DOCUMENT RESUME

ED 374 136 TM 021 991

AUTHOR de Leeuw, Edith Desiree
TITLE Data Quality in Mail, Telephone and Face to Face Surveys.
SPONS AGENCY Netherlands Organization for Scientific Research.
REPORT NO ISBN-90-801073-1-X; NUGI-659
PUB DATE 92
CONTRACT 500278008
NOTE 177p.
AVAILABLE FROM T. T. Publikaties, Plantage Daklaan 40, 1018CN Amsterdam ($20; 37 Dutch florins).
PUB TYPE Reports - Research/Technical (143) -- Books (010)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS *Adults; Comparative Analysis; *Data Collection; Foreign Countries; *Interviews; *Mail Surveys; Meta Analysis; *Research Methodology; Research Problems; Responses; *Telephone Surveys; Training
IDENTIFIERS *Empirical Research

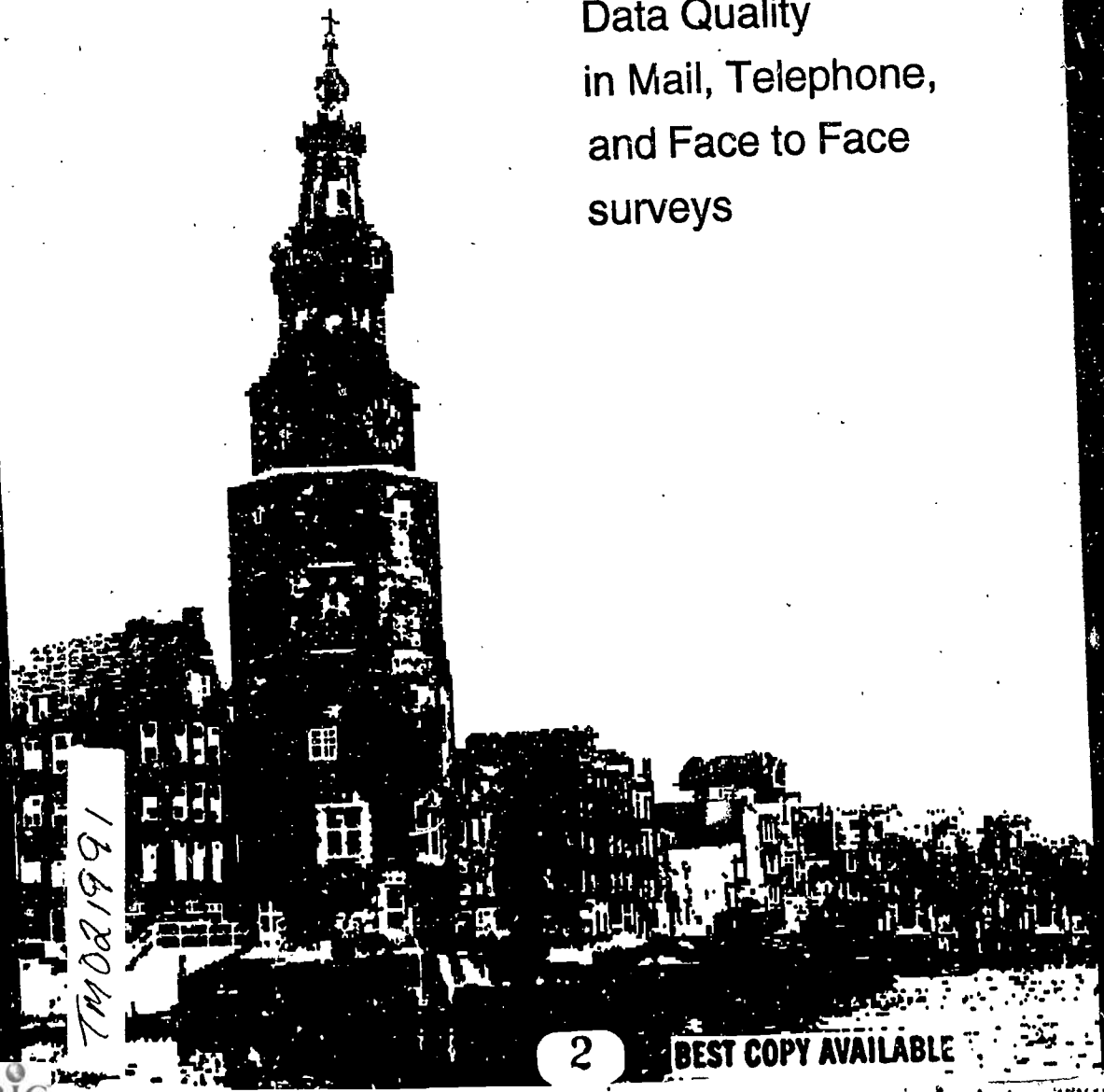
ABSTRACT

Three major methods of survey research, face-to-face interviews, telephone interviews, and mail questionnaires, are compared with respect to the quality of the data. The literature on experimental comparisons of these methods is reviewed, and the effects of the mode of data collection on aspects of data quality are examined. The effects of the data-collection method on research results are also examined with a focus on the consequences for the relations among variables and emerging empirical models. The meta analysis is followed by a field experiment with 762 responses. Meta analysis detected small differences between the modes, suggesting a dichotomy between modes with and without an interviewer. The field experiment found the lowest response rates for the face-to-face survey, with more item nonresponse in the mail survey but more self-disclosure through the mail. The mail survey was slightly superior in reliability and scalability. Results suggest that interviewer training should be adapted to the changes in data-collection mode. Five figures and 33 tables present meta analysis and survey findings. Three appendixes contain a bibliography, the questionnaire content, and marginal distributions of background variables. A summary in Dutch is included. (Contains 201 references.) (SLD)

TM

E. D. de Leeuw

Data Quality in Mail, Telephone, and Face to Face surveys



DATA QUALITY IN MAIL, TELEPHONE AND FACE TO FACE SURVEYS

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Leeuw, Edith Desirée de

Data quality in mail, telephone and face to face surveys
/ Edith Desirée de Leeuw. - Amsterdam: TT-Publikaties.
- Ill., fig.
Proefschrift Vrije Universiteit Amsterdam. - Met lit. opg.,
reg. - Met samenvatting in het Nederlands.
ISBN 90-801073-1-X
NUGI 659
Trefw.: enquêtes

Omslagontwerp: Joop Hox en Gerard Kurvers

VRIJE UNIVERSITEIT

DATA QUALITY IN MAIL, TELEPHONE AND FACE TO FACE SURVEYS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus
