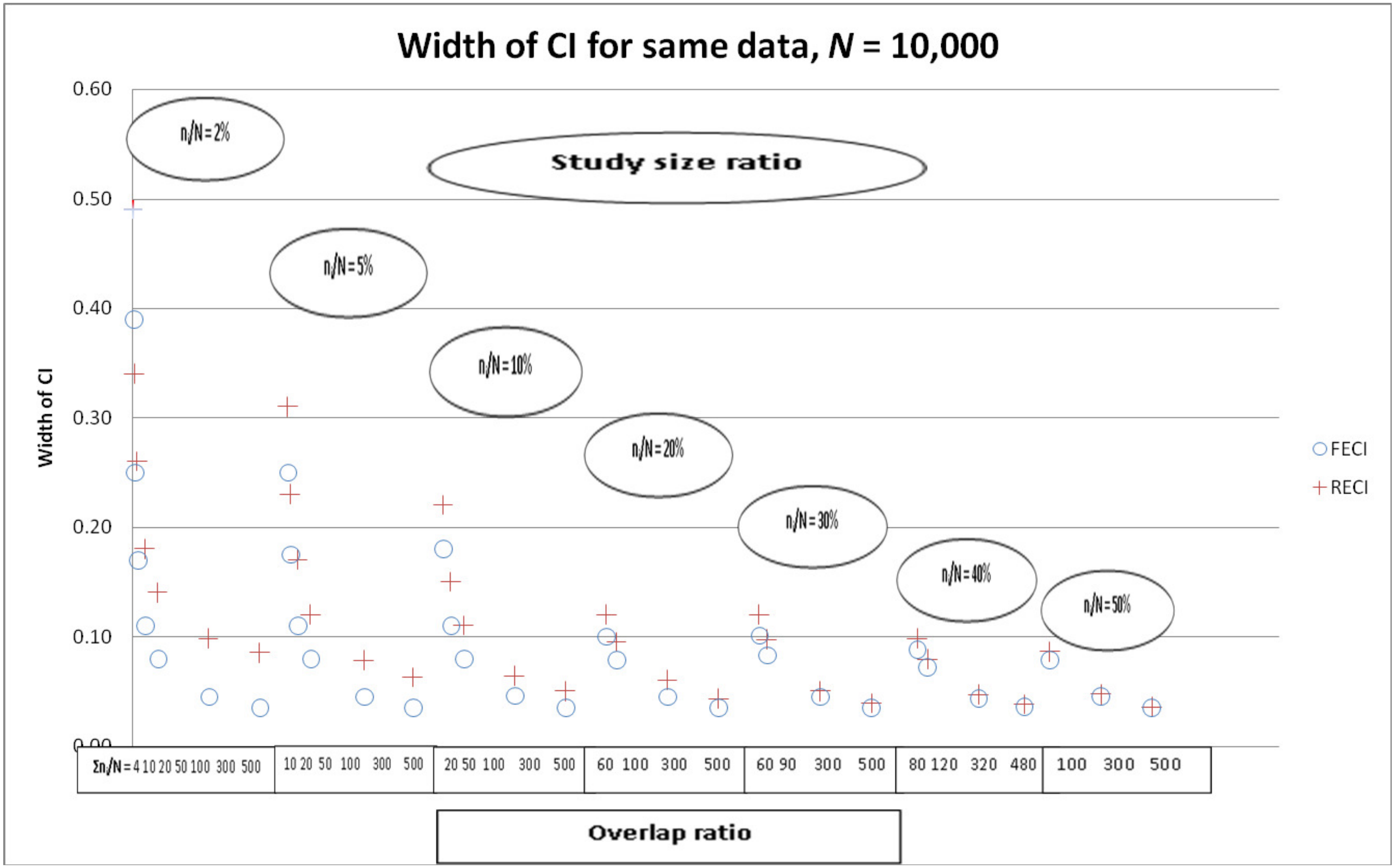


Figure 5: The width of the CI for 'same data' issue for N = 10,000.

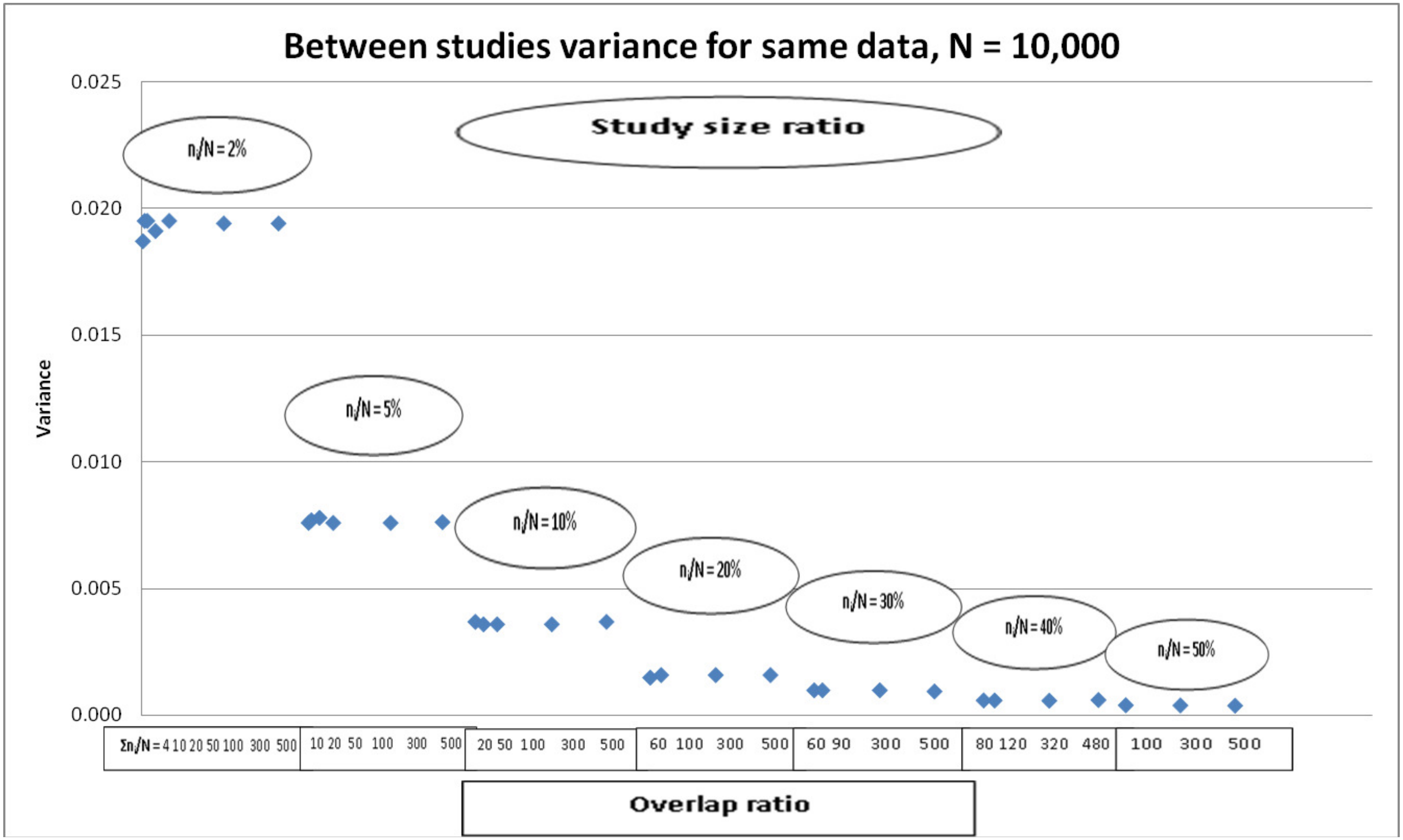


Figure 6: Between studies variance for 'same data' issue for $N = 10{,}000$.

‘Same author’ issue

Reviewers suppose that ‘same author’ studies will be similar and dependent with each other, however, it is very difficult to show their dependence. The ‘same author’ dependence is different from the dependence among papers using the ‘same data’, and the data generation will be different from the ‘same data’ situation. Instead of raw score generation, simulated hypothetical effect sizes are generated.

The assumption is that the ‘same author’ papers are nested within author. For the effect-size generation, author effects are considered using correlated effect sizes to represent hypothetical ‘same author’ papers.

Based on the literature review and case study of the ‘same author’ issue, we expect that each author will make similar choices of samples, measurement instruments, incentives for participants, and experimental conditions. However, it is difficult to generate such conditions for ‘same author’ studies in a simulation study. I will use different degrees of correlation between effect sizes to represent the ‘same author’ situation. I assume if the proposed methods in chapter 3 can detect the similarities among the correlated effects in the simulation study, the methods can also show the similarities among ‘same author’ studies in empirical meta-analyses.

For the ‘same author’ data generation, let us imagine a study of the relationship between SAT scores and college GPA. The effect size can be a correlation between SAT score and college GPA. One author may conduct several studies of this relationship using different samples.

The correlated effect sizes will be generated for the ‘same author’ situation considering study size (sample size), the number of ‘same author’ studies, the degree of correlation between the effect sizes, and the magnitude of effect size. However, I acknowledge that this correlated effect-size format is not the real covariance structure for correlational data in multivariate meta-analysis.

The objective of this study is to examine the methods proposed in chapter 3 to detect the dependence for the ‘same author’ situation, which is not the exactly the same as the multivariate meta-analysis situation, so I do not need to make the data correlated as in the Olkin and Siotani (1976) multivariate sense. In this study, I will generate equicorrelated effect sizes to represent hypothetical ‘same author’ studies. First, I describe how I generate equally correlated "true" effect sizes for the sets of studies done by each hypothetical researcher. Second, I also explain the generation of equicorrelated errors in the study effects, with variance determined by the sample size of each study.

First, to generate equally correlated "true" effect sizes, we assume that six studies are done by the same researcher. Consider the Fisher-transformed correlation Z_i, where $EZ_i = E(Z_i \mid \zeta_i) = \zeta_i$, and Z_i is the observed effect size for the *i*th study for an author; Z_i is a Fisher *Z* transformed correlation coefficient (Hedges & Olkin, 1985), and ζ_i is the true effect size for the *i*th study for an author.

In the hypothetical effect-size generation, several effect sizes per author are generated, and we consider each true effect size as a linear function of two variables *X* and Y_i. For example, $\zeta_1 = (aX+bY_1)$, $\zeta_2 = (aX+bY_2)$, ….. , $\zeta_k = (aX+bY_k)$. *X* and Y_i are both distributed as standard normal variables: $X, Y_i \sim N(0,1)$.

So, $E(aX+bY_i) = a*E(X) + b*E(Y_\mathrm{i}) = 0$,

$Var\,(aX+bY_i) = a^2*V(X) + b^2*V(Y_i) = a^2 + b^2 = \sigma^2$.

The covariance of a pair of these effect sizes is

$$Cov\,(aX+bY_{i,}\ aX+bY_\mathrm{j}) = a^2 .$$

The correlation of these effect sizes is

$$Corr\,(aX+bY_i,\ aX+bY_{i'}) = \frac{a^2}{a^2+b^2} = \rho .$$

In the above effect-size generation process, a and b are determined by the formulas,

$a = \sigma\sqrt{\rho}$ and

$b = \sqrt{\sigma^2 - a^2}$.

By specifying the ρ and σ values in the simulation program in given Appendix I, the values of a and b are determined and used to generate the true effect sizes ζ_1 through ζ_k.

Second**,** I generate equicorrelated sampling errors with variance determined by sample size. The observed effect sizes (Z_i) are determined by the true effect size (ζ_i) and error term (e_i). Consider a case with 6 effect sizes. In such a case,

$$\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \end{bmatrix} \sim \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \\ \zeta_5 \\ \zeta_6 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix} .$$

Therefore,

$$\begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \zeta_3 \\ \zeta_4 \\ \zeta_5 \\ \zeta_6 \end{bmatrix} \sim \mathrm{N}_6 \left(\begin{bmatrix} \zeta \\ \zeta \\ \zeta \\ \zeta \\ \zeta \\ \zeta \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix} \right),$$

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix} \sim \mathrm{N}_6 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{(n_i - 3)} \begin{bmatrix} 1 & \rho_2 & \rho_2 & \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & 1 & \rho_2 & \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & 1 & \rho_2 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \rho_2 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & \rho_2 & \rho_2 & 1 \end{bmatrix} \right).$$

The above true effect sizes are generated by specifying the ρ and σ values given in the R program in the Appendix I. ζ_i represents the true effect size for the *i*th study in a set of 'same author' papers, and consists of a common overall effect (ζ), and varies based on random error (σ^2) plus the correlation between studies. In other words, our example shows six studies for one author, and those six studies have overall effect ζ and vary based on variance σ^2 with correlation ρ which I specify. The variance (σ^2) represents the variance in overall effects due to the different designs of experiments, different conditions, samples and instruments for the 'same author' studies.

Next, the error term (e_i) is determined by correlation ρ_2 which I specify in the program in Appendix I, and the variance determined by the sample size ($v_i = (\frac{1}{n_i - 3})$) for the Fisher *Z* transformation.

The correlations of both true effect sizes and error terms show the different degree of similarities of studies by the same author. (In this research, I use the same correlation values for these two correlations to simplify the simulation conditions).

These two correlations represent the relatedness of two errors in meta-analysis: random between-studies error and sampling error. Correlated error represents the similarities that arise when similar samples are selected by one author, whether it is intended or unintended. However, two correlated errors were specified with the same value for my simulation studies, so considering two errors is not the big issue for my simulation study. The correlation among the effect size is the main factor to investigate for my simulation.

The simulation factors in this study are:

a) Study size (*n* = 100, 1000, and 10000),
b) The number of studies for each author (*k* = 2, 3, 4, 5, 10, and 30),
c) The magnitude of effect size (δ = 0.2, 0.5, and 0.8), and
d) The magnitude of correlation within 'same author' studies: (ρ = 0.0, 0.2, 0.5, and 0.8).

As a result, data were generated for 216 simulation conditions (3*6*3*4) in this research. For each simulation condition, the hypothetical syntheses will be replicated 1000 times for each condition.

Examination of proposed statistical methods

For evaluation of each condition of the generated 'same author' studies, I treat the different correlated effect sizes as hypothetical 'same author' studies. These different correlated effect sizes are investigated using the methods proposed in chapter 3. I summarize and make graphs for *H,* the Birge ratio*,* the percent of significant *Q* values*,* the average *Q,* the width of the CI and the between studies variance for each condition based on 1000 replications for each correlation scenario: No correlation and low (0.2), medium (0.5), or high (0.8) correlations among ‘same author’ studies.

I anticipate that the more highly correlated effect sizes are among the ‘same author’ studies, the more similar the outcomes will be. However, this will not occur in the uncorrelated studies. I investigate the results of these generated data sets to see if the analysis results are consistent with the results of the empirical analyses in chapter 5.

Results

After generating hypothetical effect sizes for each condition, the proposed indicators were calculated. The magnitudes of correlation and number of studies within 'same author' studies are investigated as main factors for this 'same author' issue simulation. Table 3 presents the characteristics of data generation for the 'same author' issue with study-size $n_i = 100$.

Table 3: Characteristics of 'same author' data generation (n = 100)

ID[A,B]	Number of studies per author	Sample size (n_i)	Magnitude of ζ	Magnitude of correlation
20	2	100	0.2	0.0
30	3	100	0.2	0.0
40	4	100	0.2	0.0
50	5	100	0.2	0.0
100	10	100	0.2	0.0
300	30	100	0.2	0.0
22	2	100	0.2	0.2
32	3	100	0.2	0.2
42	4	100	0.2	0.2
52	5	100	0.2	0.2
102	10	100	0.2	0.2
302	30	100	0.2	0.2
25	2	100	0.2	0.5
35	3	100	0.2	0.5
45	4	100	0.2	0.5
55	5	100	0.2	0.5
105	10	100	0.2	0.5
305	30	100	0.2	0.5
28	2	100	0.2	0.8
38	3	100	0.2	0.8
48	4	100	0.2	0.8
58	5	100	0.2	0.8
108	10	100	0.2	0.8
308	30	100	0.2	0.8

A: 2, 3, 4, 5, 10, and 30 represent the number of studies per author,
B: 0, 2, 5, and 8 of the last digit of the ID represent the correlation values.

Next I discuss results of the simulation for same-author issue. All figures follow the descriptions of these results.

H

Figure 7 shows the *H* values calculated for the 'same author' studies. The number of studies and different levels of inter-correlation are examined. Figure 7 shows that *H* increases as the degree of correlation increases, as expected. Figure 7 shows that *H* increases as the number of studies increases except when the correlation is equal to zero. This indicates when correlation is zero (or, when no similarities exist in the 'same author' studies), *H* decreases as the number of studies increases. This is very similar to the 'same data' simulation results in Figure 1. When the study-size ratio is small in the 'same data' situation or when 'same author' studies are uncorrelated, no similarities exist in both Figures 1 and 7.

Functions of *Q*

1) Birge ratio

Figure 8 shows that the Birge ratio is constant as the number of studies increases, Figure 8 also shows that the Birge ratio decreases as the level of correlation increases. Figure 8 shows a pattern similar to that for Figure 12 showing the between-studies variance.

2) Percent of significant *Q* values

Figure 9 shows that the percent of *Q* values significant at the .05 level decreases as the number of studies increases. Figure 9 also shows the highly correlated studies have only a small percent of *Q* values significant at the .05 level. Figure 9 shows a pattern similar to that in Figure 3 of the 'same data' study. The 2% study-size ratio effect in Figure 3 is similar to the zero correlations case in Figure 9, and the 50% study-size ratio effect is similar to the results for the .5 correlation in Figure 9.

3) Average *Q* values

Figure 10 shows that Q increases as the number of studies increases, according to statistical theory. However, Figure 10 also shows that Q decreases as the correlation increases.

The real issue here is whether the mean Q is lower than expected. Under H_0, means in Figure 10 should be equal to the *df*. So for $k = 10$, $df = 9$, for $k = 5$, $df = 4$ etc. The only case where the means are really where they should be seems to be when $\rho = 0$ in Figure 10. In all other cases they are lower than suggested by the null hypothesis distribution for independent effects.

Width of confidence interval

Figure 11 shows the effect of number of studies and correlation (ρ value) on fixed-effects and random-effects confidence interval widths. Widths of both the fixed-effects and random-effects confidence intervals decreases as the correlation and number of studies increase. When the correlation is .8, both confidence intervals are exactly the same in Figure 11.

Between studies variance in effect sizes

Figure 12 shows that the between studies variance was almost consistent across the different numbers of studies, but the between studies variance decreases when correlation increases. This is similar to Figure 6 in the 'same data' situation.

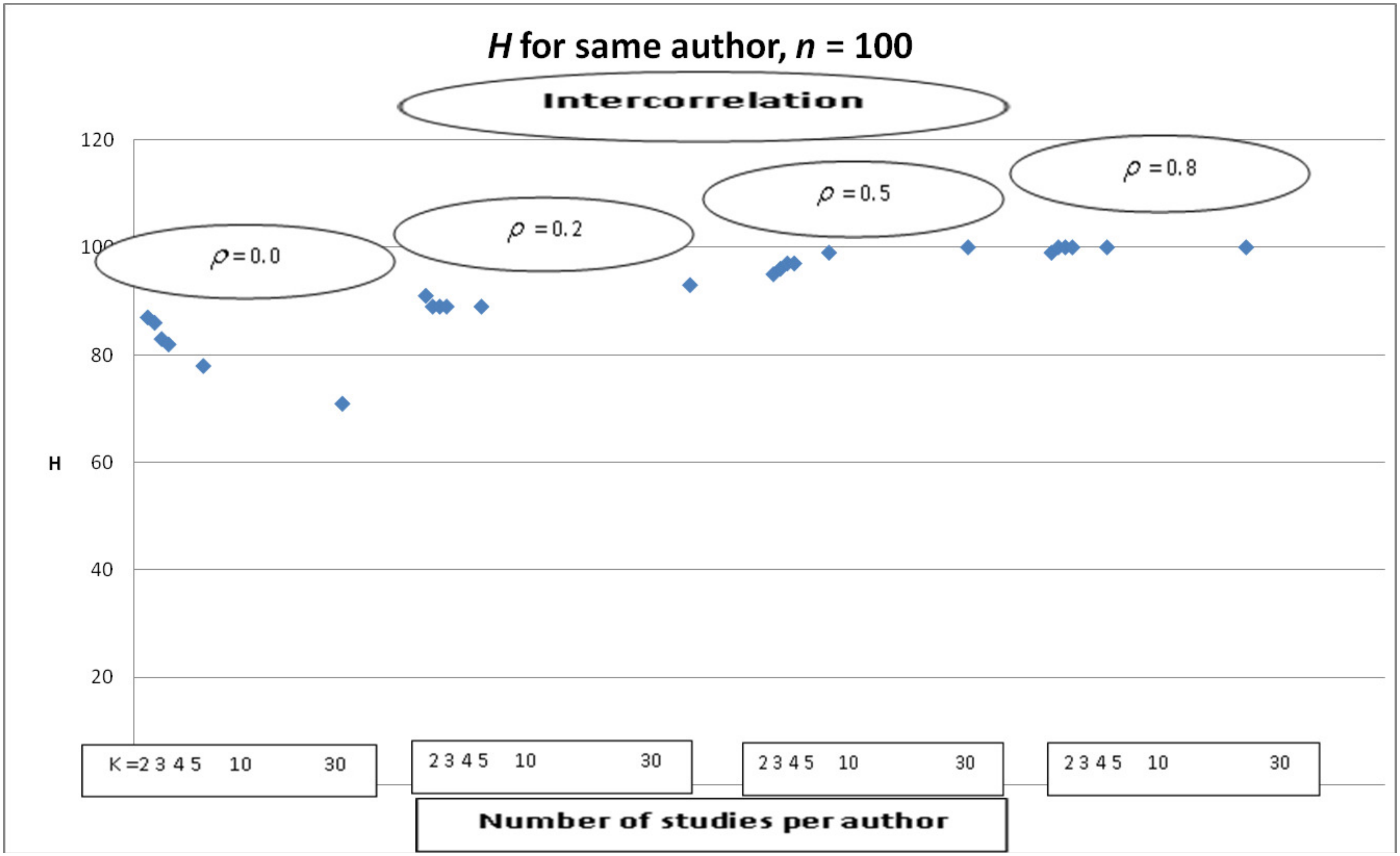


Figure 7: *H* for same author, *n* = 100.

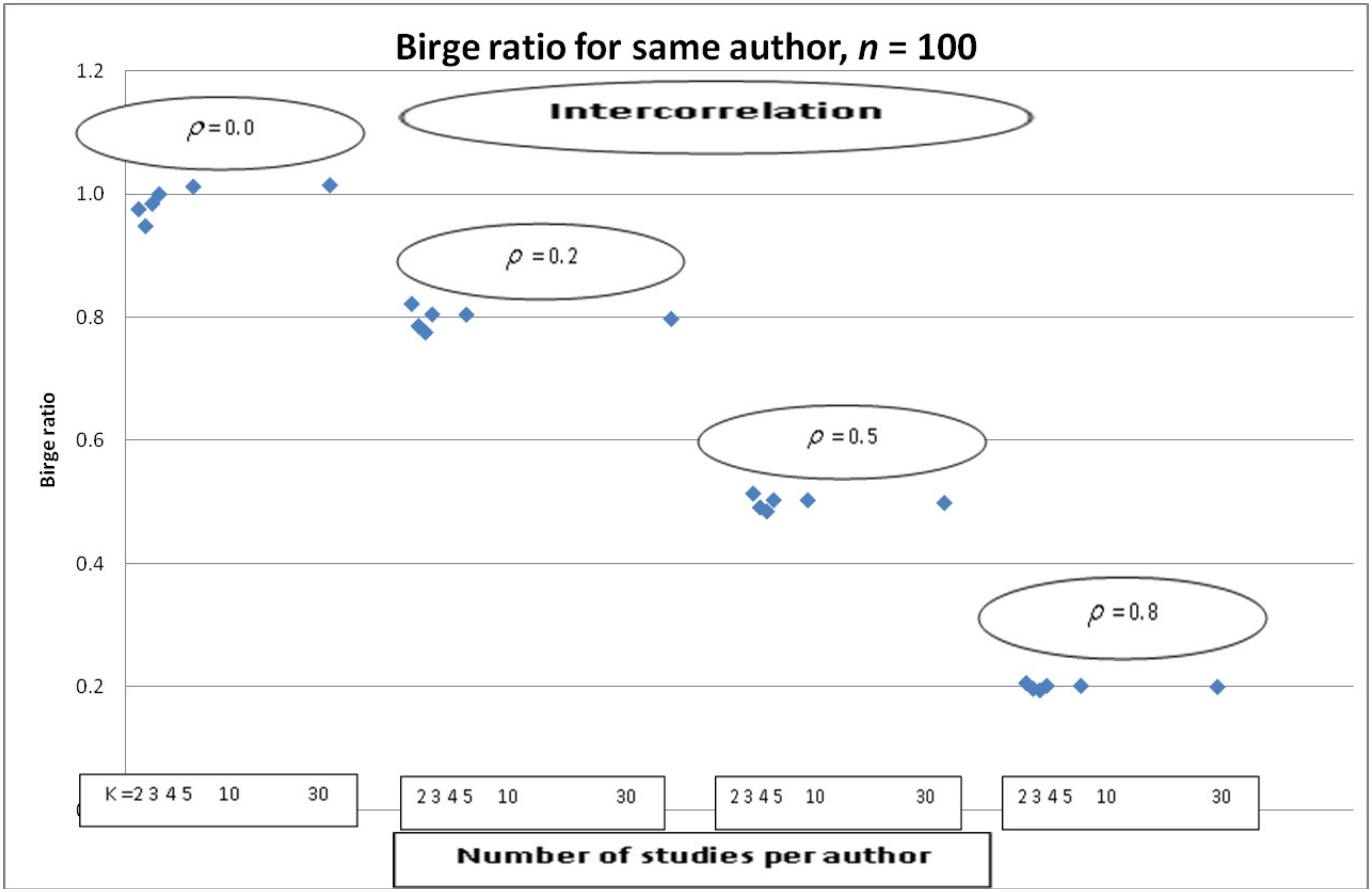


Figure 8: Birge ratio for same author, *n* = 100.

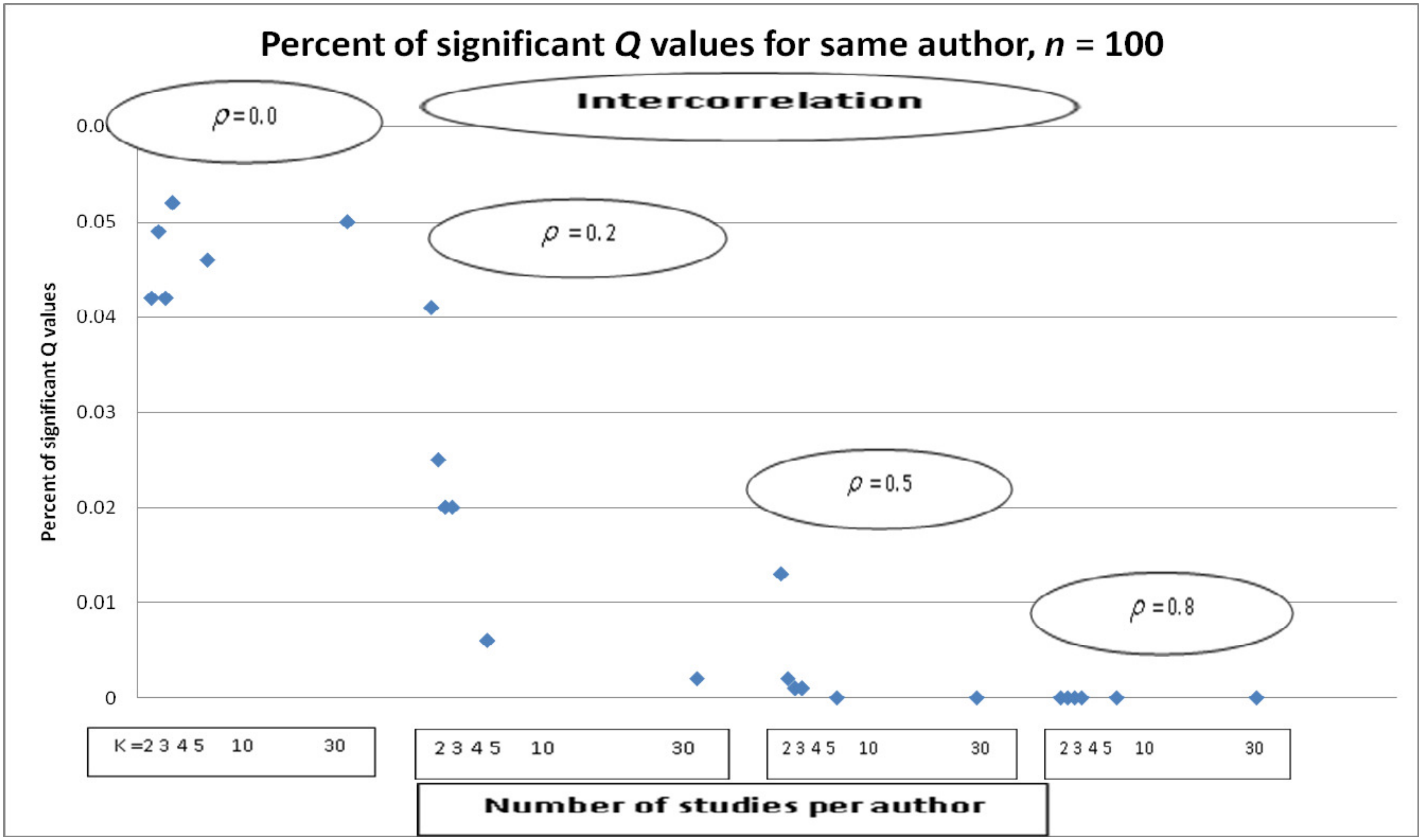


Figure 9: Percent of significant *Q* values for same author, *n* = 100.
* In the Y-axis, 0.01 to 0.07 represent 1% to 7%.

