

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**The Effect of Clustering on the  
Precision of Estimation**

A thesis presented in partial  
fulfilment of the requirements

for the degree

of Master of Business Studies  
in Marketing at  
Massey University

Zhengping Guan

1997

## ABSTRACT

The effect of clustering interval on design effect may be important in selection of alternative sampling designs by evaluating the cost-efficiency in the context of face-to-face interview surveys. There has been little work in investigating this effect in New Zealand. This study attempts to investigate this effect by using data from a two-stage sampling face-to-face interview survey.

Seventeen stimulated samples are generated. A simple

method,  $design\ effect = \frac{ms_b}{ms}$ , is developed to estimate design

effects for 81 variables for both the simulated samples and the original sample. These estimated design effects are used to investigate the effect of clustering interval.

This study also investigates the effect of cluster size.

The results indicate that clustering interval has little influence on design effect but cluster size substantial

influence. The evaluation of the cost-efficiency in

alternative clustering intervals is discussed. As an

improvement in the efficiency of a sample design by an

increase in clustering interval can not be justified by

the increase in cost, it seems that the sample design with

the smallest clustering interval is the best. An

alternative method  $design\ effect \approx mr^2$  is also discussed and

tested in estimating design effects. The result indicates

that the applicability of  $design\ effect \approx mr^2$  is the same as

that of  $design\ effect = \frac{ms_b}{ms}$ .

## **ACKNOWLEDGMENT**

I would like to thank ACNielsen McNair for providing data from a face-to-face interview survey.

Thanks are also due to Mr Nick Jones, Managing Director of ACNielsen McNair, for his helpfulness and kind cooperation, and due to Mr James Reilly for his kind assistance in preparing the data.

# CONTENTS

	Page
<b>ABSTRACT</b> -----	<b>ii</b>
<b>ACKNOWLEDGMENT</b> -----	<b>iii</b>
<b>LIST OF TABLES</b> -----	<b>vii</b>
<b>LIST OF FIGURES</b> -----	<b>viii</b>
<b>1. INTRODUCTION</b> -----	<b>1</b>
<b>2. METHODS OF ESTIMATING SAMPLING VARIANCES</b> -----	<b>4</b>
<b>2.1 Standard (Mathematical) Methods</b> -----	<b>4</b>
<b>2.2 Subsampling Methods</b> -----	<b>11</b>
2.2.1 Random Group Methods -----	11
2.2.2 Balanced Repeated Replication Methods (BRR) --	12
2.2.3 Jackknife Methods -----	14
2.2.4 Bootstrap Methods -----	15
<b>2.3 Modelling Methods</b> -----	<b>15</b>
2.3.1 The Taylor Linearization Method -----	15
2.3.2 The Generalized Variance Function Method -----	17
<b>2.4 Discussion of Variance Estimation Methods</b> -----	<b>17</b>
<b>3. DESIGN EFFECT</b> -----	<b>24</b>
<b>3.1 Introduction</b> -----	<b>24</b>
<b>3.2 Design Effects for Different Statistics and</b> <b>Variables</b> -----	<b>25</b>

3.2.1 Design Effect for Different Statistics -----	25
3.2.2 Design Effect for Different Variables -----	27
<b>3.3 Design Effect and Stratification -----</b>	<b>27</b>
<b>3.4 Design Effect and Clustering -----</b>	<b>27</b>
3.4.1 Design Effect and Cluster Size -----	28
3.4.2 Design Effect and Clustering Interval -----	30
<b>4. METHOD -----</b>	<b>32</b>
4.1 Procedure -----	32
4.2 Samples -----	33
4.2.1 Original Sample -----	33
4.2.2 Simulated Samples -----	36
4.3 Estimation for Design Effect -----	38
4.3.1 Considerations of Simplicity -----	38
4.3.2 Estimation Method for Design Effect -----	39
4.3.3 An Alternative Method of Estimating Design Effect -----	40
4.4 Significance Tests -----	42
4.5 Evaluation of Cost-Efficiency in the Sample Designs with Alternative Clustering Intervals --	43
<b>5. RESULTS -----</b>	<b>44</b>
5.1 Design Effects -----	44
5.1.1 The Effect of Cluster Size -----	44
5.1.2 The Effect of Clustering Interval -----	48
5.2 Applicability of <i>design effect</i> $\approx mr^2$ -----	54

5.3 The Effect of Clustering Interval on Cost- Efficiency of Sample Designs. -----	55
6. DISCUSSION -----	58
7. CONCLUSION -----	61
APPENDICES -----	62
Appendix A. Definition of Variables Selected -----	63
Appendix B. Formation of Simulated Samples -----	68
Appendix C. the Mathematical Derivation of $design\ effect = \frac{ms_b}{ms}$ -----	70
Appendix D. Design Effects of Variables in Different Clusterings -----	72
Appendix E. Homogeneity -----	77
Appendix F. Comparison of Two Variance Estimation Methods. -----	86
REFERENCES -----	90
BIBLIOGRAPHIES -----	94