dr. C. Datema,
hoogleraar aan de faculteit der letteren,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der sociaal-culturele wetenschappen
op donderdag 22 oktober 1992 te 13.30 uur
in het hoofdgebouw van de universiteit, De Boelelaan 1105

door

Edith Desirée de Leeuw

geboren te Amsterdam

TT-Publikaties Amsterdam

1992

Promotoren: prof.dr. J. van der Zouwen
prof.dr. G.J. Mellenbergh
Referent: prof.dr. D.A. Dillman

PREFACE

Que serai-je sans toi . . .
Louis Aragon, Le roman inachevé

Writing a book needs a lot of stubbornness and love. I am very stubborn.

I would like to thank my parents who always stimulated and accepted me, my teachers who taught me that science is fun, my friends who always listened, and my husband who believed in me. To them I dedicate this book, for what would I be without their love?

In this book, titled Data Quality in Mail, Telephone, and Face to Face Surveys, I studied three main data collection modes. I tried to summarize our knowledge of mode differences and bring the literature together. I also undertook to expand the existing knowledge by designing an experimental comparison to investigate how these data collection modes influence the way items scale together and how they affect multivariate models.

I wrote this book mainly for survey researchers and survey methodologists. Researchers who occasionally use survey methods and graduate students interested in survey methods may find this study useful too.

This book will be publicly defended as last fulfillment of the requirements for the degree of Doctor in the Social and Cultural Sciences. According to the rules of the Vrije Universiteit I added a summary in the Dutch language and a separate brochure with defendable theses.

This research has been partly funded by the social and cultural sciences foundation of the Netherlands Organization for Scientific Research (NWO) under grant number 500278008. I gratefully acknowledge the organizational support of the Department of Education, University of Amsterdam, and the Department of Social Research Methodology, Vrije Universiteit. I had the opportunity to stay as a Fulbright scholar at the Social and Economic Sciences Research Center at Washington State University, and as a visiting scholar at the Social Statistics Program of the Department of Psychology of the University of California, Los Angeles. I thank both organizations for their hospitality and stimulating research environment.

I thank my two supervisors, Don Mellenbergh and Hans van der Zouwen, who skillfully guided me through a tangled maze, and my referent Don Dillman, who inspired me to draw out the best in me.