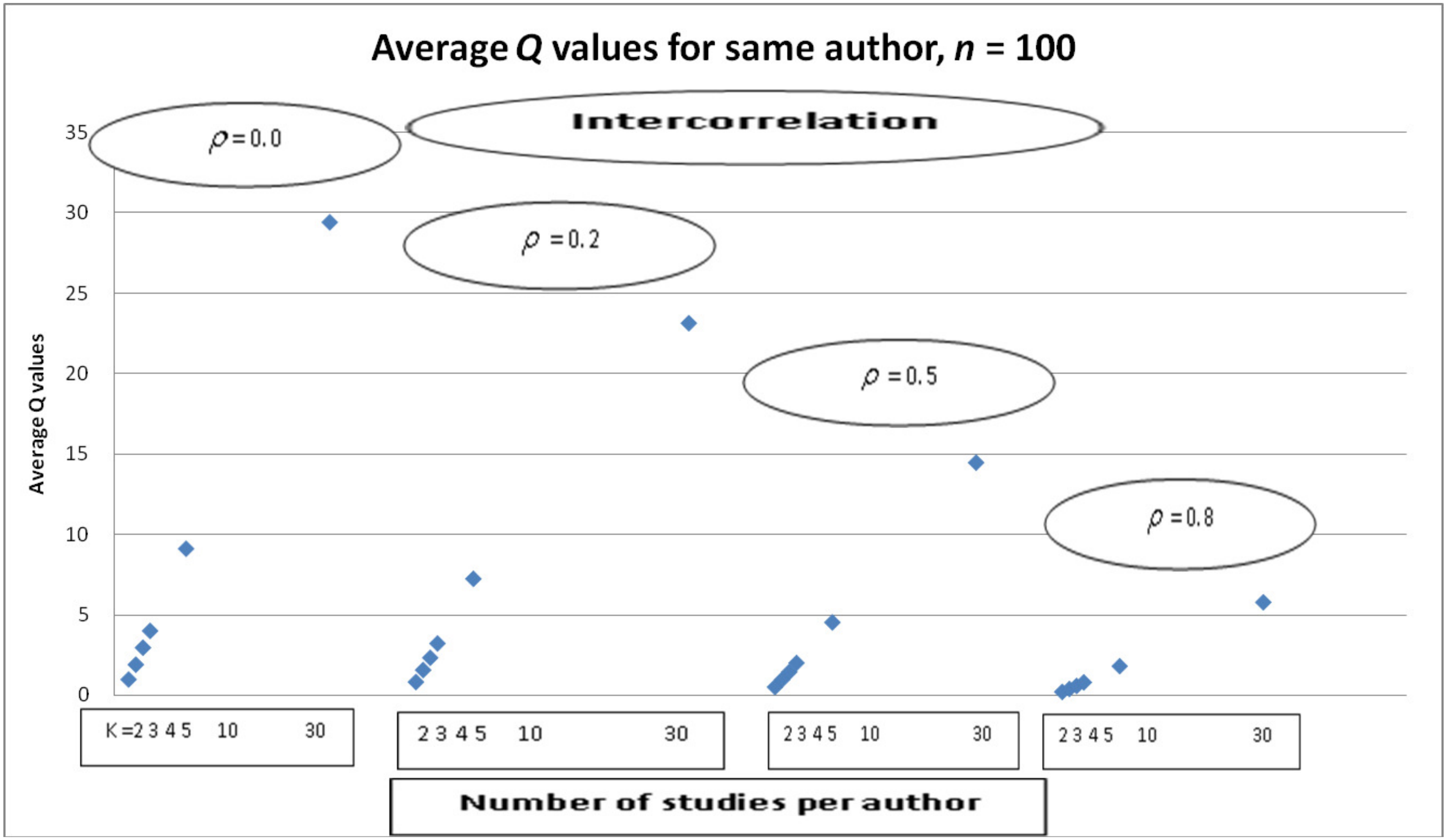


Figure 10: Average *Q* for same author, *n* = 100.

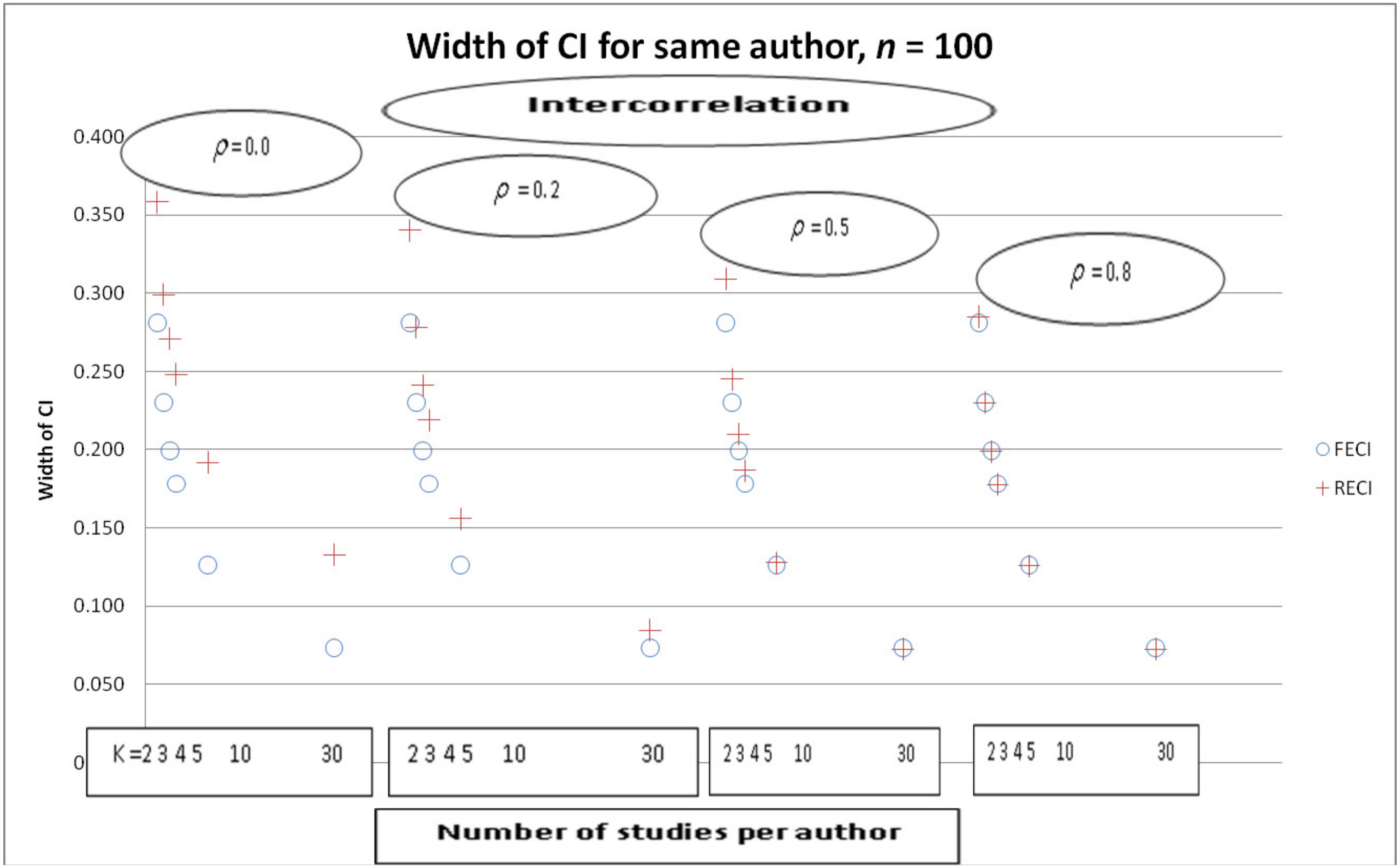


Figure 11: The width of the CI for same author, *n* = 100.

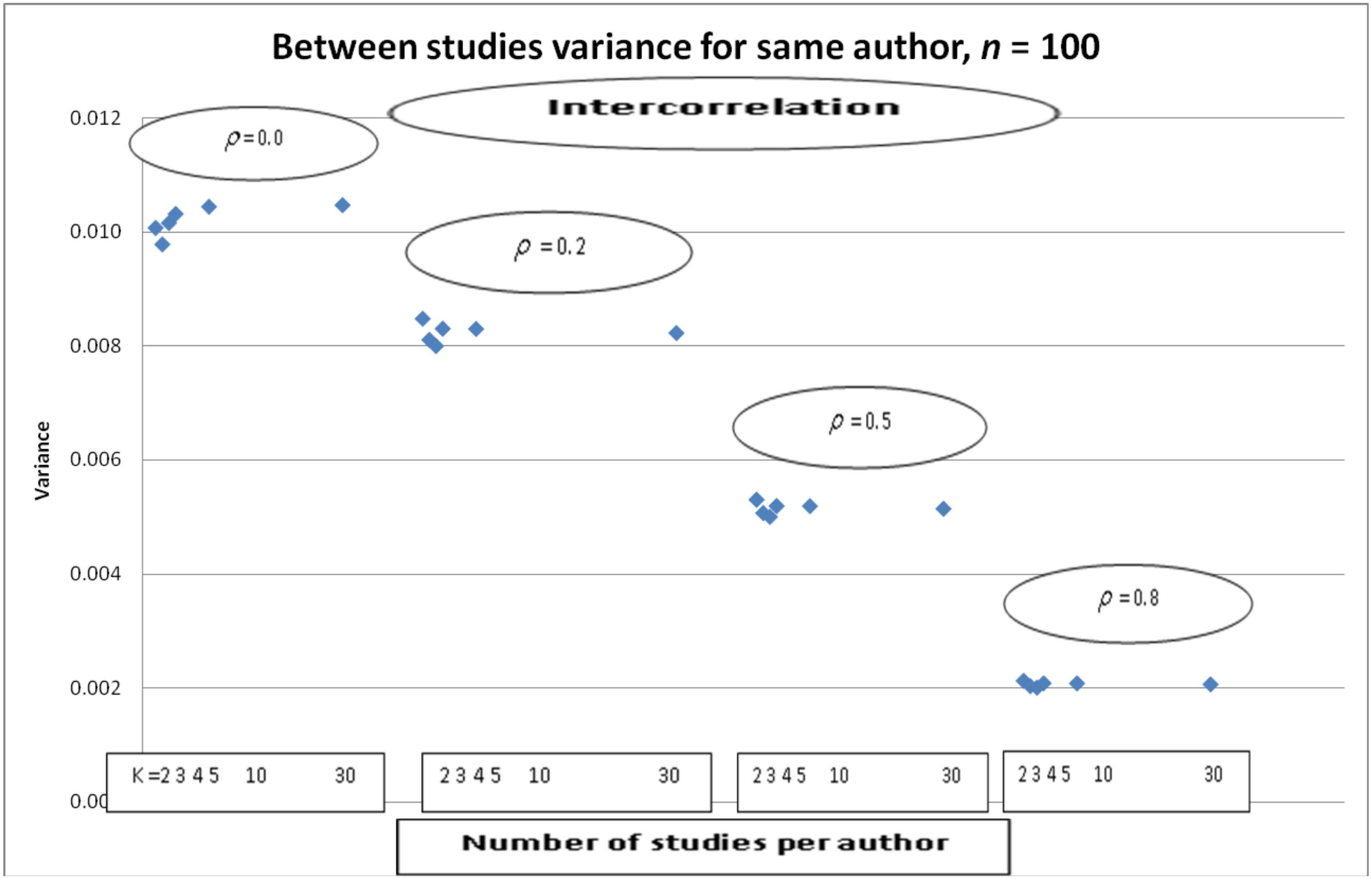


Figure 12: Between studies variance for same author, *n* = 100.

Summary

All indices (*H*, Birge ratio, percent of significant *Q* values, the width of the CI, and between studies variance) show the similarities of ‘same author’ studies. Highly correlated studies show more similarity. The average *Q* is dependent on the number of studies, but its mean is considerably lower than its theoretical expected value when $\rho > 0$.

H and the percent of significant *Q* values show no dependence for the small study-size ratio (2%) and for uncorrelated ‘same author’ studies because no similarities exist among ‘same data’ studies and ‘same author’ studies in these conditions.

The other results for sample sizes 1,000 and 10,000 for the ‘same author’ simulation are attached in Appendix K. The results show the same pattern as found for studies with sample size of 100.

Uncorrelated studies show a pattern similar to that found for a study-size ratio of 2%, and medium and highly correlated studies (with .5 and .8 correlations) show patterns similar to those for large study-size ratios (such as 0.4 and 0.5). This makes sense because the correlation and study-size ratio both represent similarities among ‘same author’ and ‘same data’ studies.

Limitation

This simulation investigated the effect of degree of overlap or magnitude of inter-correlation, number of studies by one author, and study-size. The magnitude of correlation does not directly represent the ‘same author’ condition, but it is used to approximate the effect of ‘same author’ similarities.

CHAPTER V

EMPIRICAL ANALYSIS

USING TWO EXAMPLE META-ANALYSES

I have proposed several approaches to exploring dependence in meta-analysis: homogeneity tests, fixed-effects categorical analysis, intraclass correlations, sensitivity analysis and HLM analysis. In chapter IV, I generated data and investigated various conditions to check whether the proposed methods worked or not.

In this section, I report on two empirical analyses with real meta-analysis data to examine the proposed approaches. The first one is a class-size and student achievement meta-analysis (Shin, 2008), and the second uses ESL (English as a Second Language) meta-analysis data (Ingrisone & Ingrisone, 2007). The first example examines the "same data" issue, and the second one explores the "same author" issue.

Class-size meta-analysis as an example of the ‘same data’ issue

The relationship between class-size and student achievement has been studied for a long time. However, the findings are still inconsistent among primary studies. In this analysis, I investigated studies from the 1990s to the present, because we have previous meta-analyses for this topic: Glass and Smith’s class-size and student achievement meta-analysis (1979) covers very early studies, the McGiverin, Gilman, and Tillitsk (1989) review is not actually a meta-analysis study, and Goldstein, Yang, Omar, Turner, and Thompson’s (2000) meta-analysis focused on methodological issues.

In the class-size studies, there are several large scale experiments including STAR (Student/Teacher Achievement Ratio Study in Tennessee 1985 - 1989). When I gathered the literature for this meta-analysis of class-size and student achievement, I found many studies using STAR data. The data examined in this empirical analysis are from a meta-analysis of class-size and student achievement by Shin (2008). The main goal of this analysis is to show the dependence of STAR data papers compared to other papers using different data sets. To show the dependence, I examined the homogeneity test, the fixed-effects categorical analysis, intraclass correlation, sensitivity analysis and HLM analysis. I also investigated similarities between meta-analysis using HLM and the conventional meta-analysis approach. I begin by introducing the results of the fixed-effects model and the random-effects model.

1. Fixed-effects and random-effects models

In the fixed-effects model, the effect size is 0.15, and standard error of effect size is 0.0036. The homogeneity test statistic *Q* is significant (*p* value < .05), so we can reject the null hypothesis, and it means the effect sizes are heterogeneous.

Under the fixed-effects model, the homogeneity test is significant. The random model is appropriate for this analysis. Sometimes, even though the homogeneity test is not significant, researchers use the random-effects model because the homogeneity test has less power for small numbers of studies in meta-analysis (Higgins et al., 2003, p. 557).

Table 4: Comparison of fixed-effects model and random-effects model results

	Mean	Standard Error
Fixed-effects model	0.15	0.004
Random-effects model	0.18	0.017

* SE: Standard Error

The random-effects mean is slightly larger (0.18 vs. 0.15) and the random-effects standard error (SE) is also larger (0.017 vs. 0.004). The random-effects confidence interval is much wider (a width of 0.066 vs. 0.014). This mean effect size from the random model (0.18) is similar to that in the HLM unconditional model output (0.18) . To examine the sources of variability, a fixed-effects categorical analysis is next conducted.

2. Fixed-effects categorical analysis

In this fixed-effects categorical analysis, the main interest is to investigate the difference between papers using STAR data and papers using different data. In the same-author studies and same-data studies, I used dummy coding to distinguish same-data studies and different-data studies. For example, if the papers use the same data (here, the STAR data), the coding is 1, if not, it will be 0. For the same-data issue, the researcher needs to consider the difference of two Q_{within} statistics: STAR vs. Others. STAR studies have a larger mean effect size (0.22) than the other studies (-0.07). This research focuses on the difference in the homogeneity in these two groups. When Q_{Between} is significant as it is here (Q_{Between} = 1118.2, df = 1, p < 0.001), it means the two groups are different on average, and if Q_{within} is also significant as it is here (Q_{within} =872.9, p < 0.001) it means that there is still variability within studies. Q_{within} is significant if one group has variability within studies, such as the “other” group in this example. I expected that the other studies would show more variability than STAR studies, and I show the results using quantification of the homogeneity by H and I^2 in Tables 6-8.

Table 5: Fixed-effects categorical model for class STAR

STAR	*k*	Q	*p*-value	LL	Effect size	UL	Variance	Birge Ratio
STAR	78	300.82	<.0001	0.208	0.217	0.225	0.000018	3.9
Others	31	572.07	<.0001	-0.079	-0.065	-0.051	0.000053	18.5

*LL: lower bound of confidence interval for effect size, UL: upper bound of confidence interval for effect size

3. Comparison of homogeneity by STAR variable

The goal of this research is to show the dependence in the studies using the same data, so I examined the homogeneity of the two groups of studies using the *H* and I^2 indices proposed by Higgins and Thompson (2002). *H* can be defined as follows:
H = (CI width for the fixed-effects model)/(CI width for the random-effects model) * 100.

Table 6: *H* of all class-size studies

Confidence Interval	LL	UL	Width	Ratio (H)
Fixed model	0.139	0.153	0.014	
Random model	0.142	0.208	0.066	22%

*LL: lower bound of confidence interval, UL: upper bound of confidence interval

In Table 6, the confidence interval of the fixed model is 22% of the size of the confidence interval from the random model. This indicates much variability exists between studies. I next compared the *H* index in the two groups separately: STAR studies vs. Others in Table 7.

Table 7: *H* of STAR studies vs. other studies

STAR	LL	UL	Width	Percentage
Fixed model	0.208	0.225	0.017	
Random model	0.210	0.245	0.034	48%
Other Studies	LL	UL	Width	Percentage
Fixed model	-0.079	-0.051	0.029	
Random model	-0.107	0.036	0.143	20%

*LL: lower bound of confidence interval, UL: upper bound of confidence interval

The *H* index represents a 20% confidence interval width overlap between the fixed-effects model and the random-effects model in the other group. The *H* index shows 48% confidence interval overlap between the fixed-effects model and the random-effects model in the STAR group. The STAR group's confidence intervals overlap more than those of the Other group. The STAR group is more homogenous than the Other group.

Next, I investigated the I^2. I^2 is calculated based on the following formula, I^2 = 100 %*($Q - df$)/Q.

Table 8: I^2 for STAR effect

Class	k	Q	I^2
STAR	78	300.82	74.0%
Others	31	572.07	94.6%

To quantify the homogeneity for these two groups: STAR vs. Other, I calculated I^2. The Other group I^2 is 94.6 %, while the STAR group I^2 is 74%. This means STAR studies are more homogeneous than Other studies, while the STAR group also has much variability.

4. Graphical approach

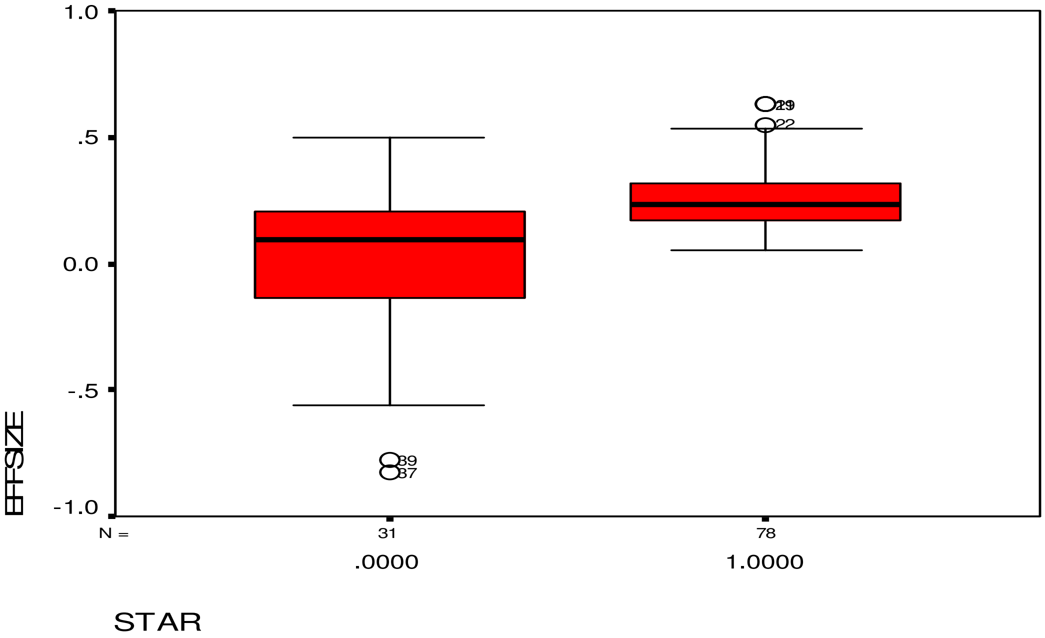


Figure 13: Fixed-effects categorical analysis box plot comparison between STAR studies and other studies.

Figure 13 indicates that variation in the effects from STAR studies is lower as shown by narrower box and whiskers compared to the other group. The ends of box represent the first and third quartiles of the distribution, and the horizontal line within the box represents the median point**.** The whiskers show the smallest and largest values of the distribution, however, if there are outliers, SPSS does not always use the maximum and minimum value. The graphical approach shows the same result as the homogeneity test, *H,* and I^2.

In this graphical approach, meta-analysts can indirectly assess and show the similarities or relatedness of same-data studies. In the fixed-effects categorical analysis, meta-analysts can examine whether same-data studies are less variable than different data studies.

5. ICC

In the class-size example (Shin, 2008), the total variance is 0.0479, and the sampling variance is 0.00997. The variance component in the unconditional model is 0.038. The total variance (0.0479) is made of the sum of sampling variance (0.00997) and systematic variance (0.038). To calculate the ICC for the class-size example, I divide systematic variance by the sum of sampling variance and systematic variance as shown in the formula below:

$$\hat{\rho} = \hat{\tau}_{00} / (\hat{\sigma}^2 + \hat{\tau}_{00})$$

$$= 0.038/0.0479 = 0.793.$$

The parameter variance (τ_{00}) is estimated as 0.038, and the proportion of systematic variance is 0.793. This means that 79.3% of the variance in the effect sizes is from differences between studies. The ICC is next examined for the two groups: STAR studies and Other studies. First, Other studies were examined. The total variance is 0.103, and sampling variance is 0.022. Thus, the systematic variance is 0.081. The ICC is 78.6% (0.081/0.103). Second, STAR studies were examined. The total variance is 0.0149, and sampling variance is 0.0052. The systematic variance is 0.0097, and the ICC is 65% (0.0097/0.0149). In Other studies, 79% of the variance in effect sizes is from variance between studies, however, in the STAR studies, only 65% of the variance in effect sizes is between studies. In conclusion, the between studies variance of the STAR studies is smaller than that of Other studies. This is similar to the findings of the homogeneity test and graphical approach.

6. HLM

Using the HLM analysis for meta-analysis, we can get the intraclass correlation and 'variance explained' information. Intraclass correlations in the unconditional model indicate how much variance in effect sizes lies between studies, and the 'variance explained' indicates the percentage of variance accounted for by study characteristics like "same author" and "same data" factors.

The unconditional model produces estimates of the grand mean, γ_0, and Level-2 variance, τ. The estimated grand-mean effect size is small, $\hat{\gamma}_0$= 0.18. It means that, on average, small-class students score about 0.18 standard deviation units above the large-class students. However, the estimated variance of the effect parameters is $\hat{\tau}$ = .038. This corresponds to a standard deviation of .62, which implies that there is important variability in the true effect sizes. This also is the same value obtained as part of the computation of the ICC.

Table 9: Conditional model for the meta-analysis of class-size on student achievement

Fixed-effects	Coefficient	Standard Error	*t* Ratio	
Intercept, γ_0	-0.037	0.03	-1.1	
STAR, γ_1	0.276	0.04	7.3	
Random-effects	Variance Component	*df*	χ^2	*p*-value
True effect size, δ_j	0.023	107	913.42	<0.000

The Table 9 results show that the class-size effects of STAR studies are larger. In comparison with other studies class-size effect on student achievement, STAR studies students achievement is larger by 0.276 effect size (i.e., $\hat{\gamma}_1$ = 0.276, *t* = 7.3). The HLM analysis has two goals: the first one is to assess the variability in the true effect parameters, and a second is to account for that variation. In the class-size example, the unconditional model assesses the variability in the true effect parameters, and the conditional model accounts for that variation. As we can see in Table 10, the STAR variable explained variability in this example using HLM analysis.

Table 10: Proportion of variance explained by the STAR dummy variable

Class	HLM	Unconditional model variance	Conditional model variance	Explained variance
		0.038	0.023	39%

Proportion of variance explained $= \dfrac{\hat{\tau}_{00}(unconditional) - \hat{\tau}_{00}(conditional)}{\hat{\tau}_{00}(unconditional)} =$

$$\frac{0.038 - 0.023}{0.038} = 39\%$$

When we compare the two models: the unconditional model and the conditional model, 39% of the total variance is explained by the STAR variable. HLM analysis efficiently performs two goals simultaneously, and gives an estimate of variability explained in the conditional model.

7. Sensitivity analysis

This sensitivity analysis focuses on the effect size of all studies vs. the effect size of all studies except the same-data studies, while the categorical analysis pays attention to the effect size of same-data studies vs. that of different data studies. However, there are many similarities in these analyses.

Based on the sensitivity analysis results, reviewers need to investigate in depth why the results differ between the effect sizes of all studies vs. the effect size of all studies except the same-data papers because the result of sensitivity analysis depend on the topic, research area, and the relationship between 'same data' studies and studies using different data sets. This needs especially when estimating overall effect size in meta-analysis.

Table 11: Sensitivity analysis for class-size studies

	Mean	Standard Error
Random-effects model of all studies	0.18	0.017
Random-effects model without STAR studies	-0.04	0.036

STAR is the most well controlled experiment among the class-size studies, and many researchers have studied the STAR data to investigate the relationship between class-size and student achievement. The effect size without STAR studies is -0.08, suggesting that smaller class student achievement is lower by -0.08 standard deviations compared to large-class student achievement. This suggests that there is no reason to reduce class size as a way to increase student achievement. However, the overall class size studies' effect size is 0.15, indicating that the effect of small class-size is positive, and represents a 0.15 standard-deviation effect for student achievement, compared to large class-size.

In this sensitivity analysis approach, the reviewer should be very cautious to generalize meta-analysis findings. Without STAR studies, a reviewer can say that the effect size of small class size on student achievement is negative. With STAR studies, meta-analysts will say the overall effect size of small class size on student achievement have a small but positive and 0.15 standard deviation unit effect on student achievement.

A reviewer should report these two results together without losing information. How can we estimate the overall effect of class-size on student achievement considering several STAR studies simultaneously? We cannot say something without considering the characteristics of data sets, the quality of papers, the in-depth study of same-data studies, and the experimental conditions of same-data studies. For this STAR example, I will include all of the studies for overall estimation of class-size effect on student achievement because STAR is the most well controlled state-wide experiment in the history of class-size studies. Many well-known authors studied this issue, the quality of papers is better than that of other studies, and every STAR study's focus is a little bit different as I indicated in the case study report in chapter 2.

In conclusion, whenever reviewers determine an overall effect that includes the same-data studies, reviewers should report all possible results using categorical analysis and sensitivity analysis, and meta-analysts should estimate overall effect size considering the characteristics of 'same data' sets and the quality of papers using 'same data'.

ESL as an example of the “same author” issue

This example examined the effectiveness of ESL instructional methods. I used this study as a “same author” case because one author, Kubota, has several studies in this meta-analysis. This example data is from Ingrisone and Ingrisone (2007). Ingrisone and Ingrisone (2007) examined 17 studies including 23 effect sizes. One author, Kubota, contributed 4 studies and 6 effect sizes among the 17 studies and 23 effect sizes. The goal of this analysis is to show the similarities of papers by Kubota compared to other papers by different authors. I also examined the homogeneity test, the fixed-effects categorical analysis, ICC, sensitivity analysis, and HLM analysis. I show the relationship between meta-analysis using HLM and conventional meta-analysis.