## LIST OF TABLES

	Page
Table 1. Frequency of Households Interviewed-----	35
Table 2. Response Rate for Designed Sample Size 936-----	36
Table 3. Design Effects for the Quartiles of Variables-----	44
Table 4. Variability of Design Effect among Variables in different Cluster Sizes-----	47
Table 5. t-tests for Differences of Design Effects between Cluster Sizes-----	48
Table 6. Design Effects for the Quartiles of 81 Variables in Different Clusterings-----	50
Table 7. Variability of Design Effect among Variables in Different Clusterings-----	52
Table 8. t-tests for Differences of Design Effects between Clustering intervals-----	54
Table 9. Comparison of Two Design Effect Estimation Methods-----	55
Table 10. Variables Selected-----	63
Table 11. Design Effects of Variables in Different Clusterings-----	73
Table 12. Homogeneity across Variables and Clusterings-----	79
Table 13. Comparison of Two Design Effect Estimation Methods with 41 Variables-----	87



## LIST OF FIGURES

	Page
Figure 1. Relation between Design Effect and Clusterings-----	45
Figure 2. Relation between Design Effect and Cluster Size-----	46
Figure 3. Relation between Design Effect and Clustering Interval with Cluster Size 2-----	49
Figure 4. Relation between Design Effect and Clustering Interval with Cluster Size 6-----	49
Figure 5. Relation between Design Effect and Clustering Interval with Cluster Size 4-----	51
Figure 6. Relation between Homogeneity and Clusterings-----	83
Figure 7. Relation between Homogeneity and Cluster Size with a Given Clustering Interval-----	84
Figure 8. Relation between Homogeneity and Clustering with Cluster Size 6-----	85
Figure 9. Relation between Homogeneity and Clustering with Cluster Size 4-----	85
Figure 10. Relation between Homogeneity and Clustering with Cluster Size 2-----	85

## 1. INTRODUCTION

Surveys using clustered multi-stage sampling designs are common in research in business and other social sciences. For a given sample size, these sampling designs may reduce the cost of data collection. However, such designs lead to increase in the sampling variances of estimates.

This study investigates the way in which final stage clustering affects sampling variances in face-to-face interview surveys.

In view of the need to make an adjustment to a sampling variance estimate from a complex sample design, Kish (1965) proposed a measurement which he called "design effect" to describe the sampling variance increase due to the complex sample design. He held the position that sample designs affect variance estimation and statistical analysis. However, Skinner, Holt & Smith (1989 chapter 2) argued that it was population structure rather than sample designs that affected variance estimation and statistical analysis. These two positions are often consistent. For a given sample design, population structure may affect variance estimation and statistical analysis, and vice versa.

Skinner et al (1989, p 24) also proposed an alternative measurement which they called "misspecification effect" instead of design effect. That is, the measurement of sample design efficiency is sampling variance of the actual sample design over the expected value of sampling variance of a simple random sample with the same size, rather than sampling variance of the actual sample design over sampling variance of a simple random sample with the same size. However, it is difficult in practice to obtain the expected value of a sampling variance estimate. Thus, design effect is likely to be more applicable in measuring the efficiency of sample designs than misspecification

effect.

Sampling variance increase due to clustering in surveys is caused by similarity of elements within clusters. This similarity is measured by the homogeneity of within-cluster elements.

There is a voluminous body of literature concerning complex sample design, variance estimation, design effect and homogeneity. However, there has been little research into the relation between design effect and intervals of selecting elements within clusters in New Zealand. The need to evaluate the cost-efficiency of the alternative sample designs with different clustering intervals requires to conduct an investigation into the effect of clustering interval on design effect.

Data for this study is from a face-to-face interview survey conducted by ACNielsen-McNair. This is a two-stage sample (see Chapter 4 for specification of the sample). A number of simulated samples are drawn from it to investigate the effect of clustering interval (see Chapter 4 for the detailed discussion in generating simulated samples).

Based on the design effects estimated from both the original sample and the simulated samples, this study investigates the following:

- a. The relation between design effect and clustering interval;*
- b. The relation between design effect and cluster size;*

c. *The applicability of the formula:*

$$\text{design effect} \approx mr^2$$

*(see Chapter 4 for both specification and derivation of this formula);*

d. *The effect of clustering interval on cost-efficiency of alternative sample designs.*

The results for both a and b should be that design effect decreases with either increase in clustering interval or decrease in cluster size. The result for c should justify the alternative estimation method for design effect. The result for d should provide the guideline for selection of the alternative sample designs with different clustering intervals.