Special thanks are due to Gerard Kurvers and Geo-Marktprofiel for their permission to use their zip-code information. I thank Marius de Pijper and Joop Hox who wrote several computer programs, and Klaas ten Hoeve for his technical assistance. Fred Bronner, Cees van Rooij and Steef de Bie provided much needed assistance in producing the equivalent versions of the questionnaire. Colleagues all over the world were kind enough to read and comment upon chapter drafts. I am especially indebted to Fred Bronner, Jenny de Jong-Gierveld, Joop Hox, Janneke Lely, Rob Meijer, Monica Meijsing, and Tom Pettigrew.

My research has benefited from the stimulating discussions in various research committees of which I am a member. I specifically want to mention the SOMO research committees on conceptualization and research design and on data collection, and the biweekly discussion group directed by Don Mellenbergh.

For their invaluable assistance during the data collection phase I want to thank the boys and girls in the back office: Els Beyderweilen, Astrid van Hattum, Joop Hox, Akke de Leeuw, Jet Naftaniel-Joëls, and Corine Noordam. Menno Zooitjes made it possible to use the facilities of the Vrije Universiteit during evening hours and the weekend. I thank Pia Dorman for drawing the figures, and Yolande Brands-Dorst for converting my old fashioned Wordstar files into Wordperfect.

Special thanks are due to Sunil Abhelakh, Frank van As, Koen Becking, Elja Bouwmeester, Mechteld Dijkman, Carla Gavrey-Jacobs, Cisca Jonkman, Harriët Kroon, Margreet van Lookeren Campagne, Marja Morsch-Broekhuizen, Corine Noordam, Hennie Oosterom, Els van der Ploeg, Hortense Spruyt-van Latum, Yvonne Towikromo, Peterke Ubbens, Bauke Viersma, and Willeke van der Weide, who formed a great interviewer team.

EDITH D. DE LEEUW

Amsterdam
August 1992

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Could you not begin at the beginning . . .
Dorothy L. Sayers, Murder must advertise, 1975, p. 57

1.1. The Face to Face Interview and its Alternatives

The face to face interview is one of the oldest forms of data collection in surveys, and it has evolved from a short and simple inquiry in the thirties into a complex and highly flexible research instrument (Rossi, Wright & Anderson, 1983; Smith, 1987). Because of its flexibility and great potential, the face to face interview has long been considered a superior data collection technique. Although mail surveys have been extensively used -in 1981 two thirds of the U.S. federal statistical surveys used self-administered questionnaires as the only means of data collection (Thornberry, Nicholls, & Kulpinsky, 1982)- the data collected by mail surveys have often been considered suspect unless proven otherwise. This is exactly the opposite of the view held toward the accepted face to face interview (Dillman, 1978, p. 1).

In the last two decades, telephone interviews have become increasingly popular in government agencies and survey research firms (Lyberg & Kasprzyk, 1991). This is caused by improved technology, by the development of random digit dialing as a sampling technique, but, above all by the increased availability of and access to telephones for the general public. For example, in the seventies the telephone coverage for households in the Netherlands doubled from approximately 40% to 80% (Bronner, 1980). According to Dutch Telecom, in 1990 approximately 92% of all private households had a telephone, while approximately eight percent of all private numbers were unlisted (cf. Dykstra, 1990, p. 29). For an international comparison of telephone coverage, see Trewin and Lee (1988). Nevertheless, although the telephone interview has attained an increasing significance in the daily practice of data collection, it also had to prove itself against the generally accepted face to face interview (Körmendi & Noordhoek, 1989; Sykes & Collins, 1988).

The increased costs of interviewing make it virtually impossible, or at least extremely costly, to utilize the face to face survey to its full potential when national surveys or large surveys in geographically dispersed areas are done. This has led to a renewed interest in alternatives for face to face interviews, and a renewed research effort to optimize mail and telephone surveys. For instance, Dillman (1978) gives an inspired account of mail survey research, with a clear and precise description of how to optimize mail and telephone surveys by using the Total Design Method or TDM. An excellent overview of the potential of telephone surveys is given in Groves, Biemer, Lyberg, Massey, Nicholls, and Waksberg (1988).

The following statistics illustrate the relative importance of mail and telephone surveys in the Netherlands; these statistics are based on turnover figures of research institutes organized in the Netherlands Association for Marketing Research (VMO). In 1990 telephone interviews were used in 18% of all studies commissioned, and self-administered questionnaires were used in 35% of the cases. Some form of face to face interview was used in 41% of all investigations (in 27% of all studies interviews were conducted at the respondent's home, office or in shopping malls, and in 14% of all cases they took place at the premises of the research institute), while in 6% of the studies another research method was used (Bro ner, 1991).