1. Fixed-effects and random-effects model

In the fixed-effects model, the effect size is 0.67, and the standard error of effect size is 0.07. The homogeneity test statistic, *Q* is not significant (*p* = .167), so we can not reject the null hypothesis. It means the ESL studies are homogeneous. We can use the fixed-effects model for these data. However, I investigated the random-effects model, too.

Table 12: Comparison of fixed-effects model and random-effects model results

	Mean	Standard Error (SE)
Fixed-effects model	0.67	0.07
Random-effects model	0.66	0.08

The random-effects mean is slightly smaller (0.66 vs. 0.67) and the random-effects SE is slightly larger (0.08 vs. 0.07). The random-effects confidence interval is much wider (a width of 0.30 vs. 0.26). As expected, the mean effect size from the random model (0.66) is exactly the same to that from the HLM unconditional analysis (0.66).

2. Fixed-effects categorical analysis

For the same-author studies example, I used dummy coding to distinguish same-author studies from different author-studies. Specifically, if the papers are from author Kubota, the coding is 1, if not, it will be 0. Kubota's papers look more homogeneous than the other papers based on the Q_{within} statistic. Other studies show a large mean effect size (0.80) compared to that of Kubota studies (0.47). The focus of this study is the homogeneity test of difference between these two groups. Q_{Between} is significant (Q_{Between} = 5,8, df = 1, p = 0.015), and means that the two groups are different on average. If Q_{within} is not significant as it is here (Q_{within} = 22.4, p = 0.376), it means that there is no variability within the sets of studies. I expected that the other studies would show more variability, and I show the results using quantification of the homogeneity by H and I^2 in Tables 14-16.

Table 13: Fixed-effects categorical model for ESL Kubota

Kubota	*k*	*Q*	*p*-value	LL	Effect Size	UL	Birge Ratio
0	17	21.28	0.17	0.63	0.80	0.97	1.33
1	6	1.12	0.95	0.27	0.47	0.68	0.22

*LL: lower bound of confidence interval for effect size, UL: upper bound of confidence interval for effect size

3. Comparison of homogeneity by Kubota variable

The goal of this research is to explore the dependence in the studies by Kubota versus others, so I examined the homogeneity of those two groups using the *H* and I^2 statistics that are defined above (Higgins & Thompson, 2002).

Table 14: *H* of all ESL studies

Confidence Interval	LL	UL	Width	Ratio (*H*)
Fixed model	0.536	0.798	0.262	
Random model	0.515	0.813	0.298	88%

*LL: lower bound of confidence interval, UL: upper bound of confidence interval

In Table 14, the confidence interval width in the fixed-effects model is 88% of the size of the confidence interval width of the random-effects model. This indicates little variability between studies in this meta-analysis. However, I will still compare the *H* index in the two groups (Kubota vs. others) in the ESL meta-analysis.

Table 15: *H* of studies by Kubota vs. other studies

Kubota	LL	UL	Width	Percentage
Fixed model	0.267	0.677	0.410	100%
Random model	0.267	0.677	0.410	
Other	LL	UL	Width	Percentage
Fixed model	0.631	0.972	0.341	
Random model	0.575	0.966	0.391	87%

*LL: lower bound of confidence interval, UL: upper bound of confidence interval

The *H* index shows 87% confidence interval width overlap between the fixed-effects model and the random-effects model in the other group. The *H* index shows 100% of confidence interval width overlap between the fixed-effects model and the random-effects model in the Kubota group. The Kubota studies' confidence intervals overlap more than those of other studies. It means Kubota studies are more homogenous than the other studies.

Next, I investigated the I^2, calculated as above, with $I^2 = 100\ \%*(Q - df)/Q$.

Table 16: I^2 for Kubota effect

ELL	*k*	*Q*	I^2
Kubota	6	1.12	0%
Other	17	21.28	20.1%

The I^2 for other studies is 20.1% which means small variability, while the I^2 for the Kubota studies I^2 is -435.7% which is truncated to 0. It means Kubota studies are much more homogeneous than the other studies.

4. Graphical approach

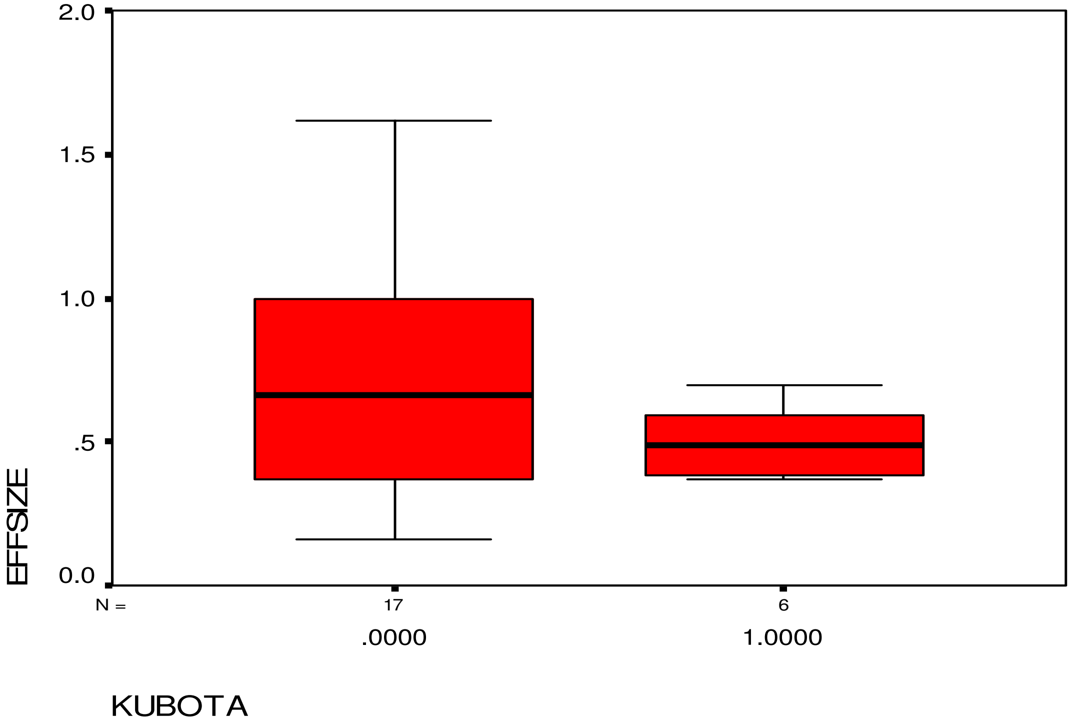


Figure 14: Fixed-effects categorical analysis box plot comparison between Kubota and other studies

Figure 14 indicates that variation in Kubota studies is lower as shown by the narrower box and whisker plot compared to the other studies. The width of the box represents the first and third quartiles in distribution, and the horizontal line within the box represents the median point. The whiskers show the smallest and largest values of distribution, however, if there are outliers, SPSS does not always use the maximum and minimum value. The graphical approach shows the same result as the homogeneity test, H, and I^2.

5. ICC

In the ESL instruction method example (Ingrisone & Ingrisone, 2007), the total variance is 0.183, and the sampling variance is 0.14. The variance component in the unconditional model is 0.043. The total variance (0.183) is from the sum of the sampling variance (0.14) and systematic variance (0.043). To calculate the ICC for the ESL example, I used the formula below:

$$\hat{\rho} = \hat{\tau}_{00} / (\hat{\sigma}^2 + \hat{\tau}_{00})$$

$$= 0.043/0.183 = 0.24.$$

The parameter variance (τ_{00}) is estimated as 0.043, and the proportion of systematic variance is 0.24. This means that 24% of the variance in effect sizes is between studies variance.

The ICC also is examined for each of the two groups: Kubota studies vs. other studies. First, other studies are examined. The total variance is 0.16, and sampling variance is 0.16. This means that the systematic variance is 0.00, and the ICC is also 0.00 (0.00/0.16). Second, the Kubota studies are examined. The total variance is 0.016, and sampling variance is 0.074, thus the systematic variance is estimated as -0.058. It will be truncated to 0.00. ICC is again 0.00 (0.00/0.016). In this ELL example, the total variance is very small; the systematic variance is also very small and essentially can not be examined. In conclusion, the between studies variance in ELL studies is almost zero. These results are similar to those of the homogeneity test and graphical approach.

6. HLM

Using the HLM analysis for meta-analysis, we can get the intraclass correlation and ‘variance explained’ information. Intraclass correlations in the unconditional model indicate how much variance in effect size lies between studies, and the ‘variance explained’ indicates the percentage of variance accounted for by study characteristics such as “same author” and “same data” factors.

The unconditional model produces the estimates of the grand mean, γ_0 , and Level-2 variance, τ . The estimated grand-mean effect size is medium, $\hat{\gamma}_0$ = 0.66. On average, ESL program students score about 0.66 standard deviation units above the control group students. However, the estimated variance of the effect parameters is $\hat{\tau}$ = .043. This corresponds to a standard deviation of .21, which implies that some variability exists in the true effect sizes. This also is the same value obtained as part of the computation of the ICC.

Based on the results of the unconditional model, we do not need to consider the conditional model because studies do not significantly vary in their effects. However, the conditional model will be investigated to examine the Kubota effect.

Table 17: Conditional model for the meta-analysis of ESL

Fixed-effects	Coefficient	Standard Error	*t* Ratio	
Intercept, γ_0	0.78	0.09	8.4	
Kubota, γ_1	-0.31	0.15	-2.0	
Random-effects	Variance Component	*df*	χ^2	*p*-value
True effect size, δ_j	0.015	21	22.45	0.374

Table 17 shows that Kubota’s studies have a negative effect on the ESL effect size for student achievement. In comparison with other author’s studies, Kubota studies effect sizes were smaller by 0.31 standard deviations (i.e., $\hat{\gamma}_1$ = -0.31, *t* = -2.0).

As we can see in the output in Table 18, the Kubota variable explains a lot variability in this example using HLM analysis.

Table 18: Proportion of variance explained by the Kubota dummy variable

Class	HLM	Unconditional model variance	Conditional model variance	explained variance
		0.043	0.015	69%

Proportion of variance explained $= \dfrac{\hat{\tau}_{00}(unconditional) - \hat{\tau}_{00}(conditional)}{\hat{\tau}_{00}(unconditional)} =$

$$\frac{0.043 - 0.013}{0.043} = 69\%$$

When we compare the two models (the unconditional model and conditional model), 69% of variance is explained by the Kubota variable. Also, even though the variability is not significant in unconditional model, the variance component is much lower in the conditional model.

7. Sensitivity analysis

This sensitivity analysis focuses on the effect size of all studies vs. the effect size of all studies except the same-author studies, while the categorical analysis pays attention to the effect size of same-author studies vs. different author studies. However, there are many similarities in these analyses. Based on the sensitivity analysis results, reviewers need to investigate why the results differ between the effect sizes of all studies vs. the effect size of all studies except same-author papers, considering the characteristics of sample, measurement instrument, and experimental conditions using by the same author such as Kubota, compared to different papers written by others.

Table 19: Sensitivity analysis for ESL studies

	Mean	Standard Error
Random-effects model of all studies	0.66	0.08
All studies with one average effect size of Kubota studies.	0.75	0.09
Random-effects model without Kubota studies	0.77	0.10

In ESL studies, the effect size of all studies is 0.66, suggesting that students received ESL instruction score higher by 0.66 standard deviations, compared to control group students. However, the effect size mean without Kubota studies is 0.80, which means the effect of Kubota studies is lower than that of other studies.

Compared to the class-size example, this ESL case show results that are very homogenous and similar to each other between ‘same author’ papers and papers written by other authors. Even though these papers are similar to each other, reviewers need to examine the effect of ‘same author’ papers when estimating overall effect sizes. Without Kubota’s studies, a reviewer can say that the effect of ESL program on student achievement is positive and roughly 0.80 standard deviation units, compared to control group students. Including Kubota studies, meta-analysts will say the overall effect of ESL on student achievement is 0.67 standard deviation units.

For this ESL case, Kubota studies are very similar and homogeneous each other, and the shifting unit of analysis can be used for Kubota studies. Just one study can be used for overall effect size estimation because the same author studies are very similar to each other. The results can be compared with the above result and reported all together. This recommendation can be applied for the same author papers even though the samples are technically independent.

For generalizable findings in meta-analysis, a reviewer should report these two results together using sensitivity analysis and categorical analysis, and will not lose information. When reviewers determine the overall effect including same-author papers, reviewers should consider differences in the quality of papers, the characteristics of samples and characteristics of experiments by the same author, compared to papers written by other authors.

Table 20: Shifting unit of analysis: Author as an analysis unit

Unit of analysis	Mean	Standard Error (SE)
Effect-size	0.66	0.08
Author	0.79	0.11

In the fixed-effects model, homogeneity test is not significant (Q = 18.5, df = 12, p = 0.101). The ESL example has 13 authors, 17 studies, and 23 effect sizes. ELL example has 13 author, and Table 20 estimates overall effect size using author as an unit. The difference between author and effect size unit analysis shows the impact of author effect for ELL example. The shifting unit of analysis is also applicable to the same data issue, too. If a meta-analysis consists of several large data sets, the data set can be a unit of analysis. It will show the effect of same data.

CHAPTER VI

CONCLUSION AND DISCUSSION

What I have learned

Whenever a reviewer conducts a meta-analysis, he or she is likely to encounter several papers by the 'same authors' and papers using the 'same data' sets. Large data sets frequently used for research can be the source of papers using the same data sets in meta-analysis. The pressure for new faculty to earn tenure encourages the writing of many papers on the same topic, and using the same data set(s) may facilitate that task. The desire for researchers to establish themselves as experts on a particular topic provides another reason for why there are so many same-author papers in meta-analysis. Reviewers will typically encounter these kinds of dependence when conducting meta-analysis, and this dilemma indicates the necessity and importance of this research.

Since Glass (1976) coined the word 'meta-analysis', many research syntheses have been conducted and the quantity of papers using meta-analysis has increased as research and knowledge have exploded in all academic areas. Reviewers have paid some attention to within-study dependence due to the repeated use of single samples, but they have not paid enough attention to "between studies" dependence due to 'same author' and 'same data' papers. I have learned several important lessons about these topics while conducting this research.

First, I found that the prevalence of same-author papers is high, and almost every meta-analysis has same-author papers. There are more same-author papers than I expected. I even learned that a journal editor may accept duplicate studies if they think the studies are beneficial to the public. This study found many possible explanations of the existence of same-author and same-data studies in journal editors' criteria, but the appropriate methods have not been applied to deal with these issues in meta-analysis.

Second, based on two case studies, I found that samples of same-data studies are not exactly the same, such as within study dependence, due to the repeated use of a single sample. The grade of students and number of samples were different in the same-data case study. I also found that same-author studies used more similar samples and measurement instruments than I initially expected. I found that the main characteristics of ‘same author’ and ‘same data’ studies reflect ‘nested’ and ‘unbalanced’ situations; therefore, HLM is a possible way for dealing with this dependence when conducting meta-analysis.

Third, my research focused on the variability of primary studies and compared the variability among same-author and same-data studies with variation for other studies. My research adopted the methods of quantifying heterogeneity proposed by Higgins and Thompson (2002). While this use is technically sound, simulation studies showed that I^2 is mostly negative for same-author and same-data studies. However, the Birge ratio and between studies variance do not show problematic behavior, and they are conceptually similar to the other measures.

Fourth, in the same-data simulation study, I found that the study-size ratio is the most important factor, in addition to the overlap ratio. I learned that the confidence interval can show the degree of heterogeneity, in addition to representing the precision of the effect-size estimate. In the same-author simulation studies, the value of the correlation showed a pattern similar to that for the study-size ratio in the same-data simulation study.

Fifth, the results of empirical studies are parallel to those of simulation studies. I adopted the idea of ‘shifting unit of analysis’ (Cooper, 1988), and I showed that the ‘author’ and ‘data’ set can be a unit of analysis when conducting meta-analysis. I proposed getting one effect size for each author or each data set in Chapter V.

Practical implications

Based on the literature review, simulation study, and empirical analysis, I recommend several things for reviewers. This practical implication can be a guideline for how to deal with same-author and same-data dependence when conducting meta-analysis.

First, reviewers should distinguish same-author and same-data dependence from the dependence due to multiple outcomes within a study. When reviewers find same-author and same-data papers, they need to keep all papers and designate a categorical variable that represents the ‘author’ and ‘source of data’ for coding. Reviewers should not discard or simply average all findings by the same author or from the same data set. If reviewers discard all papers and choose just one paper per author and data set, this will cause loss of information, which will sometimes be very severe.

Second, for the same-data studies, reviewers should check the study-size ratio and overlapping ratio for papers using the same data. If the study-size ratio is larger than 20%, the reviewers should investigate the dependence of same-data papers and show the impact of same-data papers using sensitivity analysis and shifting unit of analysis.

Third, for the same-author studies, synthesists should investigate the characteristics of the samples, such as age, location, incentive for participation, and gender, in addition to similarities of the measurement instruments and research methods. If reviewers find the relatedness in the samples and/or measurement instruments, then they should investigate the dependence of same-author studies.

Fourth, if reviewers suspect between studies dependence, then they should investigate between studies dependence using the proposed indices (such as *H*, the Birge ratio, width of confidence interval, and intraclass correlation) when comparing different-author and different-data studies. Computing *H* and the Birge ratio or graphing distributions of effect size will show whether between studies dependence exists or not.

Fifth, if reviewers find between studies dependence by using the proposed indices, they can conduct categorical analysis, sensitivity analysis, and use shifting unit of analysis by using author and data source as main factors for checking the impact of same-author and same-data studies. The ‘author’ and ‘data source’ can be a possible unit of analysis for these issues, as described in Chapter V (p. 111). Initially, the effect size is the basic unit of analysis in meta-analysis; however, the study can be a unit of analysis if multiple outcomes exist, as Cooper (1998) proposed. By the same rationale, if several authors wrote many papers, such as in the ESL example, an author can be a unit of analysis, such as in Table 20 in Chapter V. After reviewers show the different results based on different units of analysis (effect size, study, and author/data set), then they will find more generalizable findings without losing information. These methods are easy to use for reviewers who do not have a high level of statistical knowledge.

Sixth, let’s consider the practical use of the study-size ratio when study-size ratios differ from each other. If study sizes are different from each other in a real meta-analysis, reviewers need to check the overlap ratio first, and then check the average study-size ratio to determine whether average study-size ratio is big enough to suspect between studies dependence. In such cases, reviewers can adopt the shifting unit of analysis idea to estimate overall effect size and estimate effect size for sub-categories.

Limitation

This study has several limitations. It proposed several ways for dealing with between studies dependence. Each method can be a separate research topic in meta-analysis. This study focused on showing the overall methods for dealing with these issues, but each method cannot be researched in detail. For example, I argue for the quantification of heterogeneity, but the H and I^2 indices are not applicable in every situation, as Rucker et al. (2008) have reported. Computing the I^2 led to predominantly negative values in my simulation; thus, the Birge ratio should be reported instead of I^2.

Second, this study did not directly show the amount of dependence of 'same author' and 'same data' studies. If reviewers can show the amount and direction of dependence, this would be great. Stevens and Taylor (2009) showed a way to quantify hierarchical dependence, but their study needs a covariance matrix.

Third, in the simulation, the data generation is slightly different from the real situations. In particular, for the same-author dependence, this study used correlated effect sizes to represent studies by the same author, but this will be slightly different from real 'same author' dependence. I acknowledge the difficulty and limitation of generating data to represent the 'same author' situation.

Fourth, this research investigated the pattern of homogeneity using the mean difference effect size for the 'same data' problem and the correlation effect size for the 'same author' issue, however, the patterns of homogeneity found here would be expected to be similar for other type of effect sizes including odds ratio effect sizes.

Future research

This study proposed that 'author' and 'data' can be an analysis unit in Chapter V; however, the use of a different unit of analysis may lead to different results in meta-analysis. For example, in meta-analysis, reviewers can use 'effect size', 'study', and 'author' or 'data' as the unit of analysis, and the analysis results could be different using these three units. Further research is needed to determine how much difference is negligible and how much difference needs to be reported in detail.

Stevens and Taylor (2009) compared fixed-effects models, random-effects models, and a hierarchical Bayes approach for hierarchically dependent sub-studies. Their studies actually have multiple variables or groups of participants, but their hierarchical dependence is conceptually similar to the dependence of same-author and same-data studies. Their approach requires a covariance matrix; thus, further research is needed to assess whether and how their method can be applied to the dependence of same-author and same-data studies.

More empirical meta-analyses having same-author studies and same-data studies need to investigate the methods proposed in this research. I conducted case studies (Chapter 2) and empirical analyses (Chapter 5); however, re-analyses of empirical meta-analyses having same-author and same-data studies will provide more information about these issues. For example, Wang, Jiao, Young, Brooks, and Olson's (2007, 2008) meta-analyses have many same-author studies, but they do not investigate dependence. I need to search for more empirical meta-analyses having these issues and re-analyze these meta-analyses using the proposed methods.

The phenomena of 'same author' and 'same data' issues will increase as time goes on because knowledge is rapidly increasing. More research and systematic guidelines about how to deal with these issues are needed for reviewers.

APPENDIX A

PREVALENCE OF META-ANALYSIS USING SAME AUTHOR STUDIES

		Psychological Bulletin			**Review of Education Research**		
Year	Issue No	Total	MA	Review	Total	MA	Review
2008	1	6	2	0	3	0	1
	2	7	1	1			
2007	1	7	2	0	4	1	1
	2	9	0	2	5	0	1
	3	9	0	3	5	0	0
	4	7	1	0	5	1	1
	5	8	1	0			
	6	10	2	0			
2006	1	7	3	2	4	1	1
	2	6	2	0	4	2	2
	3	6	1	0	4	1	1
	4	9	3	0	8	0	0
	5	9	3	1			
	6	7	4	0			
2005	1	8	0	0	4	2	0
	2	10	1	0	4	1	1
	3	11	1	1	4	1	2
	4	7	1	1	3	0	0
	5	12	1	1			
	6	4	1	0			
2004	1	7	2	0	3	1	0
	2	9	2	0	4	1	0
	3	14	1	1	3	0	1
	4	9	2	0	4	1	0
	5	7	1	0			
	6	7	1	0			
	Total	212	39	13	71	13	12
	Percent		18%	6%		18%	17%

APPENDIX B

FREQUENCIES OF SAM AUTHORS IN META-ANALYSIS

ID	**Author**	**Year**	**Number of primary studies**	**Number of papers by same author**									**Total number of same author paper**	**% of same author paper**	**% of largest one**
				2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10 and more**			
1	Bar-Haim et al.	2007	172	8	5		2	2		1		16 paper(1)	77	45%	9%
2	Steel	2007	216	17	4	3			1		1	26 paper (1) 12 paper (1)	112	52%	12%
3	Tolin et al.	2006	290	16	2	1	2						52	18%	2%
4	Malle	2006	173	13									26	15%	1%
5	Puetz et al.	2006	66	5	1								13	20%	5%
6	Frattaroli	2006	250	13		2							28	11%	2%
7	Glasman	2006	70	4	3	1							21	30%	6%
8	Bettencourt et al.	2006	63		3			1					15	24%	10%
9	Grabe	2006	98	3	4								18	18%	4%
10	Bosch	2006	380	6	3	3	2			1	1	12(1)	72	19%	3%
11	Cepeda	2006	184	13	5	4		1	1			11(1)	81	44%	6%
12	Web et al.	2006	47	2									4	9%	4%
13	Durantini	2006	98	9	2			1					30	31%	6%
14	Weis	2006	35	4		1							12	34%	11%
15	Else-Quest	2006	189	16	1	2	1						44	23%	3%
16	Roberts	2006	92	7	1	2							25	27%	5%
17	Swanson	2006	28	4									8	29%	7%
18	Cooper	2006	32	3		1							10	31%	13%
19	Kuncel	2005	37	5	1								13	35%	8%
20	Gijbels	2005	40	3									6	15%	5%
21	Nesbit	2006	122	2	4								16	13%	2%

APPENDIX C

THE LIST OF CLASS SIZE PAPERS IN THE DATA GATHERING STAGES

	Aauthor	Year	Source	Title
1	Achilles , C. M., et al.	1996	Educational Leadership	Students Achieve More in Smaller Classes
2	Achilles , C. M., et al.	1997	Educational Leadership	using Class size to reduce the equity gap
3	Achilles, C, M., et al.	2001	AERA conference paper	Reasonable-Size Classes for the Important Work of Education in Early Elementary Years: A Manual for Class-Size Reductions So All Children Have Small Classes and Quality Teachers in Elementary Grades. Revised
4	Achilles, C. M.	2003	Conference paper	How Small Classes Help Teachers Do Their Best: Recommendations from a National Invitational Conference
5	Achilles, C. M.	1998	AERA conference paper	If not Before, At least now
6	Achilles, C. M., et al.	2002	Conference paper	Making sense of Continuing and Renewed Class-Size Findings and Interest
7	Achilles, C. M., et al.	2003	Conference paper	School Improvement Should Rely on Reliable, Scientific Evidence. Why Did "No Child Left Behind" Leave Class Size Behind?
8	Achilles, C. M.	1998	Conference paper	Small-Class Research Supports What We All Know (So, Why Aren't We Doing It?)
9	Achilles, C. M., et al.	1995	Conference paper	Success Starts Small (SSS): A Study of Reduced Class Size in Primary Grades of a Fully Chapter-1 Eligible School
10	Achilles, C. M., et al.	1993	Conference paper	The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990 - 1991): A Legacy from Tennessee's Four-Year (K-3) Class -Size Study (1985-1989), Project STAR Paper #7
11	Achilles, C. M., et al	1994	Conference paper	The Multiple Benefits of Class-Size Research: A Review of STAR's Legacy, Subsidiary and Ancillary Studies
12	Achilles, C. M., et al.	1995	Conference paper	Analysis of Policy Application of Experimental Results: Project Challenge
13	Achilles, C. M., et al.	1998	Conference paper	Attempting to understand the class size and pupil-teacher ratio (PTR) confusion: A pilot study
14	Achilles, C. M., et al.	1999	Conference paper	Some Connections between Class Size and Student Successes
15	Ahmed, A. U.	2006	Journal: Wrold Development	Do Crowded Classrooms Crowd Out Learning? Evidence from the Food for Education Program in Bangladesh
16	Akerhielm, K.	1995	Ecomomics of Education Review	Does Class matter?
17	Angrist, J, D., et al	1999	The Quartery Journal of Economics	Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement

18	Asadullah, M. N.	2005	Applied Economics letter	The effects of class size on student achievement: evidence from Bangladesh
19	Averett, S. L.	2002	International Handbook on the Economics of Education	Exploring the effect of class size on student achievement: What have we learned over the past two decades?
20	Becker, W. E., et al.	2001	Economics of Education Review	Student performance, attrition, and class size given missing student data
21	Bell, J. D.	1998	State Legislatures	Smaller = Better?
22	Besser, A. D.	2000	Dissertation	The impact of class size reduction on factors of the classroom that affect student achievement within second grade classrooms of the Venice/Westchester cluster of the Los Angeles unified school district
23	Bingham, C. S.	1994	Conference paper	Class Size as an Early Intervention Strategy in white-Minority Achievement Gap Reduction
24	Blatchford, P.	2003	Learning and Instruction	A systematic observational study of teachers' and pupil's behavior in large and small classes
25	Blatchford, P., et al.	1994	Oxford Review of Education	The Issue of Class Size for Young Children in Schools: What Can We Learn from Research?
26	Blatchford, P., et al.	1998	British Journal of Educational Studies	Research Review: The effects of Class size on classroom processes: 'It's a Bit like a Treadmill - Working hard and getting nowhere fast!'
27	Bonersronning, H.	2003	Southern Economics Journal	Class-size effects on Student Achievement in Norway: Patterns and Explanation
28	Boozer, M. A., et al.	2001	Economic growth center YALE University (Center disscussion paper)	The effects of class size on the long run growth in reading abilities and early adult outcomes in the Christchurch health and development study
29	Bourke, S.	1986	American Educational Research Journal	How Smaller Is Better: Some Relationships between Class Size, Teaching Practices, and Student Achievement
30	Brewer, D. J., et al.	1999	EEPA	Estimating the Cost of National Class Size Reductions under Different Policy Alternatives
31	Card, D., et al.	1998	Annals of the American Academy of Political and Social Science	School Resources and Student Outcomes
32	Chatman, S.	1996	Conference paper	Lower Division Class Size at U.S. Postsecondary Institutions. AIR 1996 Annual Forum Paper
33	Cooper, H. M.	1989	Educational psychology	Does Reducing Student to Instructor Ratios Affect Achievement?
34	Davis, F. E.	2000	Dissertation	The effects of Class size reduction on student achievement and teacher attitude in first grade
35	Deutsch, F. M.	2003	NASSP	How Small Classes Benefit High School Students

36	Dharmadasa, I.	1995	Conference paper	Class Size and Student Achievement in Sri Lanka
37	Driscoll, D., et al.	2000	Economics of Education Review	School district size and student performance
38	Eash, M. J.	1964	American Educational Research Journal	The Effects of Class size on Achievement and Attitudes
39	Ecalle, J., et al.	2006	Journal of School Psychology	Class-size effects on literacy skills and literary interest in first grade: A large-scale investigation
40	Finn, J. D., et al	1990	American Educational Research Journal	Answers and Questions about Class Size: A Statewide Experiment
41	Finn, J. D., et al.	1989	Peabody Journal of Education	Carry-Over Effects of Small Classes
42	Finn, J. D., et al.	1999	EEPA	Tennessee's Class Size Study: Findings, Implications, Misconceptions
43	Floger, J.	1989	Peabody Journal of Education	Evidence from Project STAR about Class Size and Student Achievement
44	Floger, J.	1989	Peabody Journal of Education	Lessons for Class Size Policy and Research
45	Friedkin, N. E., et al.	1988	EEPA	School systems and performance: A contingency perspective
46	Gilman, D. A., et al.	2003	educational Leadership	Should we try to keep class sizes small?
47	Glass, G. V et al.	1980	American Educational Research Journal	Meta-Analysis of Research on Class Size and Its Relationship to Attitudes and Instruction
48	Glass, G. V., et al.	1979	EEPA	Meta-Analysis of Research on Class Size and Achievement
49	Glass, G. V., et al	1982	Book	School Class Size research and policy
50	Goldstein, H., et al.	1998	British Educational Research Journal	Class size and Educational Achievement: A Review of Methodology with Particular Reference to Study Design
51	Goldstein, H,. et al.	2000*	Applied Statistics	Meta-Analysis Using Multilevel Models with an Application to the Study of Class-size effects
52	Grissmer, D.	1999	EEPA	Conclusion: Class-size effects: Assessing the Evidence, Its Policy Implications, and Future Research Agenda
53	Haenn, J. F.	2002	AERA conference paper	Class size and Student Success: Comparing the Results of Five Elementary Schools Using Small Class Sizes
54	Halbach, A., et al.	2001	Educational Leadership	Class Size Reduction: From Promise to Practice
55	Hallinan, M. T., et al.	1985	American Journal of Education	Class Size, Ability Group Size, and Student Achievement
56	Hanushek, E. A.	1999	EEPA	Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigation of Class-size effects

57	Hanushek, E. A.	1999	High School Magazine	Good Politics, Bad educational Policy
58	Harman, P., et al.	2002	AERA conference paper	Observing Life in Small-Class Size Classrooms
59	Harvey, B. H.	1994	Conference paper	To Retain or Not? There is No Question
60	Harvey, B. H.	1994	Conference paper	The Effect of Class size on achievement and Retention in the Primary Grades: Implications for Policy Makers
61	Hedges, L. V.	1983	American Educational Research Journal	The Effects of Class Size: An Examination of Rival Hypotheses
62	Hedges, L. V.	1994	Educational Researcher	An Exchange: Part I: Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Students Outcomes
63	Hiestand, N. T.	1994	AERA conference paper	Reduced Class Size in ESEA Chapter 1: Unrealized Potential?
64	Hood, A.	2003	Book	A Parent's Guide to Class Size Reduction
65	Hoxby, C, M.	2000	The Quarterly Journal of Economics	The Effects of Class Size on Student Achievement: New Evidence from Population Variation
66	Johnson, J., et al.	1990	Government paper	The state of Tennessee's Student/Teacher Achievement Ratio (STAR) Project, Final summary report 1985-1990
67	Keith, P. B., et al.	1993	Conference paper	Investigating the Influences of Class Size and Class Mix on Special Education Student Outcomes: Phase One Results
68	Kennedy, P. E., et al.	1996	Economics of Education Review	Class size and Achievement in Introductory Economics: Evidence from the TUCE Data
69	Kiger, D. M.	2000	Dissertation	A Case Study Evaluation of a Fortified Class Size Reduction Program
70	Kirst, M., et al.	1998	AERA conference paper	A Plan for the Evaluation of California's Class Size Reduction Initiative
71	Krieger, J.	2003	Conference paper	Class size Reduction: Implementation and Solutions
72	Krieger, J. D.	2002	Conference paper	All we need is a little class
73	Lapsley, D. K., et al.	2001	AERA conference paper	Indiana's "Class Size Reduction" Initiative: Teacher Perspectives on Training, Implementation and Pedagogy.
74	Lapsley, D. K., et al.	2002	Conference paper	Teacher Aides, Class Size and Academic Achievement: A Preliminary Evaluation of Indiana's Prime Time
75	Lee, V. E., et al.	2000	American Educational Research Journal	School size in Chicago Elementary Schools: Effects on Teacher's Attitudes and Students' Achievement
76	Lee, V. E., et al.	1997	EEPA	High School Size: Which Works Best and for Whom?
77	Lewit, E, M.	1997	The Future of Children	Class Size

78	Lindahl, M.	2005	Scandinavian Journal of Economics	Home versus School Learning: A New Approach to Estimating the Effect of Class Size on Achievement
79	Lindsay, P.	1982	EEPA	The Effects of high school size on student participation, satisfaction, and attendance
80	Lindsey, J. K.	1974	Comparative education review	A Re-analysis of Class size and Achievement as Interacting with Four Other Critical Variables in the IEA mathematics Study
81	Lou, Y., et al.	1996	Review of Educational Research	Within-Class Grouping: A Meta-Analysis
82	May, K. O.	1962	The American Mathematical Monthly	Small Versus Large Classes
83	McGiverin, J., et al.	1989*	The Elementary School Journal	A Meta-analysis of the relation between Class size and Achievement
84	Miller-Whitehead, M.	2003	conference paper	Compilation of Class Size Findings: Grade Level, School, and District
85	Mitchell, B. M.	1969	Peabody Journal of Education	Small Class Size: A Panacea for Educational ills?
86	Mitchell, D. E., et al.	2001	AERA conference paper	Competing Explanations of Class Size Reductions Effects: The California Case
87	Mitchell, D. E., et al.	1989	Peabody Journal of Education	Modeling the Relationship between Achievement and Class size: A Re-analysis of the Tennessee Project STAR Data
88	Mitchell, R. E.	2001	Dissertation	Class Size Reduction Policy: Evaluating the Impact on Student Achievement in California
89	Mitchell, R. E.	2000	conference paper	Early Elementary Class-Size Reduction: A Neo-Institutional Analysis of the Social, Political, and Economic Influences on State-Level Policymaking
90	Molnar, A., et al.	1999	EEPA	Evaluating the SAGE Program: A Pilot Program in Targeted Pupil-Teacher Reduction in Wisconsin
91	Moody, W. B., et al.	1973	Journal for Research in Mathematics Education	The Effects of Class Size on the Learning of Mathematics: A Parametric Study with Fourth-Grade Students
92	Mosteller, F.	1995	The Future of Children	The Tennessee Study of Class Size in the Early School Grades
93	Munox, M. A., et al.	2002	AERA conference paper	Voices from the Field: The Perception of Teachers and Principals on the Class Size Reduction Program in a Large Urban School District
94	Muthen, B. O.	1991	Journal of Educational Measurement	Multilevel Factor analysis of Class and Student Achievement Components
95	Nye, B. A., et al.	1992	Technical Reports	The Lasting Benefits Study. A Continuing Analysis of the Effects of Small Class Size in Kindergarten through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels: Fifth Grade.
96	Nye, B. A., et al.	2002	EEPA	Do Low-Achieving Students Benefit More from Small Classes? Evidence from the Tennessee Class Size Experiment
97	Nye, B. A., et al.	2000	American Journal of Education	Do the Disadvantaged Benefit More from Small Classes? Evidence form Tennessee Class Size Experiment

98	Nye, B. A., et al.	1999	EEPA	The Long-Term Effects of Small Classes: A Five Year Follow-Up of Tennessee Class Size Experiment
99	Nye, B. A., et al.	1993	AERA conference paper	Class-Size Research from Experiment to Field Study to Policy Application
100	Odden, A.	1990	EEPA	Class Size and Student Achievement: Research-Based Policy Alternatives
101	O'Sullivan, M. C.	2006	International Journal of Educational Development	Teaching large classes: The international evidence and a discussion of some good practice in Uganda primary schools
102	Peake, K.	2001	Dissertation	The Effect of class size: A study of second and third grade student achievement in the school district of Greenville county, South Carolina
103	Phi Delta Kappa	2002	Phi Delta kappa	Class size and its effect: Review
104	Prais, S. J.	1996	Oxford Review of Education	Class-Size and Learning: The Tennessee Experiment--What Follows?
105	Remmers, H. H.	1933	The Journal of Higher Education	Class-Size
106	Ritter, G. W., et al.	1999	EEPA	The Political and Institutional Origins of a Randomized Controlled Trial on Elementary School Class Size: Tennessee's Project STAR
107	Rountree, M. L.	1997	Dissertation	The State-initiated class size reduction program: A preliminary study of the initial district response
108	Scudder, D. F.	2002	AERA conference paper	An Evaluation of the Federal Class-Size Reduction Program in Wake County, North Carolina---1999-2000
109	Sharp, M. A.	2002	Conference paper	An Analysis of Pupil-teacher Ratio and Class Size: Difference that make a difference
110	Sharp, M. A.	2003	Conference paper	Summary of an Analysis of Pupil-Teacher Ratio and Class Size: Differences That Make a Difference and Its Implications on Staffing for Class-Size Reduction.
111	Shpason, S. M.	1980	American Educational Research Journal	An Experimental Study of the Effects of Class Size
112	Simpson, S.	1980	Journal	Comment on "Meta-Analysis of Research on Class Size and Achievement"
113	Slavin, R. E.	1989	Educational Psychologist	Class size and Student Achievement: Small Effects and Small Classes
114	Spitzer, H, F	1954	The Elementary School Journal	Class Size and Pupil Achievement in Elementary Schools
115	Stasz, C., et al.	2000	EEPA	Teaching Mathematics and language Arts in Reduced Size and Non-Reduced Size Classrooms
116	Stecher, B. M., et al.	2003	EPAA	The Relationship between Exposure to Class Size Reduction and Student Achievement in California
117	Tobin.,. et al.	1987	Comparative education review	Class size and Student /Teacher Ratios in the Japanese Preschool
118	Townsend, T.	1998	AERA conference paper	Does Money Make a Difference?

119	Urquiola, M.	2006	The Review of Economics and Statistics	NOTE: Identifying class-size effects in developing countries: evidence from rural BOLIVIA
120	Wasley, P. A.	2002	Educational Leadership	Class size, School size
121	Webb, C. E.	2003	Dissertation	The effects of a class size reduction on elementary student reading achievement in a Midwestern urban school district
122	Wilkins, W. E.	2002	Conference paper	No Simple Solution: Do Smaller Classes, More Experienced and Educated Teachers, and Per Pupil Expenditure Really Make a Difference?
123	WoBmann, L., et al.	2006	European Economic Review	Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS

APPENDIX D

LIST OF 16 STUDIES IN ANALYSIS STAGE FOR META-ANALYSIS OF CLASS SIZE STUDIES

ID	Author	YEAR	Source	sampling	DATA	Subject	STAR	*4 STATE	PUBLISH	Grade
2	Molnar, A., et al.	1999	EEPA	quasi-experimental	SAGE (Wisconsin)	math, reading, language	0	1	1	1,2
5	Mosteller, F.	1995	The Future of Children	Random assign	STAR	math, reading	1	1	1	3
7	Lapsley, D. K., et a.l	2002	conference paper	cluster sampling	PRIME TIME (INDIANA)	language, math, reading, total	0	1	0	4
9	Achilles, C. M., et al.	1994	conference paper	Random assign	STAR	math, reading	1	1	0	K,1,2,3
11	Finn, J. D., et al.	1898	Peabody Journal of Education	Random assign	STAR	math, reading, language, science, social ss, sskill	1	1	1	1
12	Dharmadas, Indranie	1995	conference paper	Not random: pretest, posttest method	Sri Lanka	math, mother tongue	0	0	0	4
13	Haenn, J. F.	2002	AERA conference paper	Matched sample	North Carolina	letter, reading	0	0	0	K,1
14	Peake, K.	2001	Dissertation	Not random: pretest, posttest	South Carolina	language, OLSAT, reading	0	0	0	2,3
15	Davis, F. E.	2000	Dissertation	Random assign	Georgia	math, reading	0	0	0	1
16	Ecalle, J., et al.	2006	Journal of School Psychology	Random assign	France	total score	0	0	1	1
17	Spitzer, H. F.	1954	The Elementary School Journal		Iowa	math, reading, language, study skill	0	0	1	3,4
18	Goldstein, H., et al.	1998	British Educational Research Journal	Random assign	STAR	math, reading	1	1	1	K,1

19	Johnston, et al.	1990	Report	Random assign	STAR	Math, Reading	1	1	0	K, 1, 2, 3
20	Nye, et al.	1992	Report	Random assign	STAR	R/M/S, Lang, Study skill, SS	1	1	0	5
21	Finn and Achilles	1990	American Educational Research Journal	Random assign	STAR	R/M, Study skill	1	1	1	1
22	Finn and Achilles	1999	Educational Evaluation and Policy	Random assign	STAR	Reading, Math	1	1	1	K, 1, 2, 3

* Four states include Tennessee, California, Wisconsin, and Indiana. These states had state-wide class size reduction policy.

APPENDIX E

CHARACTERISTICS OF 8 STAR CLASS SIZE STUDIES

Author/Year	Mosteller 1995	Achilles 1994	Finn 1989	Goldstein1998	Johnston 1990	Nye, 1992	Finn/Achilles 1990	Finn/Achilles 1999
ID	2	5	9	11	19	20	21	22
Journal	The Future of Children	Conference paper	Peabody Journal of Education	British Educational research journal	Report	Report	American Educational Research Journal	Educational Evaluation Policy
Sample	First grade	K, 1, 2, 3	4th grade	K, 1	K, 1, 2, 3	5	First grade	K, 1, 2, 3
Research method	Summary of STAR project				Quasi-experimental	Follow up study	Experiment	Experiment
Sample size	3892	110-136	2662	4197-6871	4200 (1900, 2300)	3045	804-5192	4744-6572
Small class sample size	1620	63-75	1412	1429-2762	1900	1578	346-2233	2040-2826
Large class sample size	2272	46-61	1250	2768-4109	2300	1407	458-2959	2704-3746
Subject	Reading, Math	Reading, Math	Reading, Language, Math, studyskill, Science	Reading, Math	Reading, Math	Reading, Math, Lang, Science, Social Science, Study Skill	R/M/Study Skill	Reading, Math

Sample selection	Using total 1 st grade student	N/A	Follow up study: 1 year after 4th grade	N/S: reanalysis	Four years Data	5th Grade, LBS analysis	1st grade for two years small class	1985-1989, K-3
Reporting	Effect size directly, with sample size	MEAN, SD, Sample size, effect size	Mean, SD, sample size	Effect size, N	Effect size, N	Effect size, N	Mean, SD, N	Effect size, N
Experiment Period	1st year result	N/A	One year after first 4 year experiment	First four year result summary	1985-1989	1985-1989	Two years Small class (K-1)	After 4 years Experiment (1985-1989)
DATA Source	Finn, J.D., and Achilles, C. M. 1990 “Answers and questions about class size” American Educational Research Journal	Classes included all assigned students regardless of years of participation of STAR: not clear	Follow up stud data base for 4th grade students from 58 schools	Goldstein & Blatchford (1997) “Class size and student achievement: a methodological review” presented for UNESCO	First year and subsequent 4 years	1985-1989 (K-3) small class, and follow up 1990-1991 school year 5th grade data	End of two Years (K-1)	K-3 (1985-1989)

Strong point	First year has very good experiment condition. After first year, condition changed	K-3 result, enough information: mean, SD, N	Whole data, enough info for follow up study	Methodological issue in Class size like RCT	Government Summary Report	LBS executive Summary	First Two year experienced focus	Summary and New Analysis in 1999
Weak point	Cited effect size, sample size	No mention on small N, source	Follow-up vs original Diff	Cited effect size	SES, Minority Effect size without N	Minority without N	Region, Minority without N	Overall
Unit	Aggregated data	No mention	Aggregated data	Aggregated data	Aggegated data	Aggregated data	Aggregated	Aggregated

APPENDIX F

REVIEW OF 26 'SAME AUTHOR' PAPERS IN META-ANALYSIS

ID	Paper	Year	Sample								Instrument	Result
			N	women/men	Age	Class/Race	incentive	race	grade	Place		
1	Ferrari et al	1998	546		19					Northeastern	PASS, QAE	Mean, SD
2	Ferrari & Dovidio	2001	58	40/18	21	intro psych				Northeast	DP	Mean, percent
2	Ferrari & Dovidio (2)	2001	100	74/26	21	intro psych				Midwestern		
3	Ferrari & Patel	2004	160	103/57	20	intro psych			first /sopho.	Midwestern	AIP	Corr.
4	Ferrari et al	1997	61		19	intro psych	extra credit			Midwestern	DP, AIP, DNAQ	Mean, SD, t test
4	Ferrari et al (2)	1997	58	40/18	22	social psych	extra credit	68 % Caucasian	70% junior or senior	Midwestern	DP, AIP, SDSCM	Mean, SD, Corr, ANOVA
5	Ferrari, J. R	2000	142	80/62	21	intro psych	extra credit	61 % Caucasian		Midwestern	DP, AIP, GP, ADDHC, BPS	t-test, Corr
6	Ferrari & Dovidio	2000	130	105/25	20	intro psych	extra credit				DP	Corr
7	Ferrari & Scher	2000	37	30/7	20	intro psych	35/50 dollar			Midwestern	FIAR, PIAC	ANOVA
8	Ferrari, J. R	2001	93	51/42	20	intro psych	extra credit			Midwestern	AIP	M, SD, ANOVA
8	Ferrari, J. R (2)	2001	226	178/48	20	intro psych	extra credit		80% first/second	Midwestern	AIP	M, SD, ANOVA
9	Ferrari & Tice	2000	59	40/19		intro psych	extra credit			Midwestern	GP	Corr
9	Ferrari & Tice (2)	2000	88	48/40		intro psych	extra credit			Midwestern	GP	Mean, SD

10	Ferrari & Becke	1998	103	88/35	19	psych	extra credit	80W Caucasian	first/second	Northeast	PASS, QAE	M, SD, MANOVA
11	Ferrari, J. R	1991	54	37/16	19	intro psych	extra credit			Rural private college	PS, DMQ, ISI, CAS	Corr, F, T
12	Harriot, J & Ferrari, J. R.	1996	211	122/89	48		volunteers	adult sample			APS, APS, IS	M, SD, CORR
13	Ferrari, J. R	1989	116	80/36				College student			PA, GP	
14	Ferrari, J. R.	1991	241					46 pro/ 52 non-pro			DP, BP	
14	Ferrari, J. R. (2)	1991	287					48 pro/ 54 non-pro			Intelligence, ISI	M, SD
15	Ferrari, & Emmons	1995	277	205/72	18-21	intro psych	extra credit	75% first grade		Northeast, small private	DP, AIP	CORR, Regress
16	Ferrari, J. R	1995	262	211/51	18	intro psych	extra credit				AIP, PCI	Corr, M, SD
16	Ferrari, J. R.	1995	136	114/22	18			78% first grade			PCI, DP	Corr, M, SD
16	Ferrari, J. R.	1995	65	39/25	44			8 years of therapy			DP, PCI	Corr, M, SD
17	Ferrari, J. R.	1992	52, 59	30/22, 44/15	32, 20	intro psych				Public univ. in New York	AIP, GP	Corr
17	Ferrari, J. R.	1992	215	134/81	34						AIP, GP	Corr
18	Ferrari, et al	1995	324, 375, 171	238/86, 270/105, 137/34				three undergraduate institute		North east	PASS, ISI	M, SD, CORR
19	Ferrari, et al	1994	84		19, 47				Students, Parents		DP, AIP	CORR, T
20	Effect & Ferrari	1989	111	84/27		intro psych	extra credit				DP	CORR
21	Ferrari, J. R	1992	307	204/103	22		volunteers				PS	M, SD, Corr
22	Ferrari, J. R.	1991	120		18	intro psych		57 pro / 63 non-pro			DP, BP	M, SD

23	Ferrari & Emmons	1994	202, 161	112/49, 93/23	18-21	intro psych	extra credit				DP, AIP	Corr
24	Ferrari, J. R	1994	263	202/61	21	intro psych	extra credit				DP, AIP	CORR, Regress
25	Ferrari, J. R	1992	319	241/78	18					North east	MBTI, PASS	t-test, Corr

APPENDIX G

LIST OF 17 STUDIES IN ANALYSIS STAGE FOR META-ANALYSIS OF ESL STUDIES

ID	Authors	Year	PUBLISH	Kubota	N	NE	NC	T
1	Bouton	1994	Published	1	46	14	32	1.09
2	Davidson	1995	Unpublished	1	36	17	19	1.62
3	Doughty	1991	Published	1	14	8	6	0.42
4	El-Bana	1994	Unpublished	1	97	46	51	1.19
5	Ellis	2003	Published	1	28	14	14	0.16
5	Ellis	2003	Published	1	28	14	14	0.68
5	Ellis	2003	Published	1	28	14	14	0.76
6	Ellis	1999	Published	1	34	16	18	0.66
7	Fukuya	1998	Unpublished	1	16	8	8	1
7	Fukuya	1998	Unpublished	1	19	11	8	0.85
8	Kim	1996	Published	1	26	13	13	0.42
9	Kubota	1994	Published	2	40	20	20	0.51
9	Kubota	1994	Published	2	40	20	20	0.59
10	Kubota	1995	Published	2	84	42	42	0.37
10	Kubota	1995	Published	2	84	42	42	0.38
11	Kubota	1996	Published	2	80	40	40	0.47
12	Kubota	1997	Published	2	48	24	24	0.7
13	Mackey	1999	Published	1	13	7	6	0.37
14	Master	2002	Published	1	46	34	12	0.26
15	Polio	1998	Published	1	30	14	16	0.62
16	Shiinichi	2002	Published	1	25	11	14	0.36
16	Shinichi	2002	Published	1	26	12	14	0.34
17	White	1991	Published	1	108	79	29	1.15

APPENDIX H

R CODE FOR ‘SAME DATA’ ISSUE

```
# Insoo2 study 1<-
# function(nschool = 10, nsmallclas = 2, nbigclas = 2, sizsmall = 15, sizbig = 25, effsiz = 10,
mubig = 300,  sigschool = 2, sigclas = 5, sigstud = 50, nsmpsmall = 10, nsmpbig = 10)

# For later use.
allquantfn<-function(effsiz,wt,Calpha){
numstud<-length(effsiz)
seTdotFE<-sqrt(1/sum(wt))
Tdot<-sum(wt*effsiz)/sum(wt)
# For random effect case.
s2theta<-pmax(var(effsiz)-mean(1/wt),0)
seTdotRE<-sqrt(1/sum(wt) + s2theta)
Qval<-sum(terms<-(effsiz-Tdot)^2*wt)
CIRE<-c(CIRE.low=Tdot-Calpha*seTdotRE,CIRE.hi=Tdot+Calpha*seTdotRE)
CIFE<-c(CIFE.low=Tdot-Calpha*seTdotFE,CIFE.hi=Tdot+Calpha*seTdotFE)
pval<-pchisq(Qval,df=numstud-1,lower.tail=FALSE)
ves<-var(effsiz)
c(Tdot,Qval,pval,CIFE,CIRE,ves)
}

effsizfn<-function(x,y){
nx<-length(x)
ny<-length(y)
mx<-mean(x)
my<-mean(y)
sdx<-sd(x)
sdy<-sd(y)
pooledsd<-sqrt(((nx-1)*sdx^2+(ny-1)*sdy^2)/(nx+ny-2))
effsiz<-(mean(x)-mean(y))/pooledsd
condvar<-(nx+ny)/(nx*ny)+effsiz^2/(2*(nx+ny))
c(nx,ny,mx,my,sdx,sdy,pooledsd,effsiz,condvar)
}

nschool<-49                        # make total population 10,000 (205*49)
nsmallclas<-6
nbigclas<-5
sizsmall<-15
sizbig<-25

nrep<-1000 #1000
nsmpsmall<-7                        # make a sample size 200 7*15 = 105
nsmpbig<-4                         # make a sample size 200 4*25 = 100
nstud<-2                          # make a ratio equal to 0.04

stuff<-matrix(0,nstud,9)
dimnames(stuff)[[2]]<-c("nx","ny","mx","my","sdx","sdy","pooledsd",
                    "effsiz","condvar")

simresults<-matrix(0,nrep,8)
dimnames(simresults)[[2]]<-c("Tdot","Qval","pval","CIFE.L",
                    "CIFE.U","CIRE.L","CIRE.H","Vefsiz")

ntotal<-nschool*(nbigclas*sizbig+nsmallclas*sizsmall)

score<-school<-clas<-small<-numeric(ntotal)

meandiff<-10                    # to make effect size .2
mubig<-300
sigschool<-0 #2
sigclas<-0 #5
sigstud<-50
totsigma<-sqrt(sigschool^2+sigclas^2+sigstud^2)
trueeffsiz<-meandiff/totsigma

alpha<-.05
Calpha<-qnorm(alpha/2,lower.tail=FALSE)

for(irep in 1:nrep){

# Simulating a new population from which to sample.

j<-0  # student number
k<-0  # school number
h<-0  # class number

for(ischool in 1:nschool){
  k<-k+1
  schooleff<-rnorm(1,0,sigschool)
  for(iclas in 1:nsmallclas){
    h<-h+1
    claseff<-rnorm(1,0,sigclas)
    for(istud in 1:sizsmall){
      j<-j+1
      score[j]<-mubig+meandiff+schooleff+claseff+rnorm(1,0,sigstud)
      school[j]<-k
      clas[j]<-h
      small[j]<-1
    }
  }

  for(iclas in (nsmallclas+1):(nsmallclas+nbigclas)){
    h<-h+1
    claseff<-rnorm(1,0,sigclas)
   for(istud in 1:sizbig){
     j<-j+1
     score[j]<-mubig+schooleff+claseff+rnorm(1,0,sigstud)
     school[j]<-k
```

```
      clas[j]<-h
      small[j]<-0
    }
  }
}

# New code
avescoresmall<-mean(score[small==1])
avescorebig<-mean(score[small==0])
trueeff<-avescoresmall-avescorebig
score[small==1]<-score[small==1]-trueeff+meandiff

# End new code.

# sampling from the data

indsmall<-unique(clas[small==1])
indbig<-unique(clas[small==0])


for(istud in  1:nstud){
smpsmallclas<-sample(indsmall,nsmpsmall)
smpbigclas<-sample(indbig,nsmpbig)
smpclas<-sort(c(smpsmallclas,smpbigclas))

# Data consists score,school,clas,small.

smp<-clas %in% smpclas

scoresmp<-score[smp]
schoolsmp<-school[smp]
classmp<-clas[smp]
smallsmp<-small[smp]


# The generated population.
# cbind(score,school,clas,small)
# Sample of rows of data.
# cbind(scoresmp,schoolsmp,classmp,smallsmp)
# Effective size (ignoring class and school effects)

smp1<-scoresmp[smallsmp==1]
smp2<-scoresmp[smallsmp==0]
stuff[istud,]<-effsizfn(smp1,smp2)
}

effsiz<-stuff[,8]
wt<-1/stuff[,9]

simresults[irep,]<-allquantfn(effsiz,wt,Calpha)
}
```

```
options(width=130)

simresults

apply(simresults,2,mean)

apply(simresults,2,var)

# Coverage of fixed effect confidence intervals.

L<-simresults[,4]
U<-simresults[,5]
cover.FE<-mean((L<trueeffsiz)&(trueeffsiz<U))

# Coverage of random effect confidence intervals.

L<-simresults[,6]
U<-simresults[,7]
cover.RE<-mean((L<trueeffsiz)&(trueeffsiz<U))

trueeffsiz

cover.FE

cover.RE

save.image(file="1samedata.Rdata")
```

APPENDIX I

R CODE FOR 'SAME AUTHOR' ISSUE

```
# Simulating hypothetical effect sizes for researchers conducting
# multiple studies.
# set.seed(333)

equicorr<-function(sig,rho,k){
# Generate random vector of length k having an equicorrelation matrix.
a<-sig*sqrt(rho)
b<-sqrt(sig^2-a^2)
a*rnorm(1)+b*rnorm(k)
}

# Generates equicorrelated sampling error with variance determined
# by sample size.

equicorrsamperr<-function(rho,k,n){
# rho is the correlation between the sampling error in each study.
# k is the number of studies.
# n is the vector of sample sizes (of length k).
a<-sqrt(rho)
b<-sqrt(1-a^2)
sqrt(1/(n-3))*(a*rnorm(1)+b*rnorm(k))
}

# For computing quantities using the simulated effect sizes.

allquantfn<-function(effsiz,wt,res,Calpha){
if(length(effsiz)==1)return(effsiz)
numstud<-length(effsiz)
seTdotFE<-sqrt(1/sum(wt))
Tdot<-sum(wt*effsiz)/sum(wt)

# Corrected formulas for RE case:
s2theta<-pmax(var(effsiz)-mean(1/wt),0)
seTdotRE<-sqrt(1/sum(wt) + s2theta)
Qval<-sum(terms<-(effsiz-Tdot)^2*wt)
CIRE<-c(CIRE.low=Tdot-Calpha*seTdotRE,CIRE.hi=Tdot+Calpha*seTdotRE)
CIFE<-c(CIFE.low=Tdot-Calpha*seTdotFE,CIFE.hi=Tdot+Calpha*seTdotFE)
pval<-pchisq(Qval,df=numstud-1,lower.tail=FALSE)
ves<-var(effsiz)
ans<-c(Tdot,Qval,pval,CIFE,CIRE,ves)
if(!all(res[1]==res)){
```

```
Qdecomp<-numeric(nres)
for(i in 1:nres) Qdecomp[i]<-sum(terms[researcher==i])
ans<-c(ans,Qdecomp=Qdecomp)
}
ans
}

#######################

nres<-6
nstud<-c(30,10,5,4,3,2)

nnn<-list(
100,
100,
100,
100,
100,
100
)

# Overall effect size:

overall<-.2                    # Overall effect size:

# sig1 and rho1 are for generating equicorrelated "true" effect sizes
#    for the studies done by each researcher.

sig1<-0.1
rho1<-0.2 #0.2                    # Magnitude of correlation

# rho2 is for generating equicorrelated "errors" in the study effects
# with variance determined by the sample size of each study.

rho2<-0

# ssd[i,j] is the matrix of study effect standard deviations,
#   as determined by the sample sizes.

researcher<-rep(1:nres,nstud)
totnumstud<-sum(nstud)
effsiz<-numeric(totnumstud)
study<-1:totnumstud
sampsiz<-effsiz
for(i in 1:nres)sampsiz[researcher==i]<-nnn[[i]]
```

```
wt<-sampsiz-3
# For fixed effect case:
alpha<-.05
Calpha<-qnorm(alpha/2,lower.tail=FALSE)

# nrep is the number of simulation repetitions
nrep<-1000
allquant<-matrix(0,nrep,8+nres)
dimnames(allquant)[[2]]<-c("Tdot","Qval","pval","CIFE.L","CIFE.U",
"CIRE.L","CIRE.H","Vefsiz",paste("Q",1:nres,sep=""))

researcher.results<-list()
for(i in 1:nres){
if(nstud[i]==1){
researcher.results[[i]]<-matrix(0,nrep,1)
dimnames(researcher.results[[i]])[[2]]<-list("Tdot")
}
else{
researcher.results[[i]]<-matrix(0,nrep,8)
dimnames(researcher.results[[i]])[[2]]<-c("Tdot","Qval","pval","CIFE.L",
"CIFE.U","CIRE.L","CIRE.H","Vefsiz")
}
}

for(irep in 1:nrep){

for(i in 1:nres)
effsiz[researcher==i]<-overall+equicorr(sig1,rho1,nstud[i])+
                   equicorrsamperr(rho2,nstud[i],nnn[[i]])

allquant[irep,]<-allquantfn(effsiz,wt,researcher,Calpha)

for(i in 1:nres){
s<-(researcher==i)
researcher.results[[i]][irep,]<-
allquantfn(effsiz[s],wt[s],researcher[s],Calpha)
}
}

# allquant and researcher.results contain the accumulated results.

save.image(file="researcher_1.Rdata")

# If nrep is large, comment out the following lines to reduce output.

#allquant
```

```
Options (width = 130)

#researcher.results
# Compute means and variances (over the nrep repetitions) of all quantities.
# Only a few of these are of interest.

apply(allquant,2,mean)

apply(allquant,2,var)

# Do the same for each researcher.  (Delete this if it is of no interest.)

for(i in 1:nres){

cat(paste("\n\nFor Researcher",i,":\n"))

cat(paste("\nMeans:\n"))
print(apply(researcher.results[[i]],2,mean))

cat(paste("\nVariances:\n"))
print(apply(researcher.results[[i]],2,var))
}

# Confidence interval coverages.
# Coverage of fixed effect confidence intervals.

L<-allquant[,4]
U<-allquant[,5]
cover.FE<-mean((L<overall)&(overall<U))

# Coverage of random effect confidence intervals.

L<-allquant[,6]
U<-allquant[,7]
cover.RE<-mean((L<overall)&(overall<U))

overall

cover.FE

cover.RE

# You can also report the confidence interval coverages
#      separately for each researcher, if that is of interest.
```

APPENDIX J

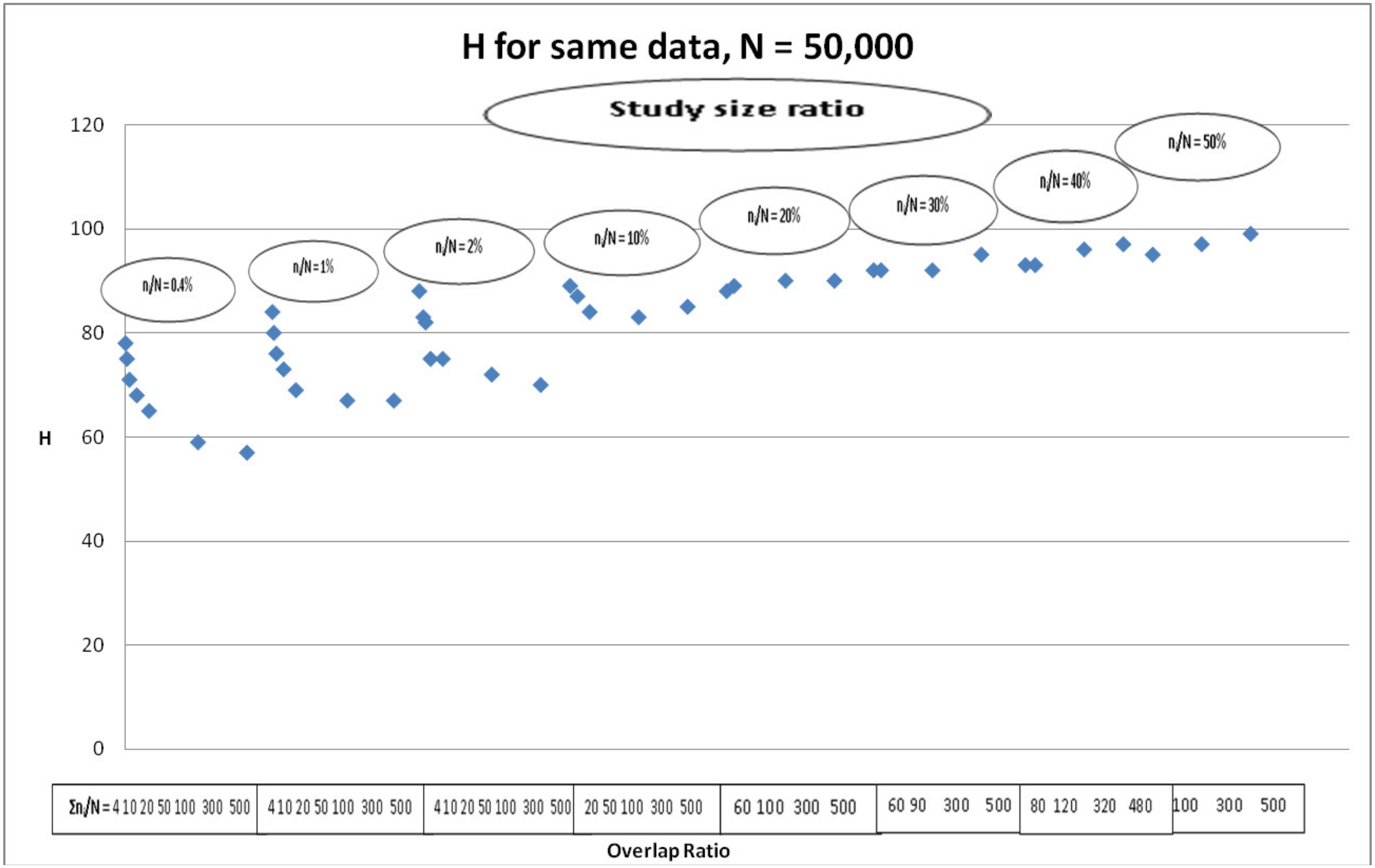
SAME DATA SIMULATION RESULTS FOR 50,000 DATA SETS

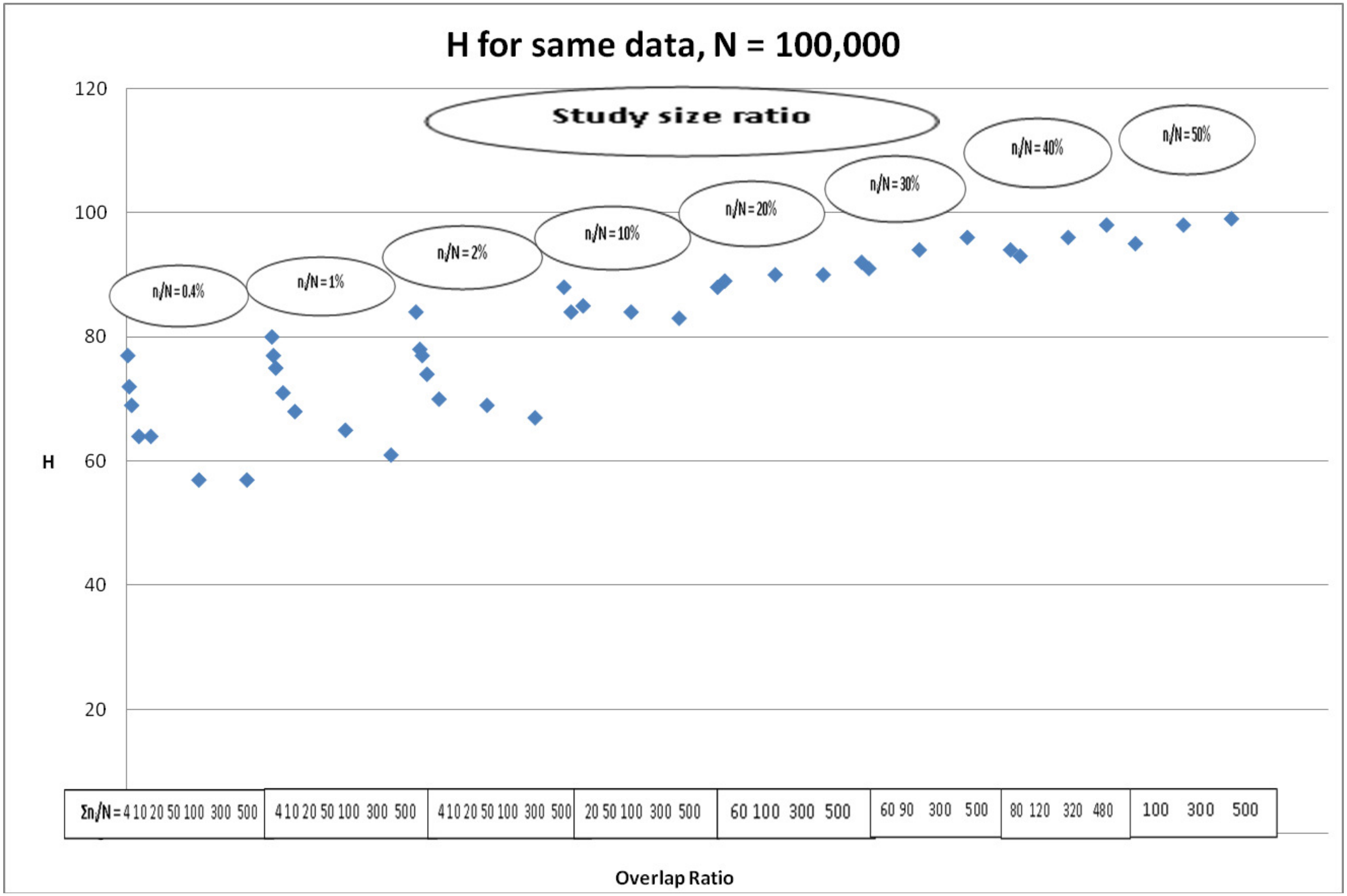
H300	50000	6	25000	3	0.2	150000
H500	50000	10	25000	5	0.2	250000
SAME DATA	Initial data set	Number of studies	Sample size	Ratio	Effect size	# of sample
A4	50000	10	200	0.04	0.2	2000
A10	50000	25	200	0.1	0.2	5000
A20	50000	50	200	0.2	0.2	10000
A50	50000	125	200	0.5	0.2	25000
A100	50000	250	200	1	0.2	50000
A300	50000	750	200	3	0.2	150000
A500	50000	1250	200	5	0.2	250000
B4	50000	4	500	0.04	0.2	2000
B10	50000	10	500	0.1	0.2	5000
B20	50000	20	500	0.2	0.2	10000
B50	50000	50	500	0.5	0.2	25000
B100	50000	100	500	1	0.2	50000
B300	50000	300	500	3	0.2	150000
B500	50000	500	500	5	0.2	250000
C4	50000	2	1000	0.04	0.2	2000
C10	50000	5	1000	0.1	0.2	5000
C20	50000	10	1000	0.2	0.2	10000
C50	50000	25	1000	0.5	0.2	25000
C100	50000	50	1000	1	0.2	50000
C300	50000	150	1000	3	0.2	150000
C500	50000	250	1000	5	0.2	250000
D20	50000	2	5000	0.2	0.2	10000
D50	50000	5	5000	0.5	0.2	25000
D100	50000	10	5000	1	0.2	50000
D300	50000	30	5000	3	0.2	150000
D500	50000	50	5000	5	0.2	250000
E40	50000	2	10000	0.4	0.2	20000
E60	50000	3	10000	0.6	0.2	30000
E100	50000	5	10000	1	0.2	50000
E300	50000	15	10000	3	0.2	150000
E500	50000	25	10000	5	0.2	250000
F60	50000	2	15000	0.6	0.2	30000
F90	50000	3	15000	0.9	0.2	45000
F300	50000	10	15000	3	0.2	150000
F500	50000	17	15000	5	0.2	255000
G80	50000	2	20000	0.8	0.2	40000
G120	50000	3	20000	1.2	0.2	60000
G320	50000	8	20000	3.2	0.2	160000
G480	50000	12	20000	4.8	0.2	240000
H100	50000	2	25000	1	0.2	50000

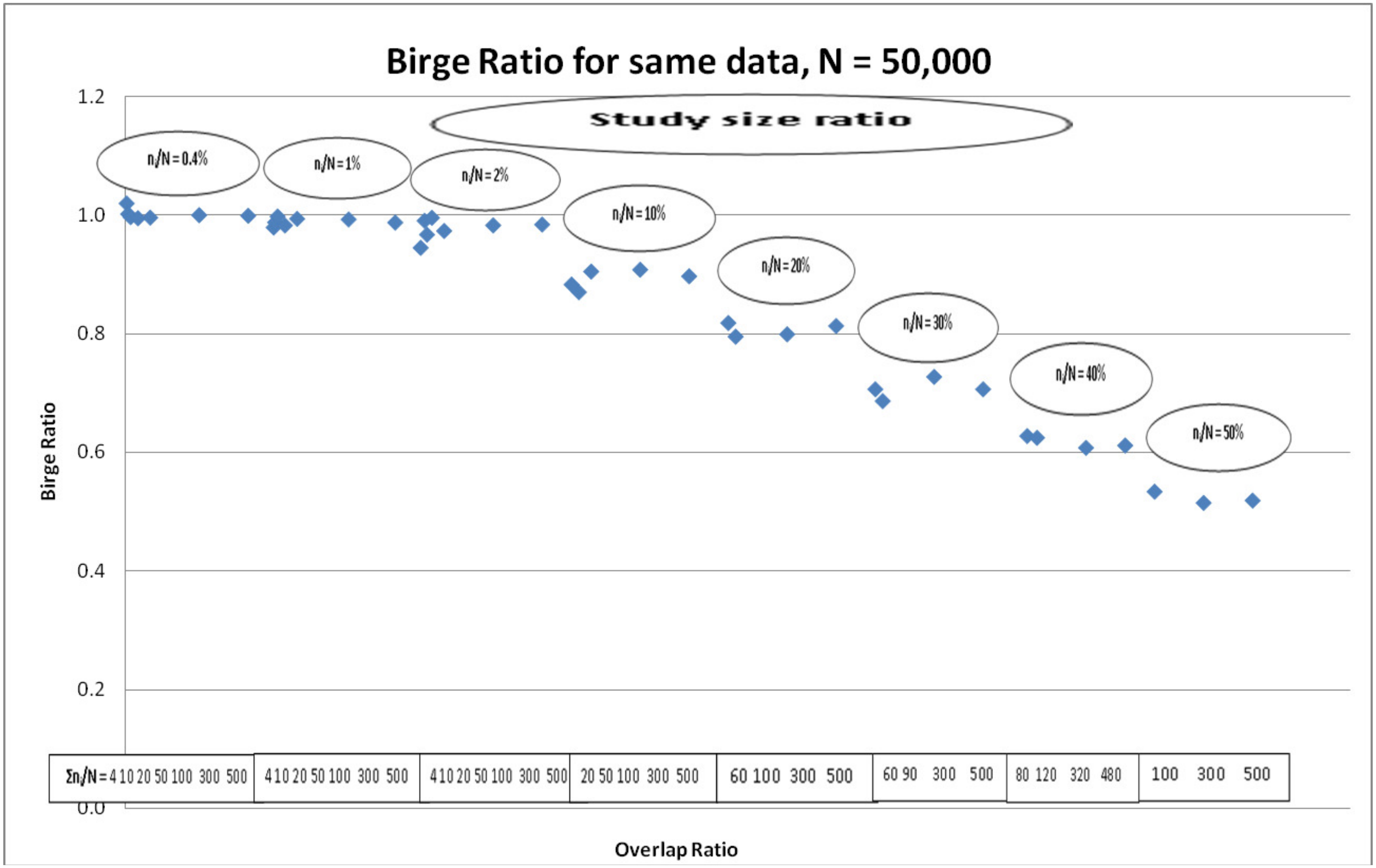
APPENDIX J-2

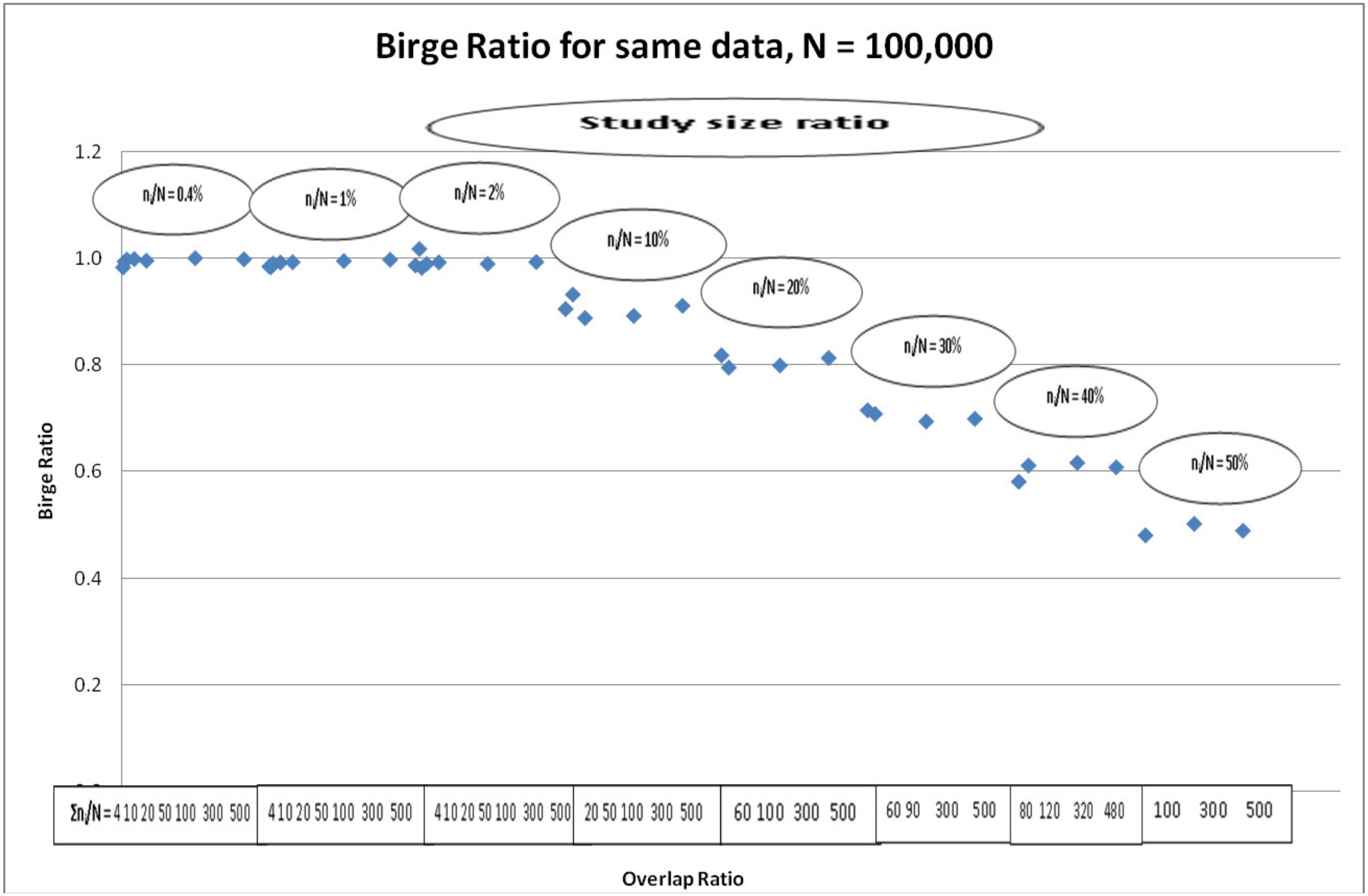
SAME DATA SIMULATION RESULTS FOR 100,000 DATA SETS

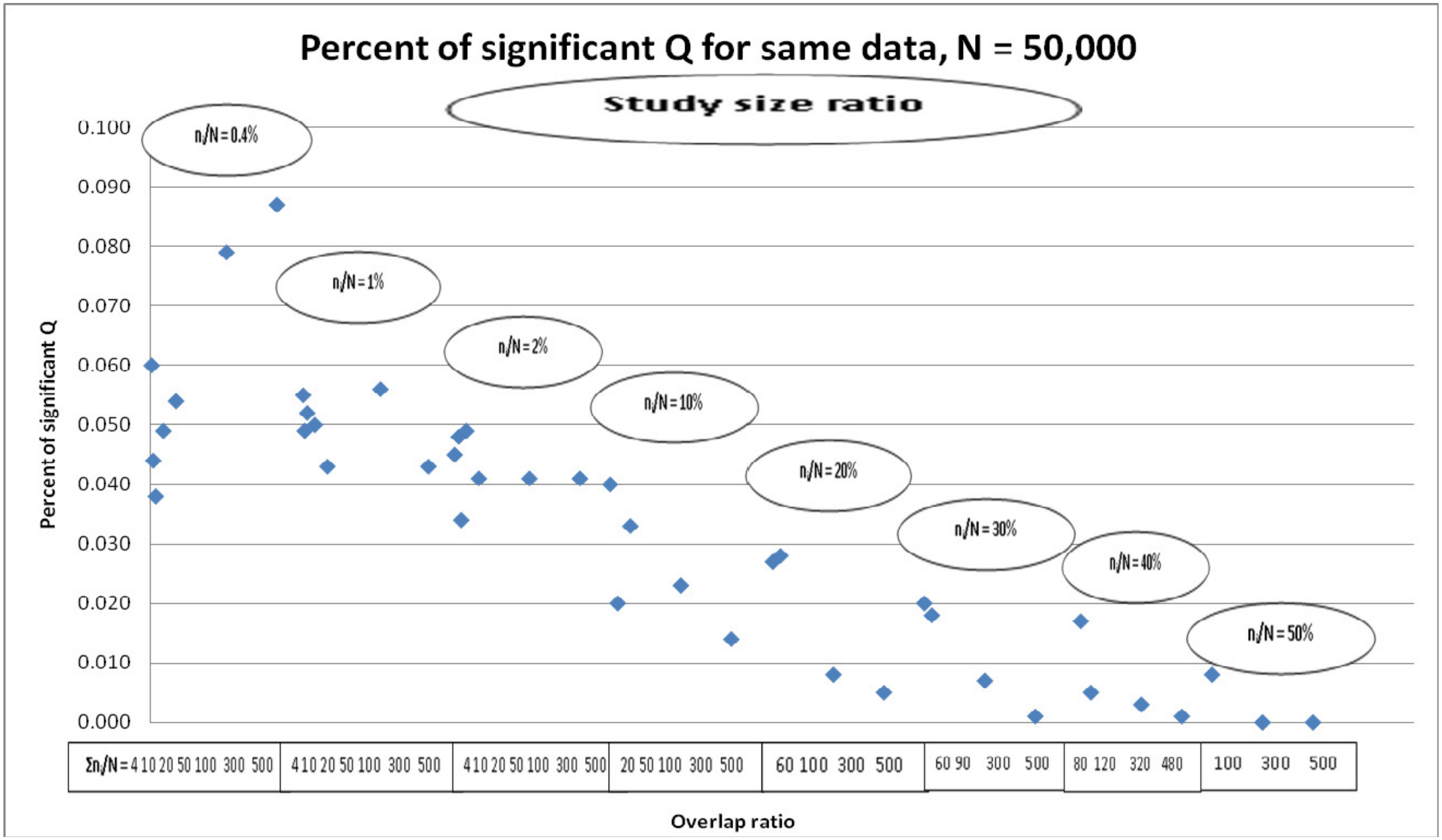
SAME DATA	Initial data set	Number of studies	Sample size	Ratio	Effect size	# of sample
A4	100000	20	200	0.04	0.2	4000
A10	100000	50	200	0.1	0.2	10000
A20	100000	100	200	0.2	0.2	20000
A50	100000	250	200	0.5	0.2	50000
A100	100000	500	200	1	0.2	100000
A300	100000	1500	200	3	0.2	300000
A500	100000	2500	200	5	0.2	500000
B4	100000	8	500	0.04	0.2	4000
B10	100000	20	500	0.1	0.2	10000
B20	100000	40	500	0.2	0.2	20000
B50	100000	100	500	0.5	0.2	50000
B100	100000	200	500	1	0.2	100000
B300	100000	600	500	3	0.2	300000
B500	100000	1000	500	5	0.2	500000
C4	100000	4	1000	0.04	0.2	4000
C10	100000	10	1000	0.1	0.2	10000
C20	100000	20	1000	0.2	0.2	20000
C50	100000	50	1000	0.5	0.2	50000
C100	100000	100	1000	1	0.2	100000
C300	100000	300	1000	3	0.2	300000
C500	100000	500	1000	5	0.2	500000
D20	100000	2	10000	0.2	0.2	20000
D50	100000	5	10000	0.5	0.2	50000
D100	100000	10	10000	1	0.2	100000
D300	100000	30	10000	3	0.2	300000
D500	100000	50	10000	5	0.2	500000
E60	100000	3	20000	0.6	0.2	60000
E100	100000	5	20000	1	0.2	100000
E300	100000	15	20000	3	0.2	300000
E500	100000	25	20000	5	0.2	500000
F60	100000	2	30000	0.6	0.2	60000
F90	100000	3	30000	0.9	0.2	90000
F300	100000	10	30000	3	0.2	300000
F500	100000	17	30000	5	0.2	510000
G80	100000	2	40000	0.8	0.2	80000
G120	100000	3	40000	1.2	0.2	120000
G320	100000	8	40000	3.2	0.2	320000
G480	100000	12	40000	4.8	0.2	480000
H100	100000	2	50000	1	0.2	100000
H300	100000	6	50000	3	0.2	300000
H500	100000	10	50000	5	0.2	500000


H for same data, N = 50,000
Study size ratio
120
100
80
60
40
20
0
H
n/N = 0.4%
n/N = 1%
n/N = 2%
n/N = 10%
n/N = 20%
n/N = 30%
n/N = 40%
n/N = 50%
Σn/N = 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap Ratio

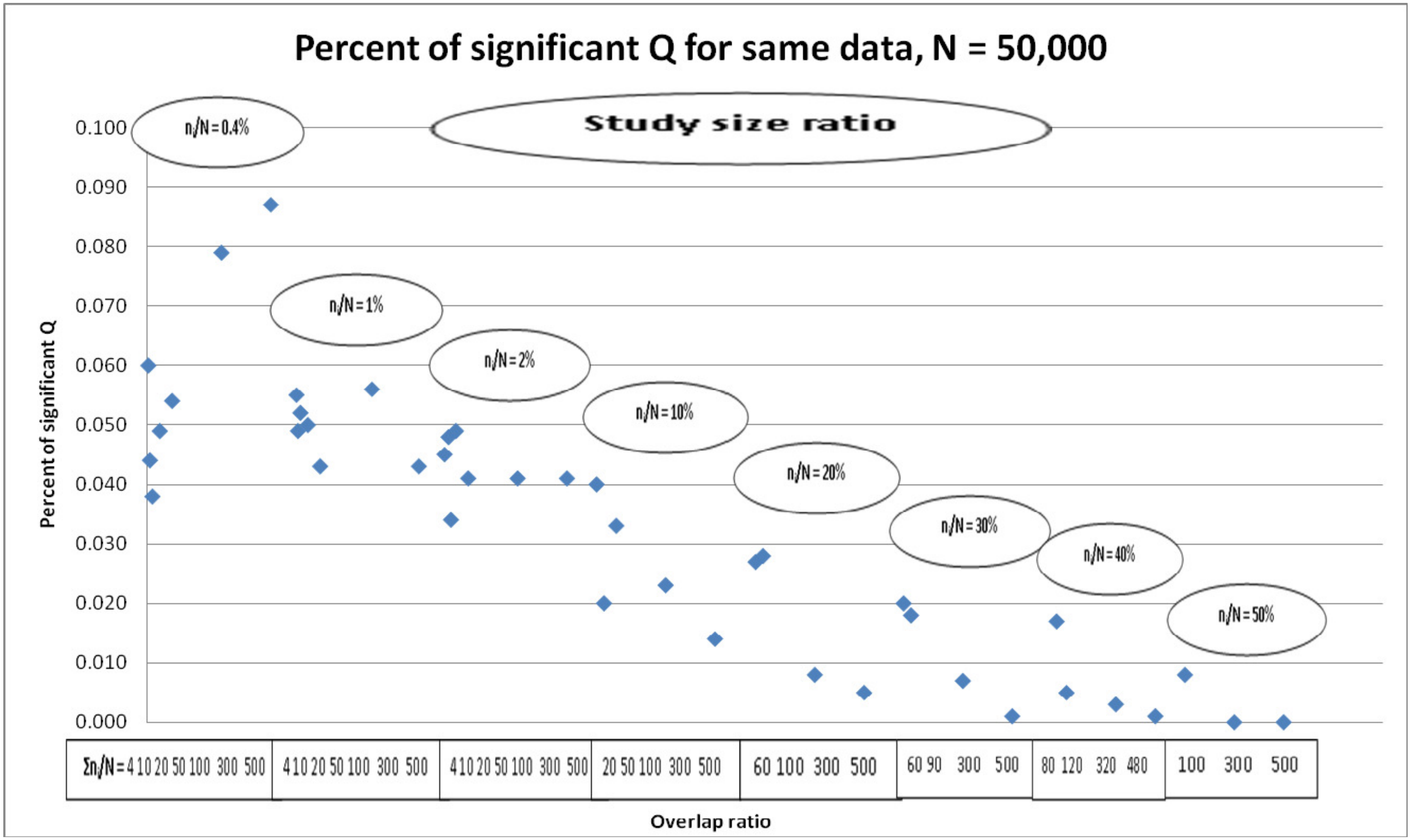
H for same data, N = 100,000
Study size ratio
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
H
120
100
80
60
40
20
$\Sigma n_i/N$ = 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap Ratio


Birge Ratio for same data, N = 50,000
Study size ratio
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Birge Ratio
$\Sigma n_i/N$ = 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap Ratio

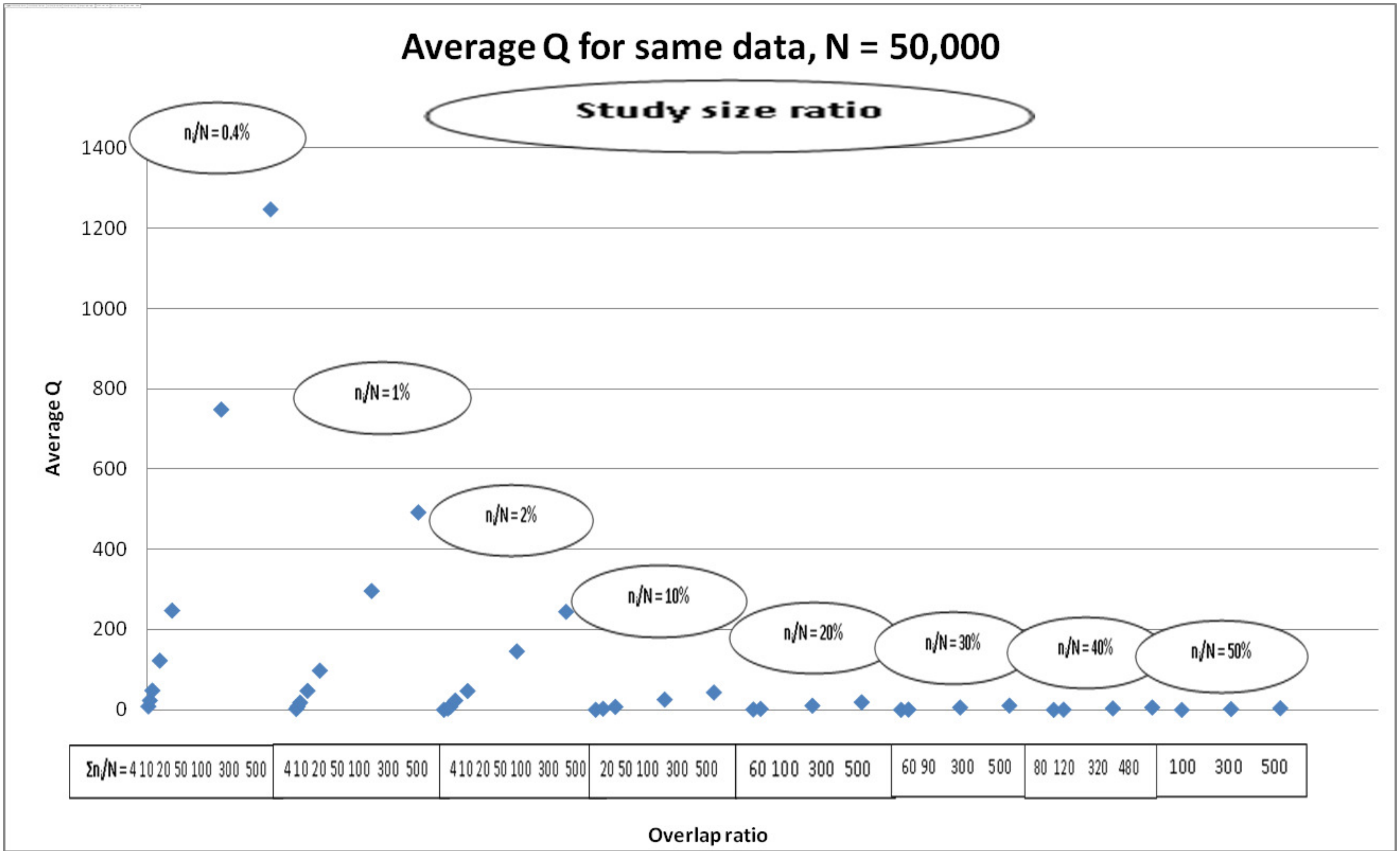

Birge Ratio for same data, N = 100,000
Study size ratio
1.2
1.0
0.8
0.6
0.4
0.2
Birge Ratio
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
$\Sigma n_i/N =$ 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap Ratio

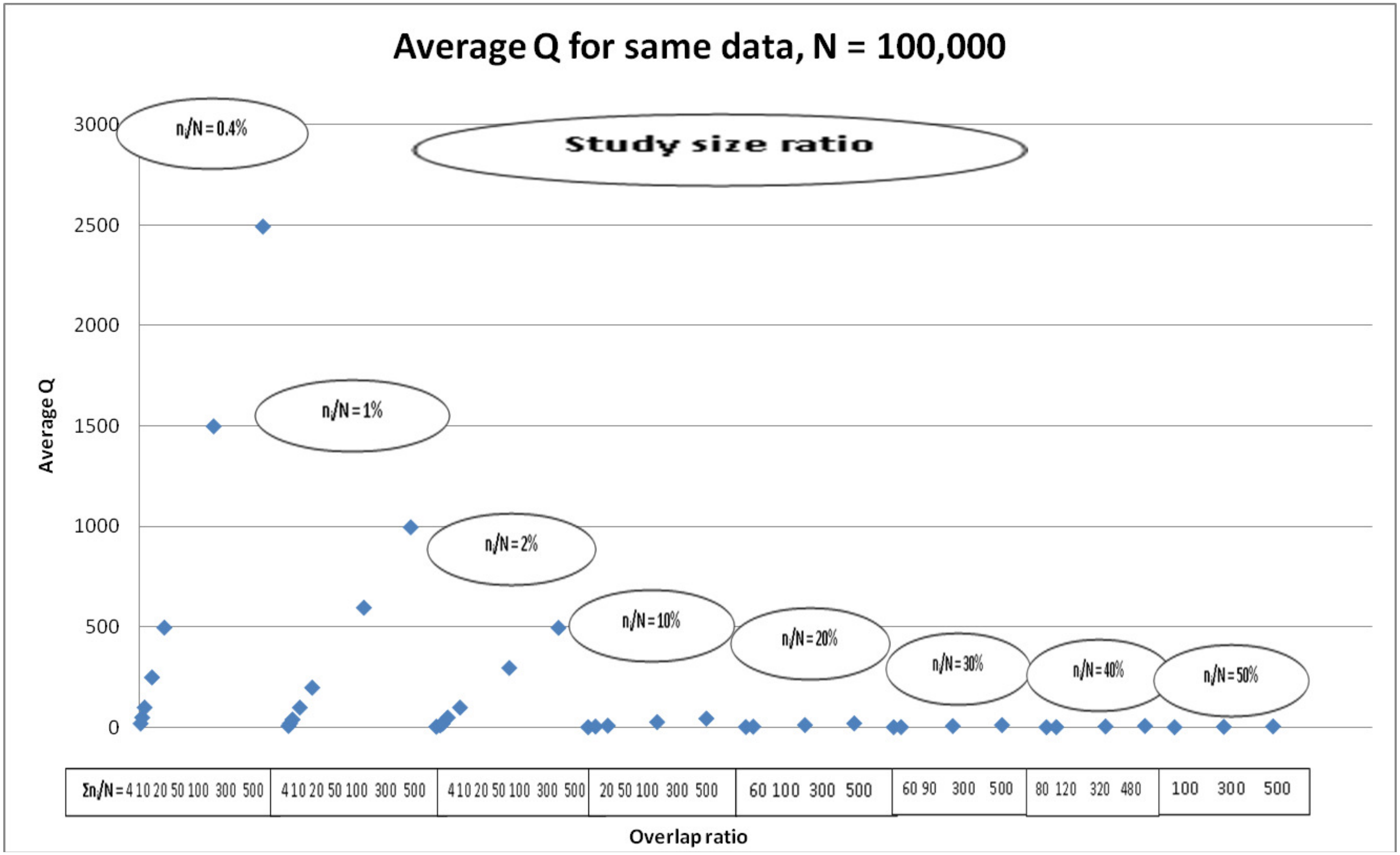


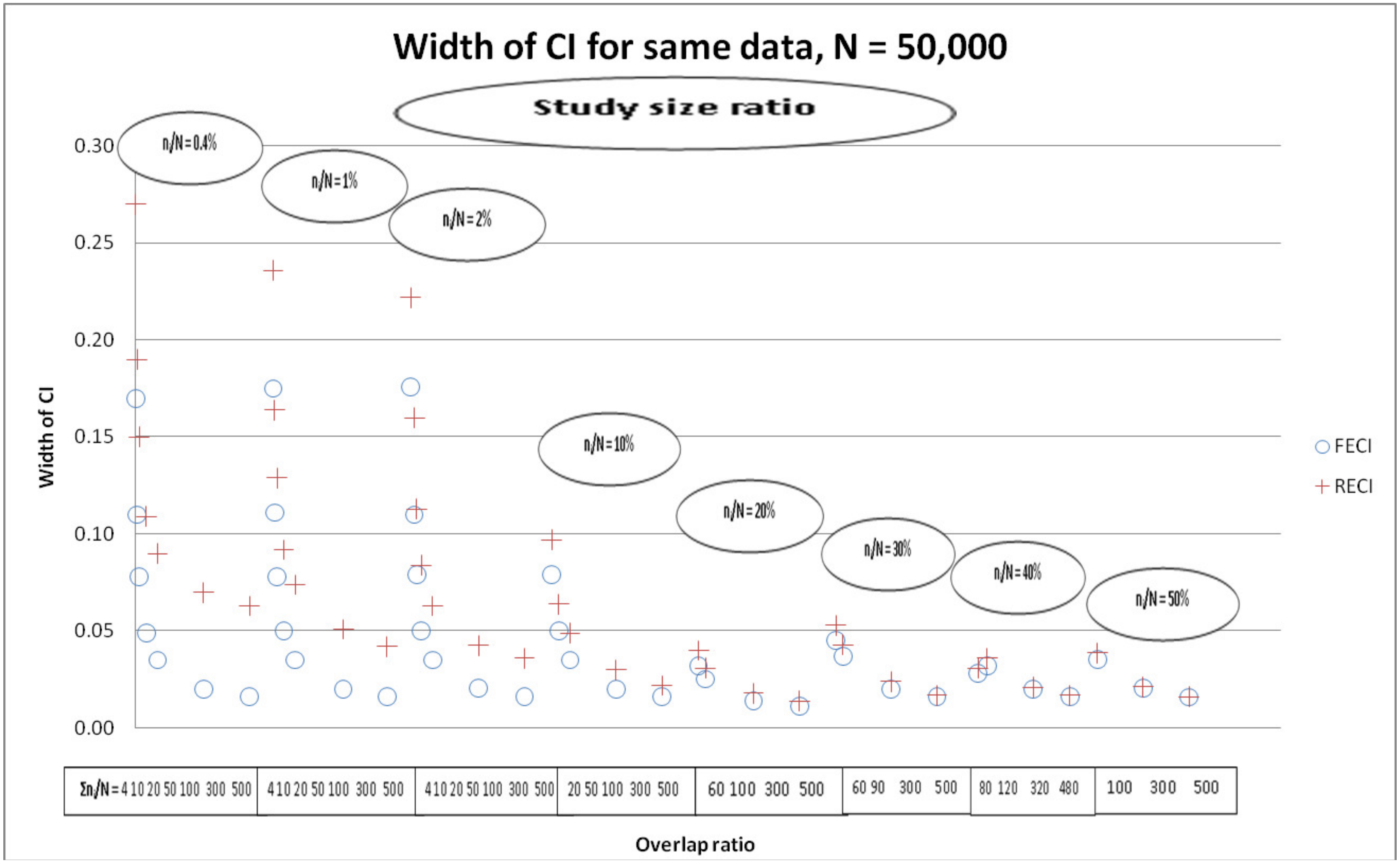
* In the Y-axis, 0.01 to 0.1 represent 1% to 10%.

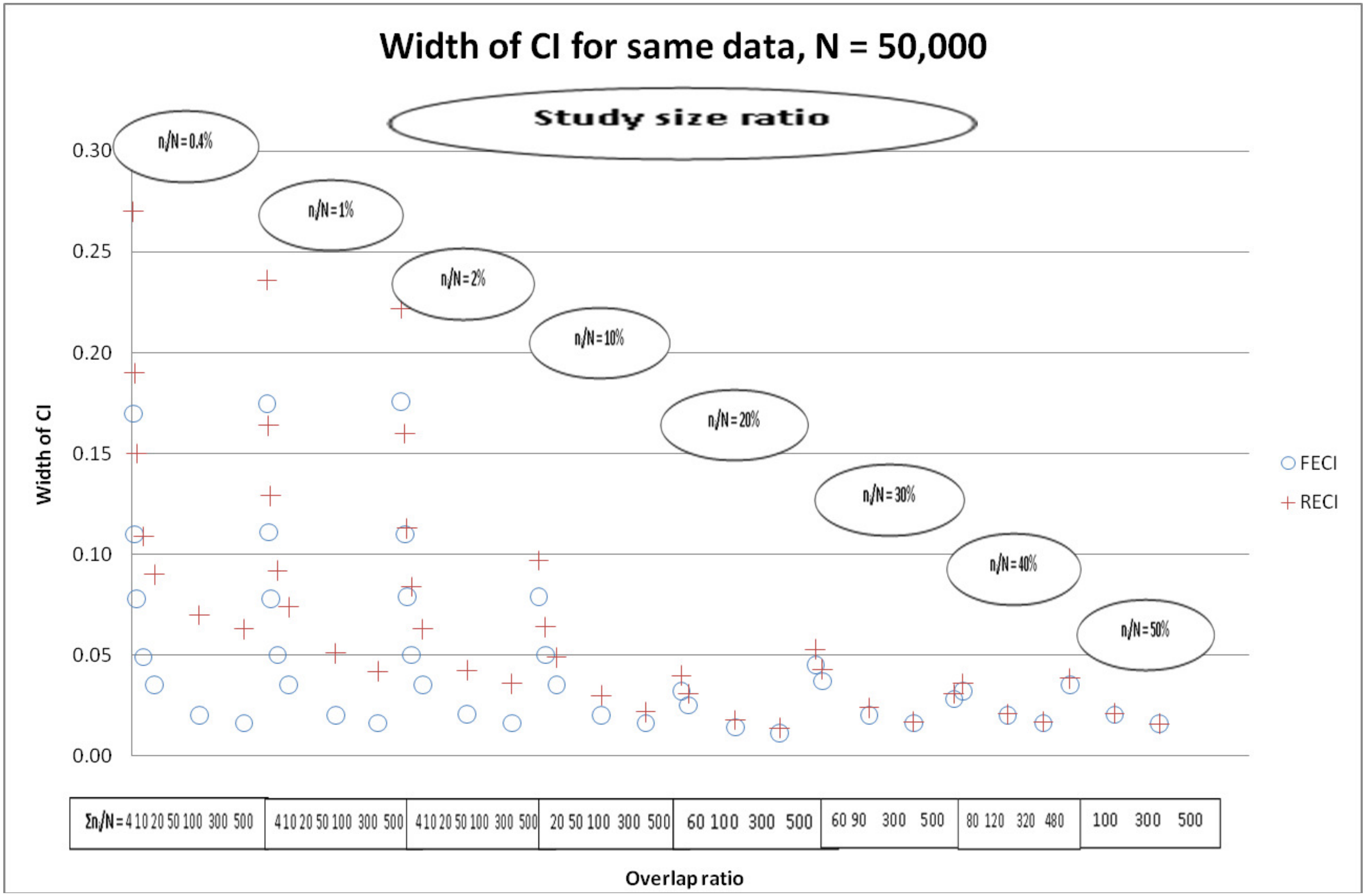


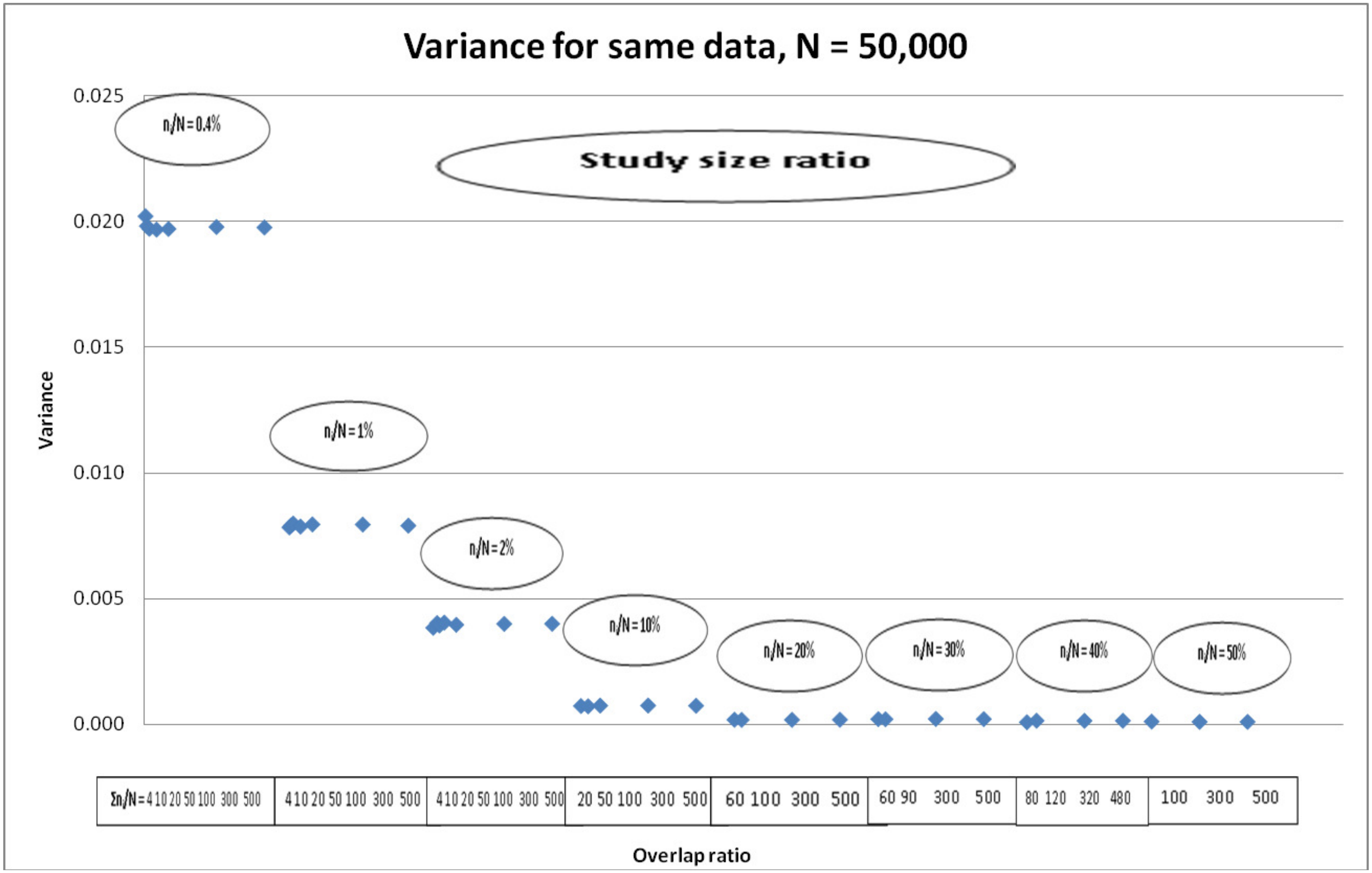
* In the Y-axis, 0.01 to 0.1 represent 1% to 10%.

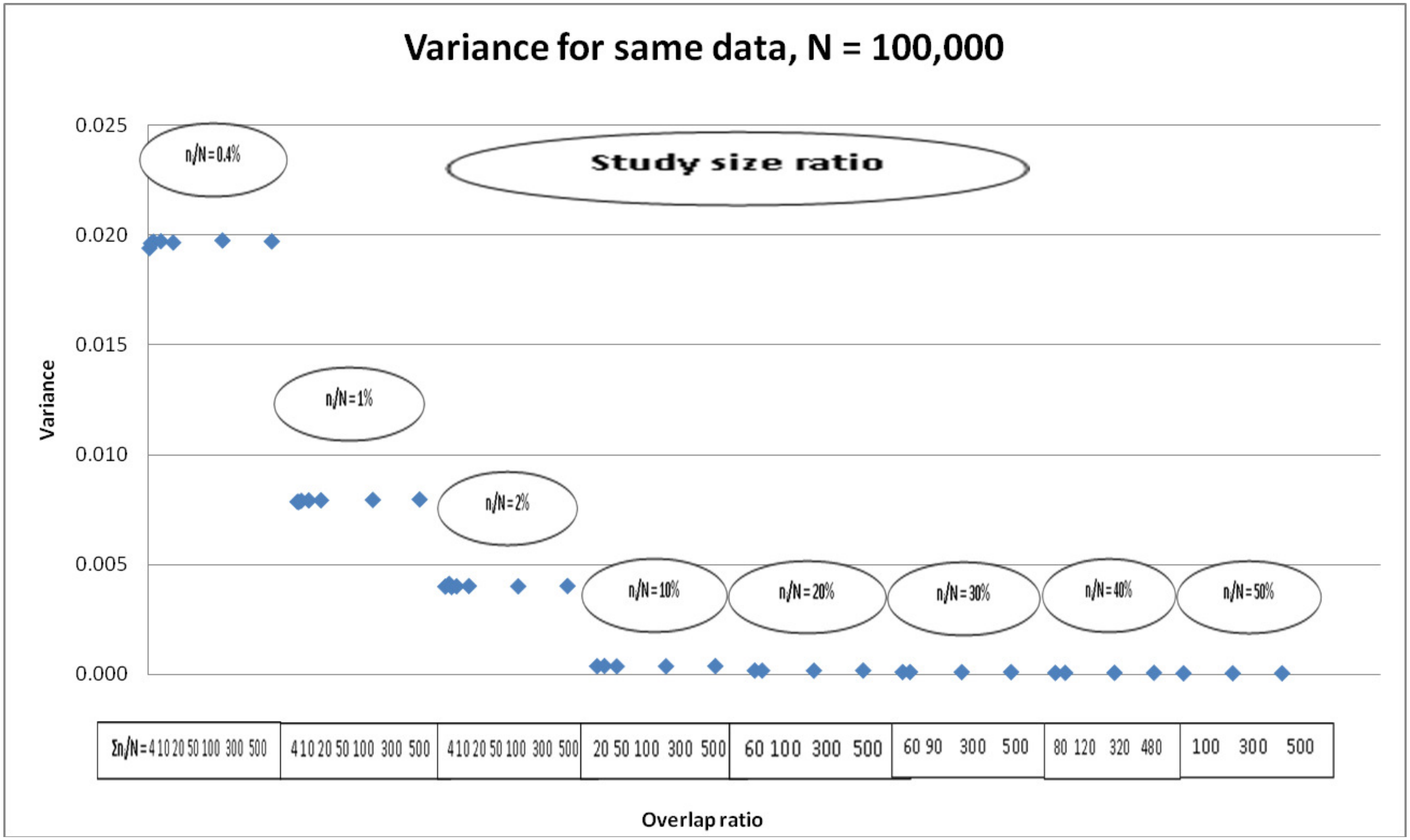

Average Q for same data, N = 50,000
Study size ratio
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
Average Q
1400
1200
1000
800
600
400
200
0
$\Sigma n_i/N = 4$ 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio


Average Q for same data, N = 100,000
Study size ratio
Average Q
3000
2500
2000
1500
1000
500
0
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
$\Sigma n_i/N =$ 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio


Width of CI for same data, N = 50,000
Study size ratio
$n_i/N=0.4\%$
$n_i/N=1\%$
$n_i/N=2\%$
$n_i/N=10\%$
$n_i/N=20\%$
$n_i/N=30\%$
$n_i/N=40\%$
$n_i/N=50\%$
Width of CI
0.30
0.25
0.20
0.15
0.10
0.05
0.00
○ FECI
+ RECI
$\Sigma n_i/N$ = 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio


Width of CI for same data, N = 50,000
Study size ratio
n/N=0.4%
n/N=1%
n/N=2%
n/N=10%
n/N=20%
n/N=30%
n/N=40%
n/N=50%
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Width of CI
○ FECI
+ RECI
Σn/N = 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio


Variance for same data, N = 50,000
Study size ratio
0.025
0.020
0.015
0.010
0.005
0.000
Variance
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
$\Sigma n_i/N =$ 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio


Variance for same data, N = 100,000
0.025
0.020
0.015
0.010
0.005
0.000
Variance
Study size ratio
$n_i/N = 0.4\%$
$n_i/N = 1\%$
$n_i/N = 2\%$
$n_i/N = 10\%$
$n_i/N = 20\%$
$n_i/N = 30\%$
$n_i/N = 40\%$
$n_i/N = 50\%$
$\Sigma n_i/N =$ 4 10 20 50 100 300 500
4 10 20 50 100 300 500
4 10 20 50 100 300 500
20 50 100 300 500
60 100 300 500
60 90 300 500
80 120 320 480
100 300 500
Overlap ratio

APPENDIX K

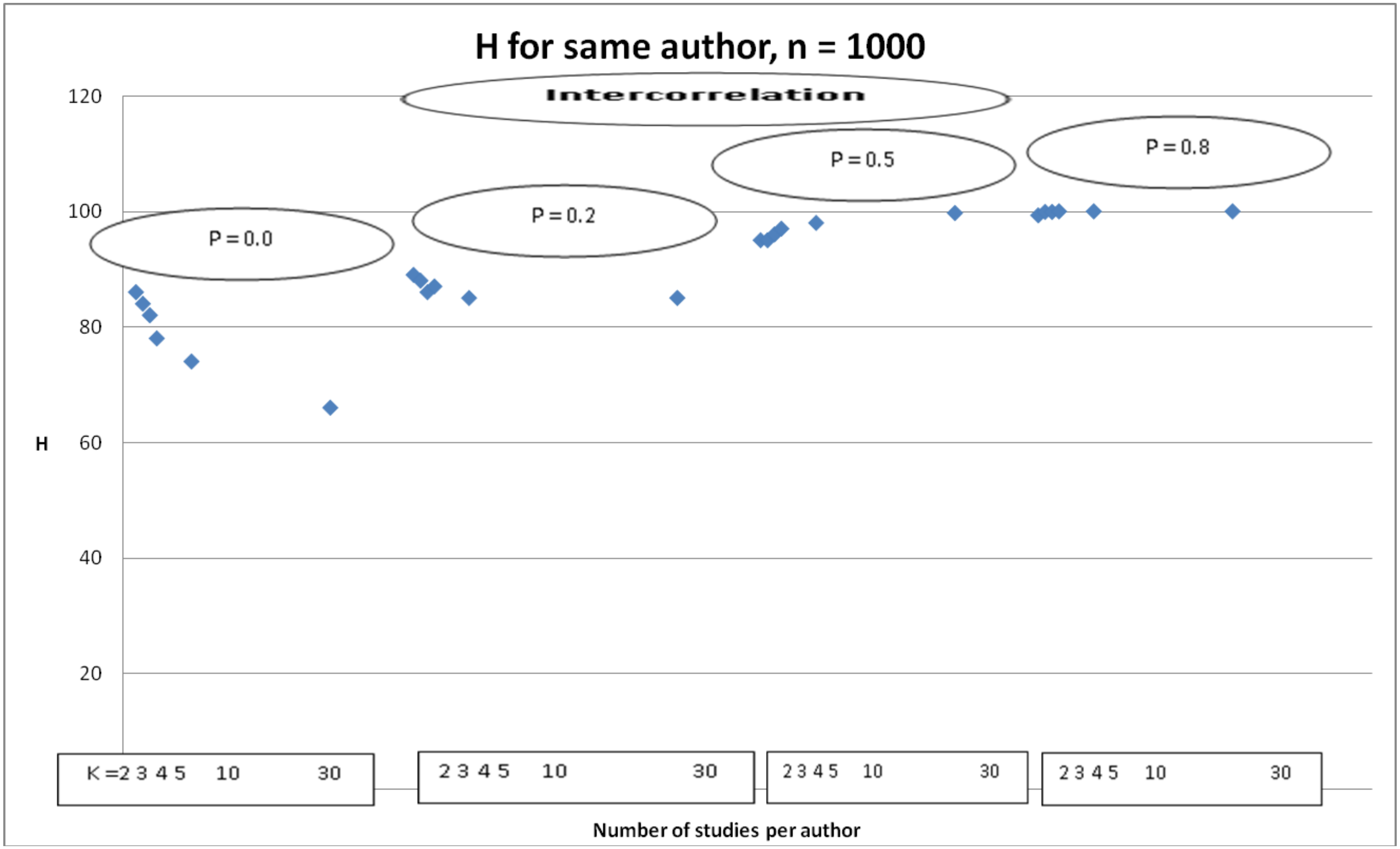
SAME AUTHOR SIMULATION RESULTS FOR 1,000 SAMPLE SIZE

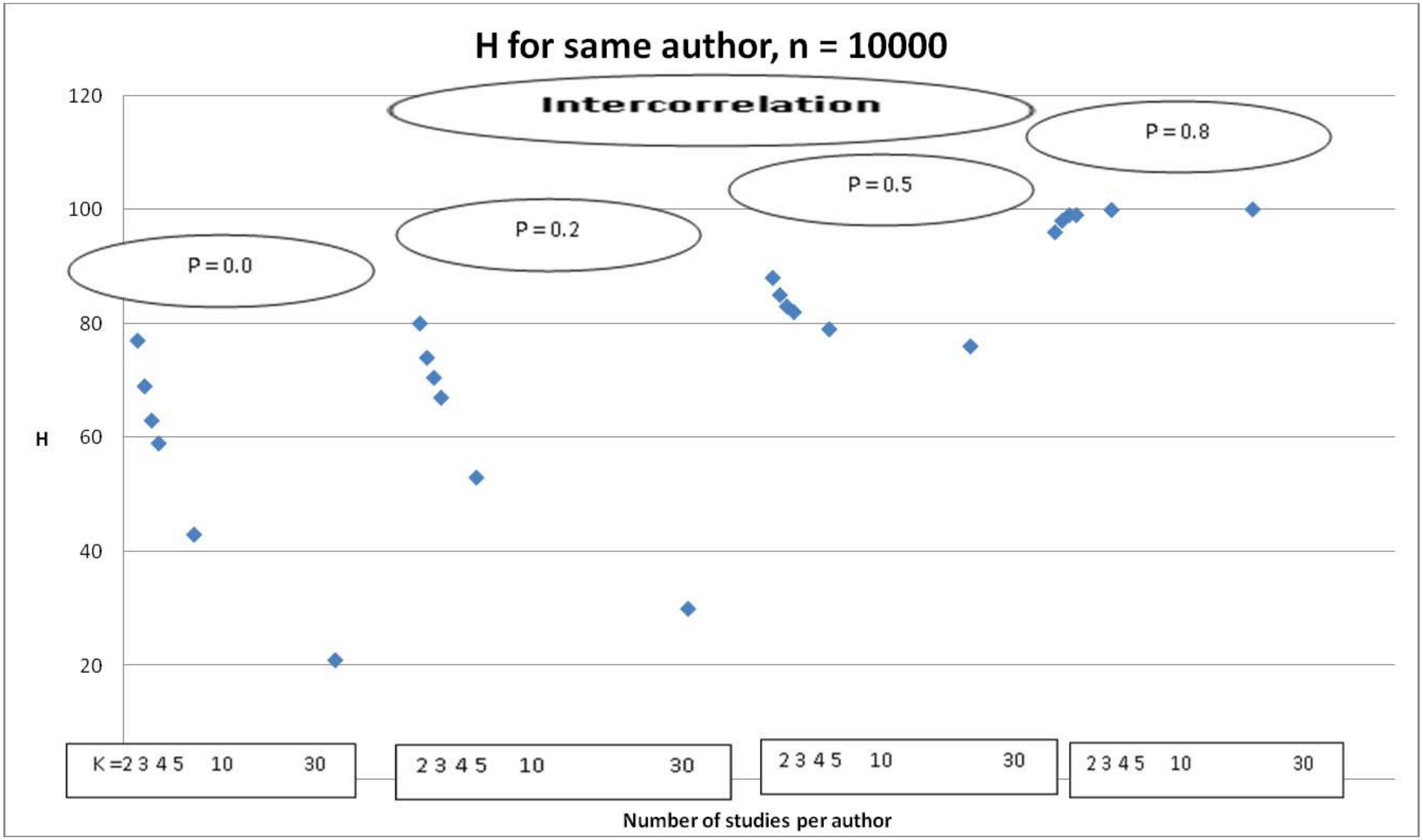
ID	Number of studies per author	Sample size (n_i)	Magnitude of ζ	Magnitude of correlation
*2**0	2	1000	0.2	0.0
30	3	1000	0.2	0.0
40	4	1000	0.2	0.0
50	5	1000	0.2	0.0
100	10	1000	0.2	0.0
300	30	1000	0.2	0.0
22	2	1000	0.2	0.2
32	3	1000	0.2	0.2
42	4	1000	0.2	0.2
52	5	1000	0.2	0.2
102	10	1000	0.2	0.2
302	30	1000	0.2	0.2
25	2	1000	0.2	0.5
35	3	1000	0.2	0.5
45	4	1000	0.2	0.5
55	5	1000	0.2	0.5
105	10	1000	0.2	0.5
305	30	1000	0.2	0.5
28	2	1000	0.2	0.8
38	3	1000	0.2	0.8
48	4	1000	0.2	0.8
58	5	1000	0.2	0.8
108	10	1000	0.2	0.8
308	30	1000	0.2	0.8

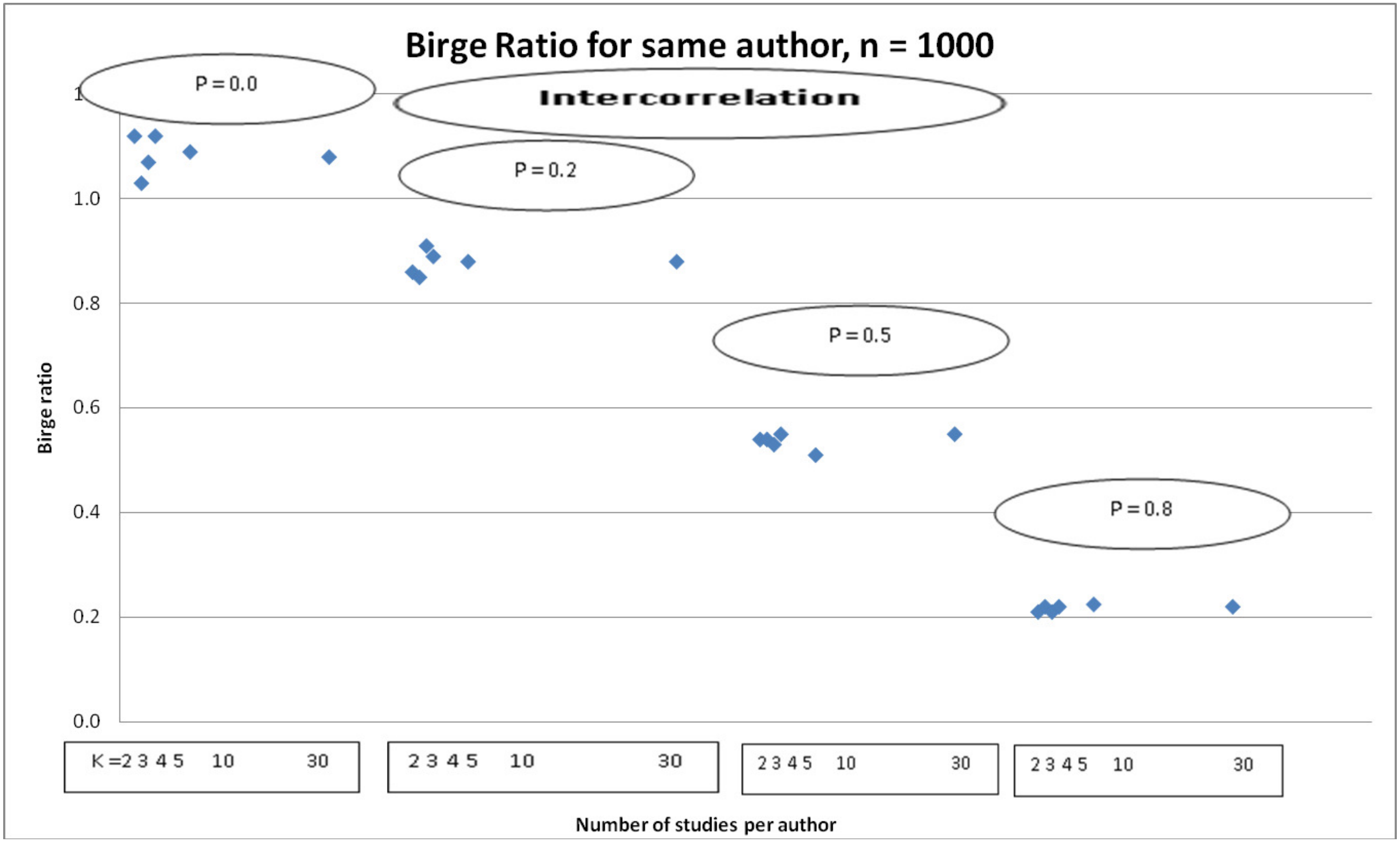
APPENDIX K-2

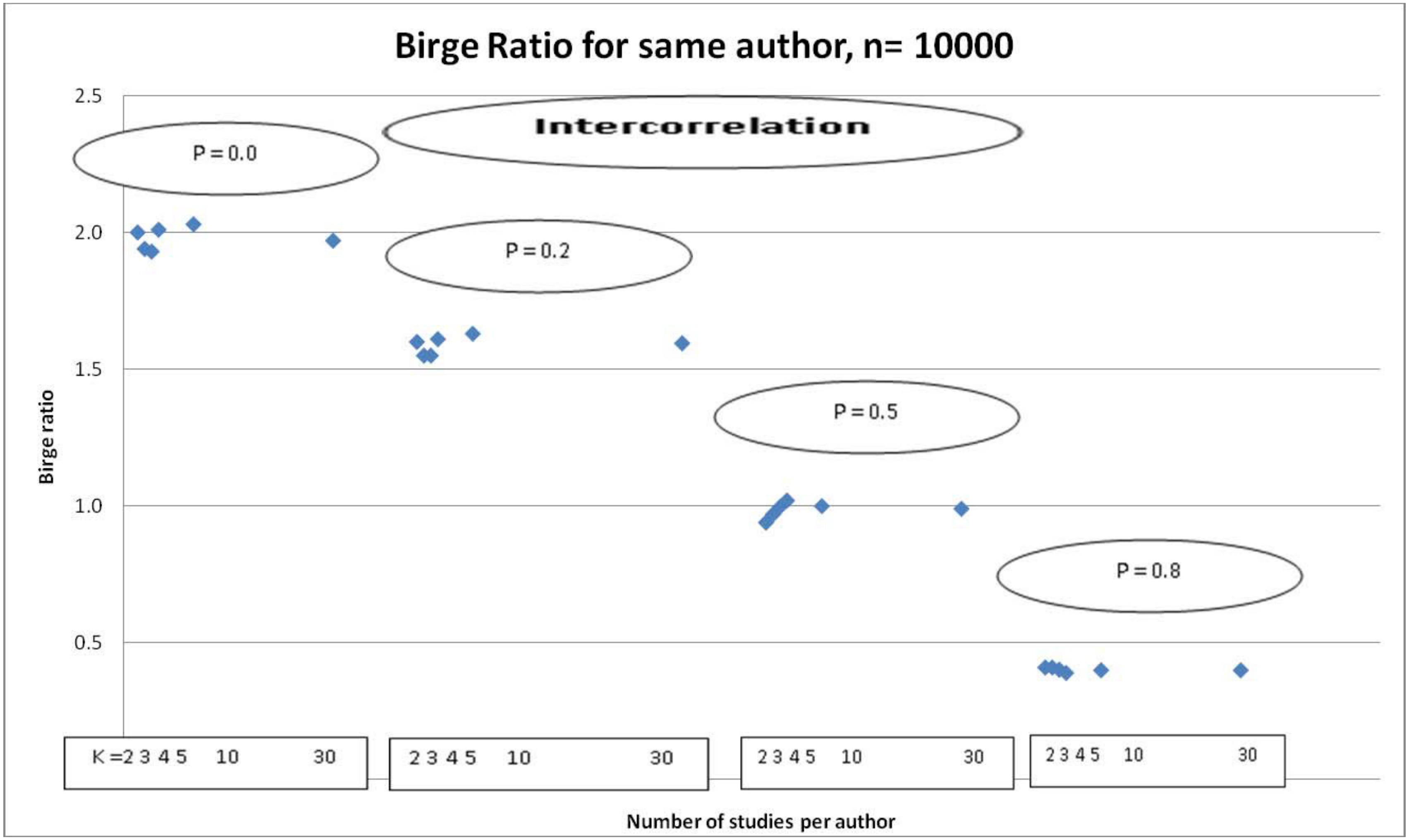
SAME AUTHOR SIMULATION RESULTS FOR 10,000 SAMPLE SIZE

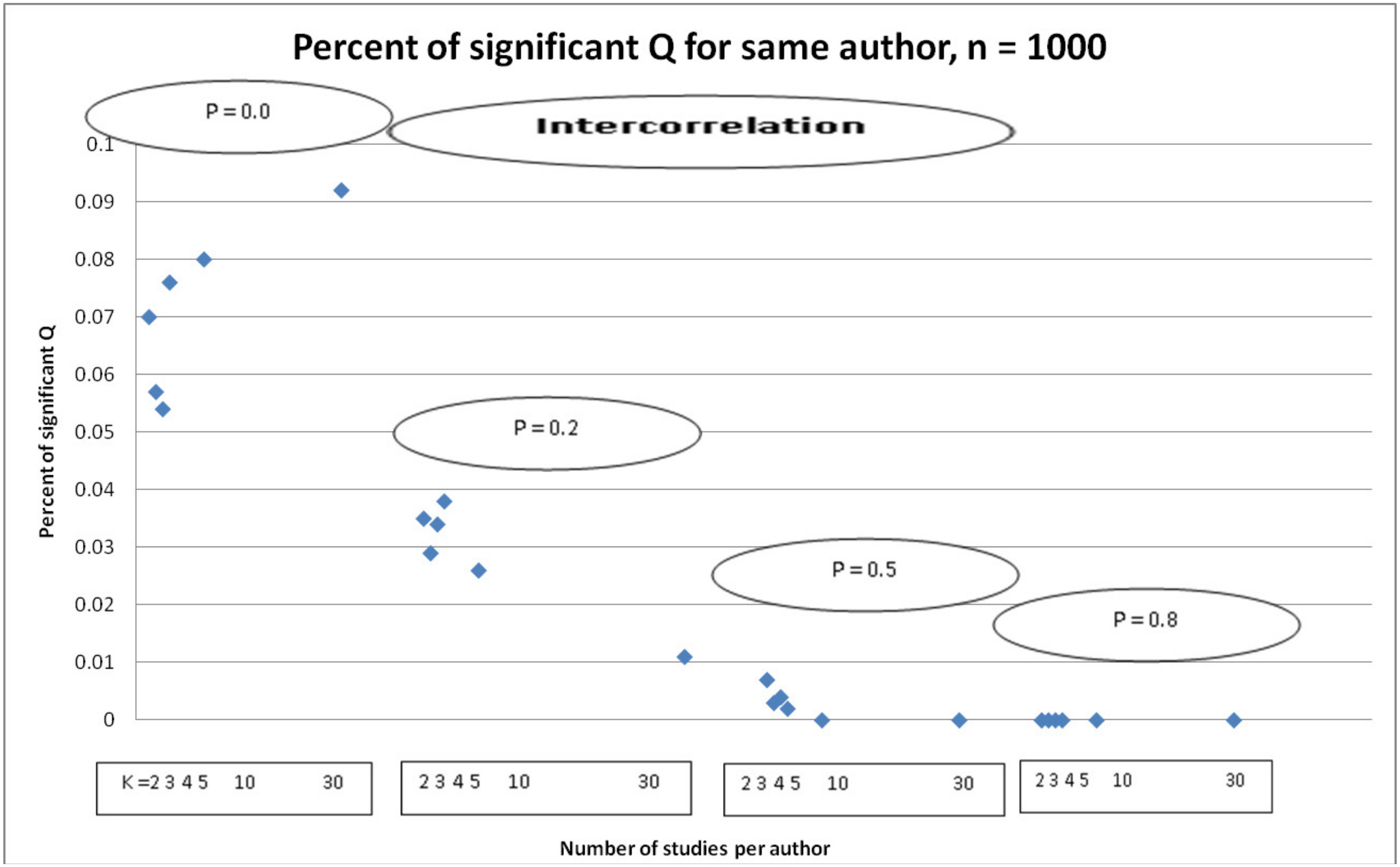
ID	Number of studies per author	Sample size (n_i)	Magnitude of ζ	Magnitude of correlation
*2**0	2	10000	0.2	0.0
30	3	10000	0.2	0.0
40	4	10000	0.2	0.0
50	5	10000	0.2	0.0
100	10	10000	0.2	0.0
300	30	10000	0.2	0.0
22	2	10000	0.2	0.2
32	3	10000	0.2	0.2
42	4	10000	0.2	0.2
52	5	10000	0.2	0.2
102	10	10000	0.2	0.2
302	30	10000	0.2	0.2
25	2	10000	0.2	0.5
35	3	10000	0.2	0.5
45	4	10000	0.2	0.5
55	5	10000	0.2	0.5
105	10	10000	0.2	0.5
305	30	10000	0.2	0.5
28	2	10000	0.2	0.8
38	3	10000	0.2	0.8
48	4	10000	0.2	0.8
58	5	10000	0.2	0.8
108	10	10000	0.2	0.8
308	30	10000	0.2	0.8


H for same author, n = 1000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
120
100
80
60
40
20
H
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author

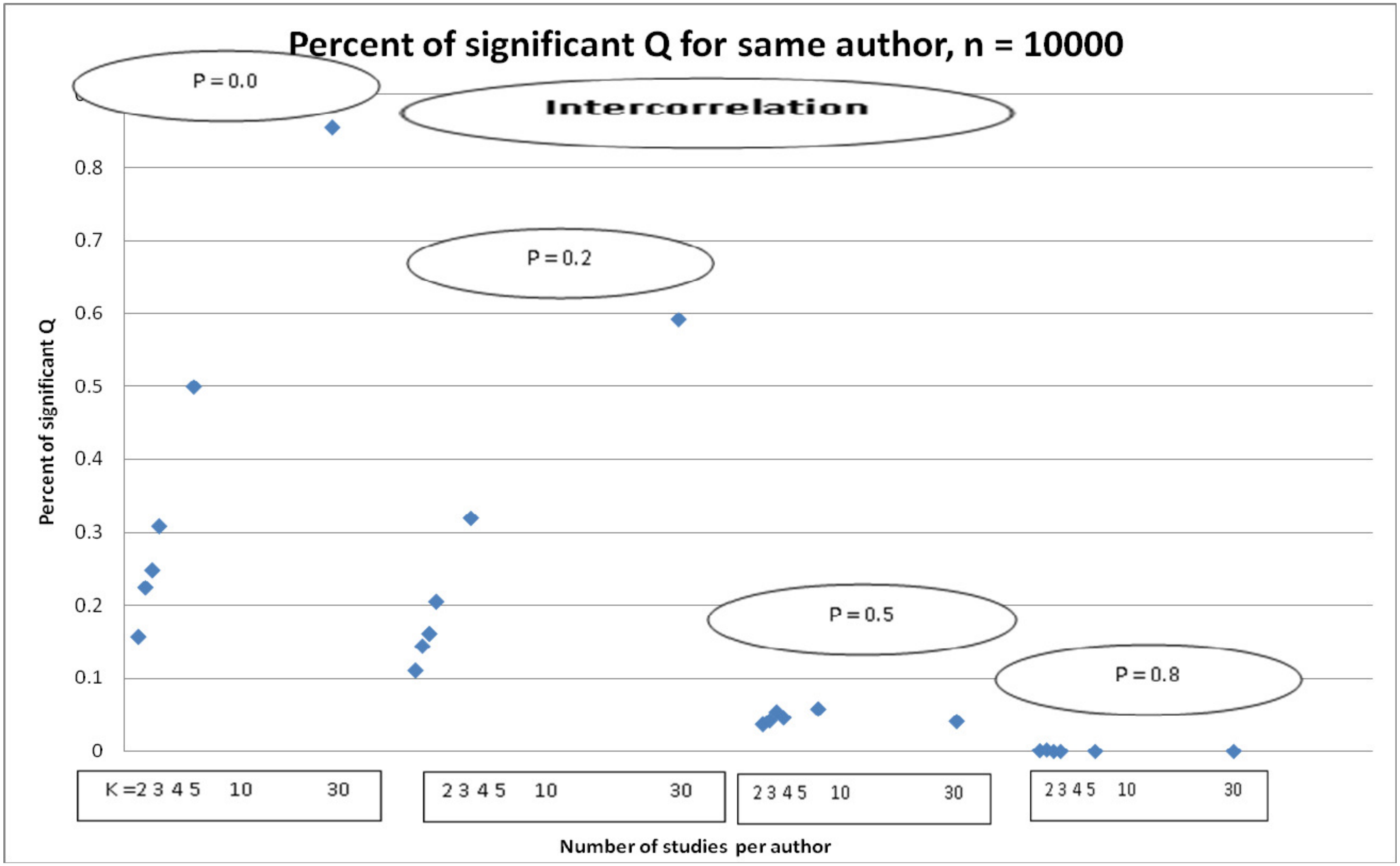

H for same author, n = 10000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
120
100
80
60
40
20
H
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author


Birge Ratio for same author, n = 1000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
1
1.0
0.8
0.6
0.4
0.2
0.0
Birge ratio
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author

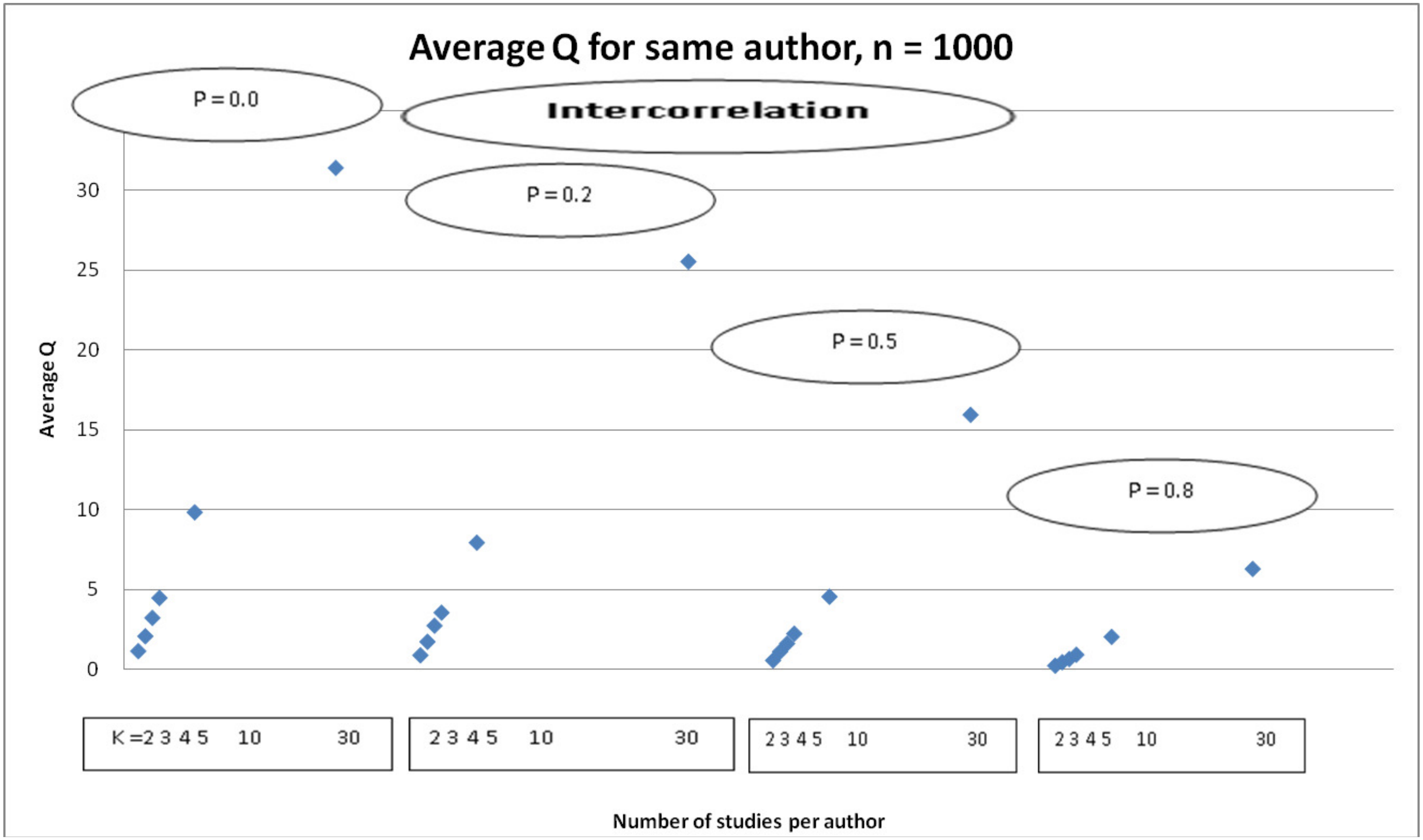

Birge Ratio for same author, n= 10000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
2.5
2.0
1.5
1.0
0.5
Birge ratio
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author

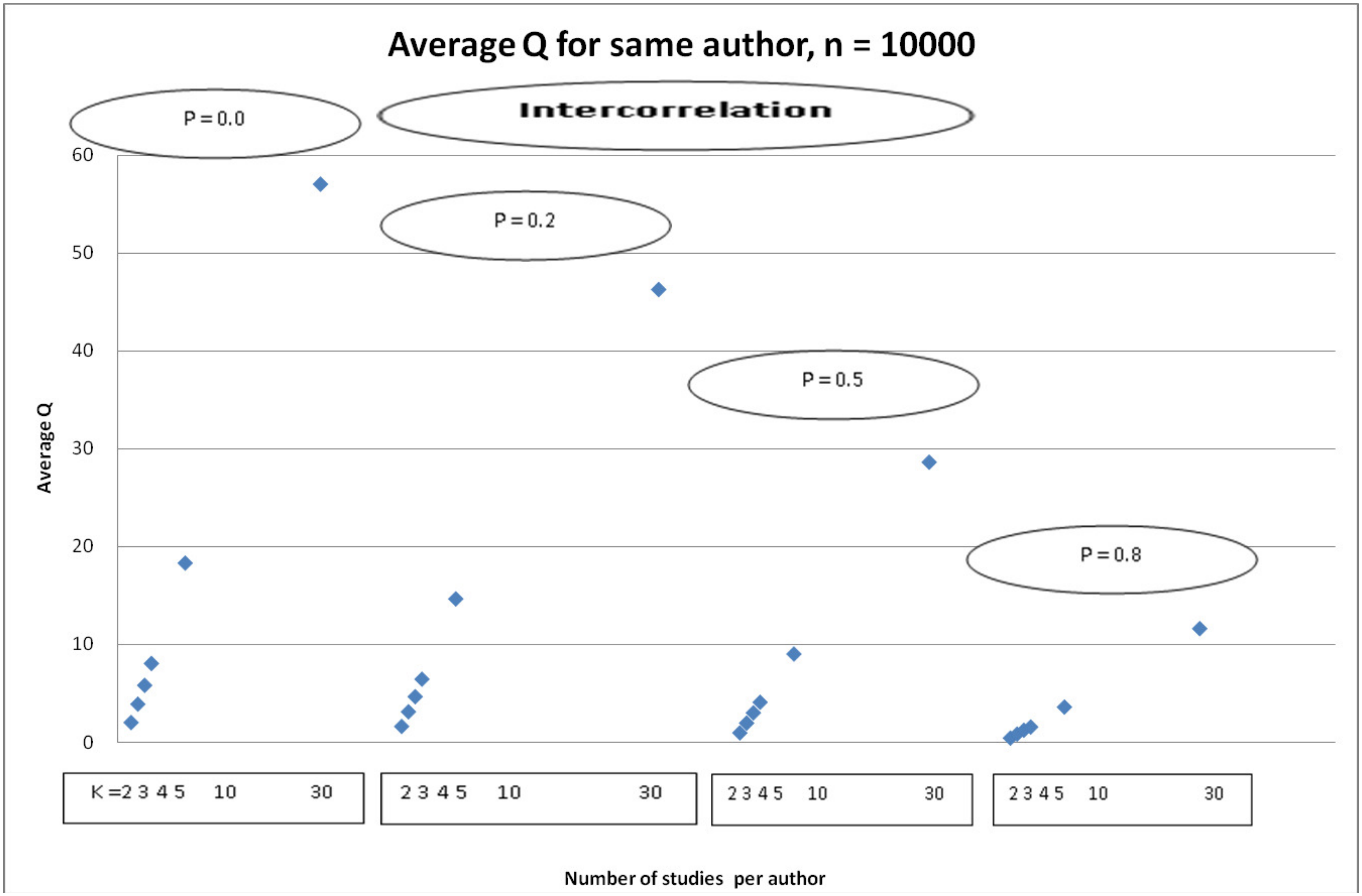


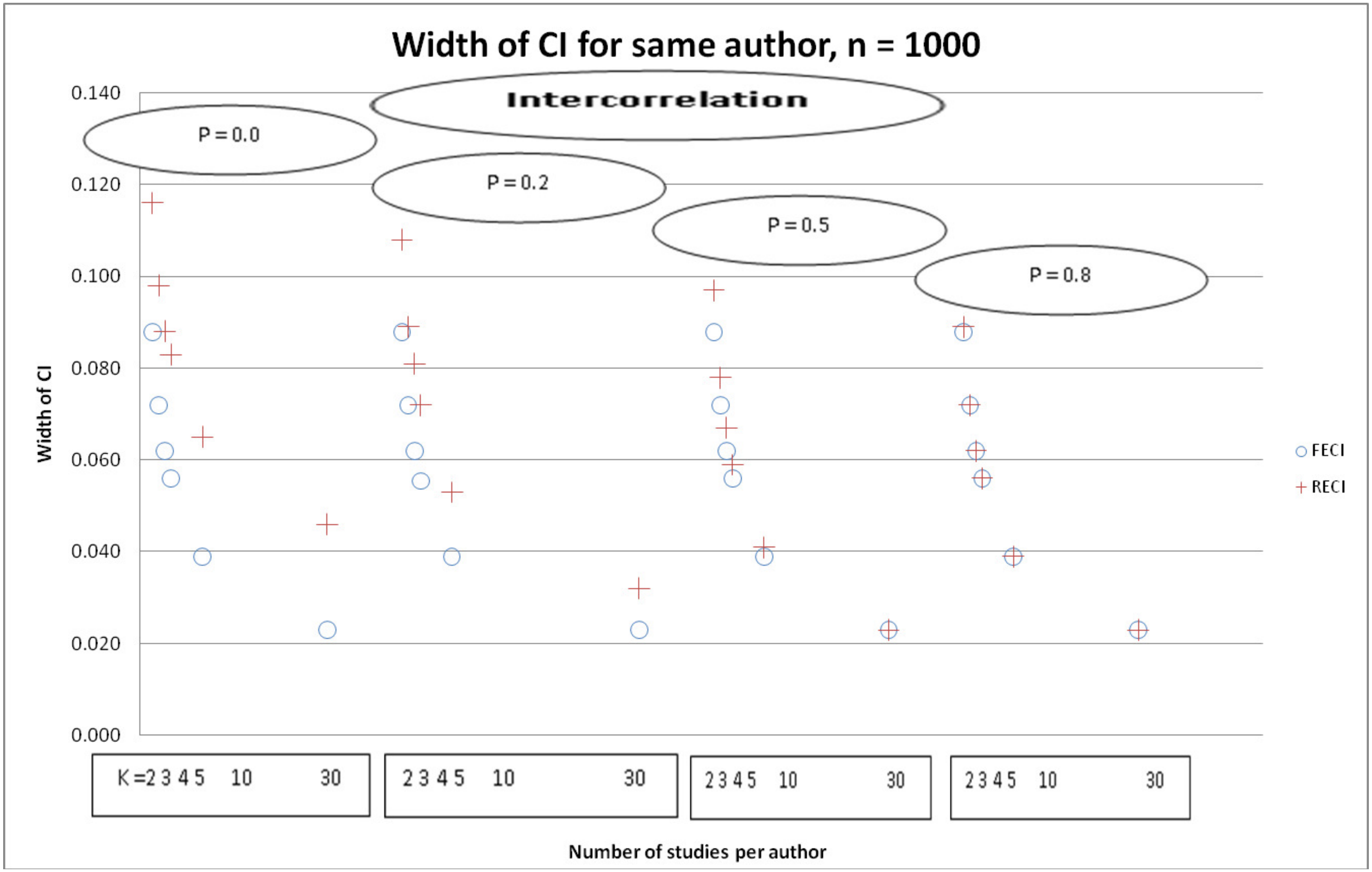
* In the Y-axis, 0.01 to 0.1 represent 1% to 10%.

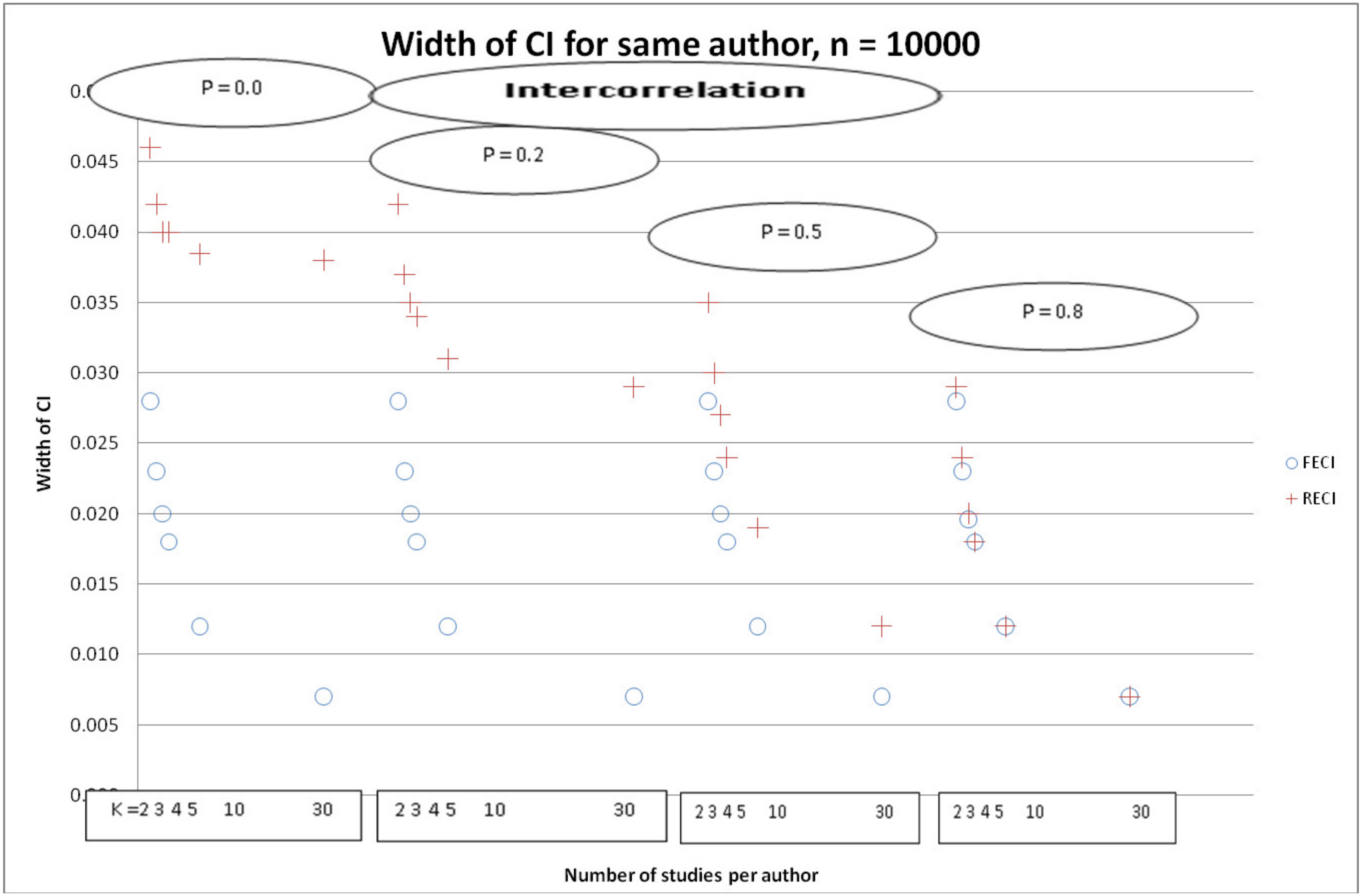


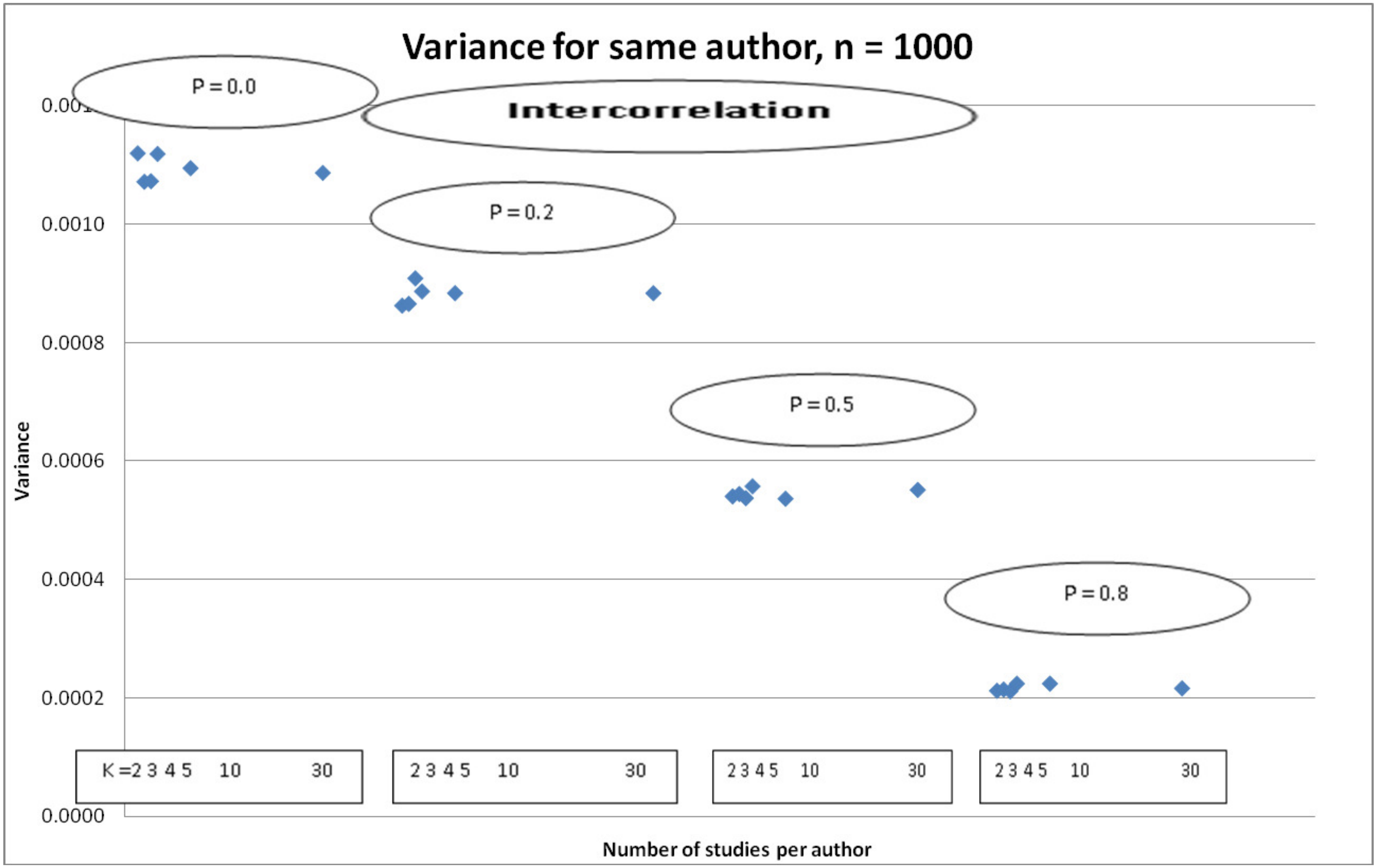
* In the Y-axis, 0.01 to 0.1 represent 1% to 10%.

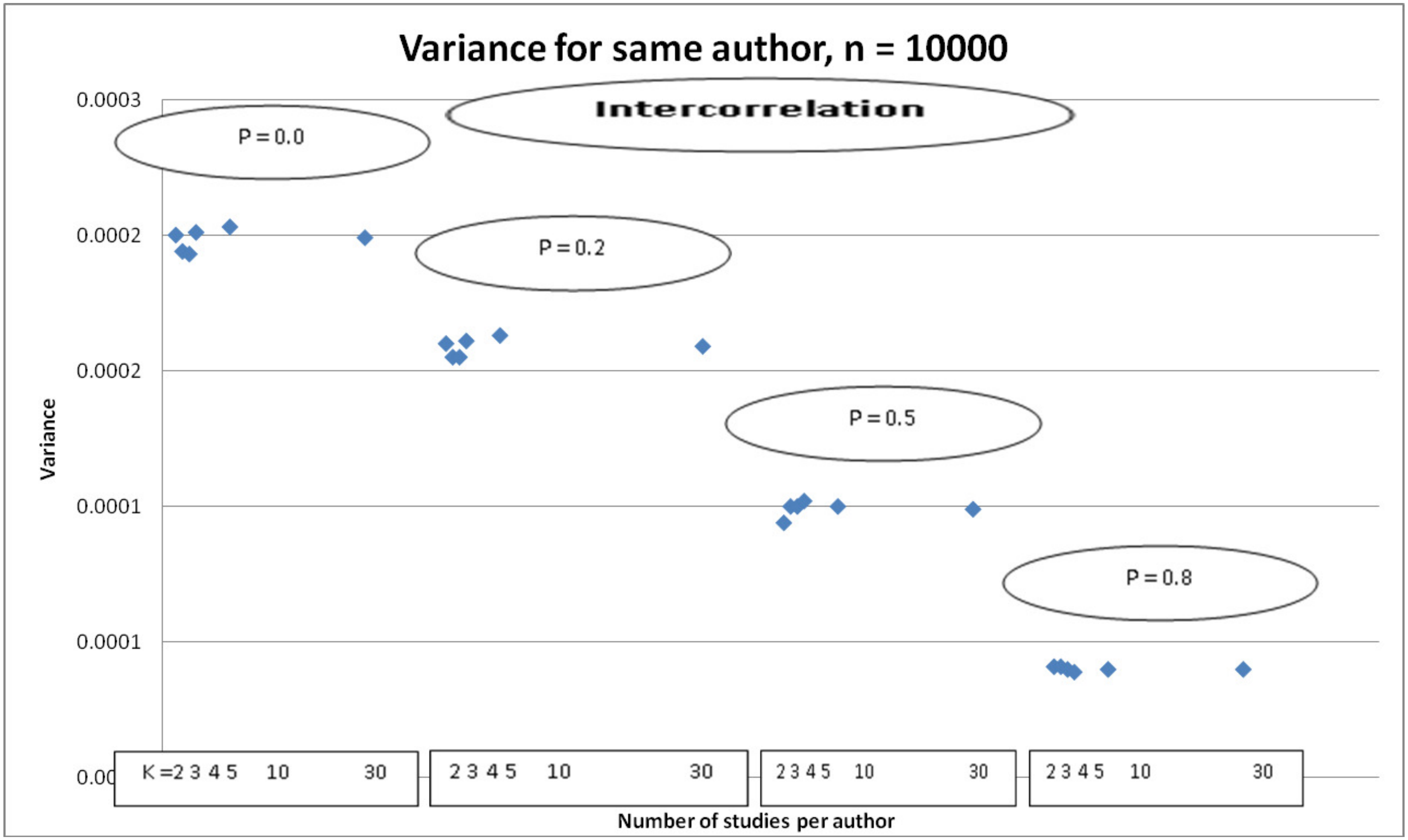

Average Q for same author, n = 1000
P = 0.0
Intercorrelation
P = 0.2
P = 0.5
P = 0.8
Average Q
30
25
20
15
10
5
0
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author


Average Q for same author, n = 10000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
60
50
40
30
20
10
0
Average Q
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author


Width of CI for same author, n = 1000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
0.140
0.120
0.100
0.080
0.060
0.040
0.020
0.000
Width of CI
FECI
RECI
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author


Width of CI for same author, n = 10000
P = 0.0
Intercorrelation
P = 0.2
P = 0.5
P = 0.8
Width of CI
0.0
0.045
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author
○ FECI
+ RECI


Variance for same author, n = 1000
P = 0.0
Intercorrelation
P = 0.2
P = 0.5
P = 0.8
0.001
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Variance
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author


Variance for same author, n = 10000
Intercorrelation
P = 0.0
P = 0.2
P = 0.5
P = 0.8
0.0003
0.0002
0.0002
0.0001
0.0001
0.00
Variance
K =2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
2 3 4 5 10 30
Number of studies per author

REFERENCES

Achilles, C. M. (1994). *The multiple benefit of class size research: A review of STAR's legacy, subsidiary and ancillary studies*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN.

Aloe, A. M., & Becker, B. J. (2007). *Teacher verbal ability and school outcome: Where is evidence?* Paper presented at the annual meeting of the American Educational Research Association in Chicago, IL

Bateman, I. J., & Jones, A. P. (2003). Contrasting conventional with multi-level modeling approaches to meta-analysis: An illustration using UK woodland recreation values. *Land Economics*, *79*, 235-258.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlation: A simulation. *Organizational Research Methods*, *1*, 355-373.

Bliese, P. D., & Halverson, R. H. (1998a). Group consensus and psychological well-being: a large field study. *Journal of Applied Social Psychology*, *28*, 563-580.

Bliese, P. D., & Halverson, R. H. (1998b). Group size and measure of group-level properties: an examination of eta-squared and ICC values, *Journal of Management*, *24*, 157-172.

Bar-Haim, Y., Bakermans-Kranenburg, M. J., Lamy, D., Pergamin, L., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin*, *133(1)*, 1-24.

Castro, S. L. (2002). Data analytic methods for the analysis of multilevel questions, *The Leadership Quarterly*, *13*, 69-93.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: U. S. Department of Health, Education and Welfare.

Commenges, D., & Jacqmin, H. (1994). The intraclass correlation coefficient: Distribution-free definition and test. *Biometrics, 50*, 517-526.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H. M., & Hedges, L. V. (Eds.). (1994) *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cooper, H. M., Robinson, J. C., & Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987-2003. *Review of Educational Research*, *76(1)*, 1-62.

Donner, A. (1986). A review of inference procedures for the ICC in the one-way random effects model. *International Statistical Review*, *54(1)*, 67-82.

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin*, *132(1)*, 33-72.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class-size: A statewide experiment. *American Educational Research Journal*, *27*, 557-577.

Finn, J. D., Fulton, D., Zaharias, J., & Nye, B. A. (1989). Carry-over effects of small classes. *Peabody Journal of Education*, *67(1)*, 75-84.

Frattaroli, J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin*, *132(6)*, 823-865.

Gijbels, D., Dochy, F., Bossche, P. V., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, *75(1)*, 27-61.

Glasman, L. R. & Albarracin, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin*, *132(5)*, 778-822.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, *1(1)*, 2-15.

Gleser, L.J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Russell Sage Foundation.

Goldberg, W. A., Prause, J., Lucas-Thompson, R., & Himsel, A. (2008). Maternal employment and children's achievement in context: A meta-analysis of four decades of research. *Psychological Bulletin*, *134(1)*, 77-108.

Goldstein, H. & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal*, *24(3)*, 255-268.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class-size effects. *Applied Statistics*, *49(3)*, 399-412.

Greenhouse, J. B., & Iyengar. S. (1994). Sensitivity analysis and diagnostics. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383-398). Russell Sage Foundation, New York.

Gurevitch, J., & Hedges, L. V. (1999). Statistical issues in ecological meta-analyses, *Ecology*, *80(4)*, 1142-1149.

Hannan, P.J., Murray, D. M., Jacobs, D. R., & McGovern, P. G. (1994). Parameters to aid in the design and analysis of community trials: Intraclass correlations from the *Minnesota Heart Health Program. Epidemiology*, *5*, 88-95.

Hedges, L. V. (1994). Fixed effects models. In H.M. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285-299). New York: Russell Sage Foundation.

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

Higgins, J. P. & Thompson, S. G. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539-1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D.G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557-560.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, and Public Policy and Law, 2*, 324-347.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effect and random effects model. *International Journal of selection and assessment*, *8*, 275-292.

Ingrisone, J., & Ingrisone, S. J. (2007). *A meta-analysis: what works in English as a second language instruction.* Paper presented at the annual meeting of the Florida Educational Research Association in Tampa, FL.

Johnston, J., Bain, H. P., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N., et al (1990). *The state of Tennessee's student/teacher achievement ratio (STAR) project: Final summary report 1985-1990*. Tennessee Department of Education.

Kalaian, S. A. (2003). Meta-analysis methods for synthesizing treatment effects in multisite studies: hierarchical linear modeling (HLM) perspective. *Practical Assessment, Research & Evaluation, 8(15)*. Retrieved September 9, 2008 from http://PAREonline.net/getvn.asp?v=8&n=15.

Kim, J. P. (2000) *An Empirical Study of the Effect of Pooling Effect Sizes on Hedge's Homogeneity Test.* Paper presented at the annual meeting of the American Educational Research Association in New Orleans, LA.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75(1)*, 63-82.

Lauer, P. A., Akiba, M., Wilkerson, S. B., Apthorp, H. S., Snow, D., & Martin-Glenn, M. L. (2006). Out-of-school-time programs: A meta-analysis of effects for at risk students. *Review of Educational Research*, *76(2)*, 275-313.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlations coefficients. *Psychological Methods*, *1(1)*, 30-46.

McGiverin, J., Gilman, D., & Tillitsk, C. (1989). A meta-analysis of the relation between class size and achievement. *The Elementary School Journal*, *90*, 47-56.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (Surprising) meta-analysis. *Psychological Bulletin*, *132(6)*, 895-919.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children*, *5(2)*, 113-127.

Murray, D. M., & Blistein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, *27*, 79-103.

Murray, D. M., Phillipas, G. A., Bimbaum, A. S., & Lytle, L. A. (2001). Intraclass correlation for measures from a middle school nutrition intervention study: Estimates, correlates, and applications. *Health Education & Behavior*, *28(6)*, 666-679.

Nestbit, J. C. & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76(3)*, 413-448.

Nye, B. A., Zaharias, J. B., Fulton, B. D., Wallenhorst, M. P., Achilles, C. M., & Hooper, R. (1992). *The lasting benefit study: A continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels: Fifth grade technical report*. Center of excellence for research in basic skills Tennessee state university, Nashville, Tennessee.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7(1)*. Retrieved September 10, 2008 from http://PAREonline.net/getvn.asp?v=7&n=1.

Rappaport, A. (1967). Sensitivity analysis in decision making. *The Accounting Review*, *42(3)*, 441-456.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology*, *76* , 85-97.

Raudenbush, S. W. (1988). Educational application of hierarchical linear models: Review. *Journal of Educational Statistics*, *13(2)*, 85-116.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical bayes meta-analysis. *Journal of Educational Statistics, 10(2)*, 75-98.

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical linear models: Applications and data analysis methods* (2nd ed.), Newbury Park, CA: Sage Publications.

Rose, A. K., & Stanley, T. D. (2005). A meta-analysis of the effect of common currencies on international trade. *Journal of Economic Surveys*, *19(3)*, 347-365.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rucker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008) Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology, 8,* 79-96

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. *Psychological Bulletin*, *86(2)*, 420-428.

Shin, I. S. (2008) *Meta-analysis: The effect of class size on student achievement.* Unpublished manuscript.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75(3)*. 417-453.

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research*, *75(2)*, 247-284.

Sliwinski, M. J., & Hall, C. B. (1998). Constraints on general slowing: A meta-analysis using hierarchical linear models with random coefficients, *Psychology and Aging*, *13(1)*, 164-173.

Sohn, S. Y. (2000). Multivariate meta-analysis with potentially correlated marketing study results. *Naval Research Logistics*, *47(6)*, 500-510.

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, *133(1)*, 65-94.

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34(1),* 46-73.

Tolin, D. F., & Foa, E. B. (2006). Sex differences in trauma and posttraumatic stress disorder: A quantitative review of 25 years of research. *Psychological Bulletin*, *132(6)*, 959-992.

Uchiyama, K., & Simone, G. (1999) *Publishing educational research guidelines and tips*. Retrieved October 3, 2008, from American Educational Research Association Website: https://www.aera.net/uploadedFiles/Journals_and_Publications/Journals/pubtip.pdf

Wampold, B. E., & Serlin, R. C. (2000). The consequences of ignoring a nested factor on measures of. effect size in analysis of variance. *Psychological Methods*, *5*, 425–433.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K-12 mathematics tests. *Educational and Psychological Measurement,* 67, 219-238

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper and pencil testing in K-12 reading assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement,* 68, 5-24

Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, *132(2)*, 249-268.

Weisz, J. R., McCarty, C. A., & Valeri, S. M. (2006). Effects of psychotheraphy for depression in children and adolescents: A meta-analysis. *Psychological Bulletin*, *132(1)*, 132-149.

BIOGRAPHICAL SKETCH

EDUCATION

Florida State University, Tallahassee, FL
Ph. D. in Measurement and Statistics, August 2009

Florida State University, Tallahassee, FL
Masters of Science in Foundations of Education, May 2004

Yonsei University, Seoul, Korea
B.A in Law, August 1996

Yonsei University, Seoul, Korea
B.A in Education, February 1995

ADDITIONAL COURSEWORK

Seoul National University, Seoul, Korea
Graduate Coursework taken in Public Administration (March 2000 to August 2002)

Young-San University, Seoul, Korea
Graduate Coursework taken in Law (March 2000 to August 2001)

PROFESSIONAL EXPERIENCE

Psychometrician: Florida Department of Education, Division of Accountability Research and Measurement, Test Development Center
November 2006 to present

· Supporting Psychometric issues for FCAT (Florida Comprehensive Assessment Test) such as review specification documents, test construction, check calibration sample for representativeness, calibration, equating, and scaling for FCAT.
· Standard setting, DIF, Design of new Writing test, Technical reports, Ad hoc Reports, Writing FT prompt Statistics, TAC committee.
· Data analysis support for CELLA (Comprehensive English Language Learner Assessment)

Graduate Student Assistant, Leon County School District
May 2004 to May 2006

· Assist the staff in the Department of Student Assessment in the coordination of the district and state assessment program
· Provide support services for classroom teachers in the proper use and interpretation of assessment procedures

Student Assistant, Florida State University Dirac Science Library
January 2004 to May 2004
· Preshelving and organizing library resources according to Dewey Decimal and Congress of Library Systems

Deputy Director, Kyonggi Province Office of Education School Building Department, Kyonggi Province, Korea
January 2001 – August 2002
· Negotiating Budget from Ministry of Education
· Redistributing budget to 25 local offices for building school for class size reduction
· Checking and reporting the process of building school process

Research Coordinator, Prime Minister's Office of Korea
April 2000 – Dec 2000
· Coordinated research for anti-corruption policy on education funded by World Bank
· Conducted Survey research for knowing public understanding on anti-corruption in educational administration
· Created policies with research team

Deputy Director, Presidential Commission for New Educational Committee, Korea
April 1999 – April 2000
· Supported Subcommittee leaders in vocational & life-long education and educational reform subcommittee
· Organized agendas for meeting and managed internet-homepage
· Reported meeting results to the presidential secretary via computer documents
· Held a panel discussion for educational reform

Director, Kang-Nung National University Planning Department
April 1997 – October 1997
· Lead committee in the long-term university development plan committee
· Discussed long-term plan with professors
· Received university evaluations from Ministry of Education
· Implemented plans to improve the quality of the university