The heightened interest in mail and telephone surveys has stimulated discussion of the relative advantages and disadvantages of these methods, and individual researchers are now faced with a difficult decision when selecting a data collection method for their survey. Besides costs, other factors enter into this complex decision process such as the population under study, the questionnaire content, and the administrative and staff resources available.

The availability of alternative methods for the rather expensive face to face survey has also increased the demand for comparative research on the influence of data collection methods on the resulting data quality. When the strengths and weaknesses of different survey methods are identified, designs can be developed that reduce both survey error and survey costs.

In this book three major methods of survey research, face to face interviews, telephone interviews and mail questionnaires, are compared with respect to the quality of the data. The purpose of this study is to: (1) review the literature on experimental comparisons of these data collection methods, (2) examine the effects of the mode of data collection on various aspects of data quality, and (3) examine the effects of mode of data collection on research results, especially on the consequences for the relationships between variables and the emerging empirical models.

In the remaining sections of this chapter I will first give a definition of the three data collection methods under comparison. This is followed by an overview of the relative strengths and weaknesses of mail, telephone, and face to face surveys concerning various practical attributes such as sampling control, nonresponse and administrative arrangements. A discussion of data quality is reserved for chapter 2 where I provide an overview of mode factors that may influence data quality. In the last section of this chapter the outline of this book is presented.

1.2. Concise Definitions of the Major Data Collection Methods

In this study three major methods of survey research, face to face interviews, telephone interviews and mail questionnaires, are compared. To avoid misunderstanding, I will start with a concise definition of these data collection methods, based on Groves and Kahn (1979) and Lyberg and Kasprzyk (1991). The *face to face interview* is the mode in which an interviewer administers a structured or partly structured questionnaire to a respondent within a limited period of time and in the presence (usually at the home) of the respondent. In a *telephone interview* the interviewer administers the questions (from a structured questionnaire and within a limited period of time) via a telephone. Telephone interviewing is often centralized; i.e., all interviewers work from a central location under direct supervision of a field manager or a quality controller. When a *mail questionnaire* is used, a respondent receives a structured questionnaire and an introductory letter by mail, answers the questions in her/his own time without any assistance (from the researcher or her/his representative) except for any written instructions in the questionnaire or in the accompanying letter, and finally sends the questionnaire back.

In the last decade computer assisted procedures for these three main data collection techniques were developed, of which CATI (computer assisted telephone interviewing) is the oldest and the best developed. Besides CATI, these procedures include CAPI (computer assisted personal interviewing), and CASAQ (computer assisted self administered questionnaires). For an introduction, see Hox, De Bie, and De Leeuw (1990), Nicholls and Groves (1986), and Saris (1989, 1991).

1.3. Practical Advantages and Disadvantages of Mail, Telephone, and Face to Face Surveys

This section is based on overviews given by Dillman (1978, chapter 2) and Tull and Hawkins (1984, chapter 5). It is organized around the following factors relevant for judging which type of survey to use in a particular situation: type of population and sample control, nonresponse, type of questions and complexity of questionnaire, and resources available.

Population of interest and possibility of sample control

When one is interested in studying the general population the face to face survey has the greatest potential. Sophisticated sampling designs for face to face surveys have been developed, which do *not* require a detailed sampling frame or a list of persons or households (cf. Cochran, 1977; Kish, 1965, 1987). For instance, area probability sampling can be used to select geographically defined units (e.g., streets or blocks of houses) as primary units and households within these areas. Elaborate techniques based on household listings (i.e., inventories of all household members derived by an interviewer) can then be used to select one respondent from those eligible in a household (Kish, 1949).

Random digit dialing techniques, which are based on the sampling frame of all possible telephone numbers, make it possible to use telephone interviews in investigations of the general population. Telephone interviewing, of course, tacitly assumes that the telephone coverage is high (cf. Lepkowski, 1988). As mentioned above, at present telephone coverage in the Netherlands is high (92%). Still, there is some evidence that certain subpopulations (the unemployed, the elderly, students and young adults (18-25 years)) are relatively more difficult to reach by telephone because they are less likely to own one (Kerssemakers, De Mast & Remmerswaal, 1987). This can lead to biased estimates in telephone surveys, especially when these special groups are the target populations (cf. Snijkers, 1992).

In telephone interviews, as in face to face interviews, the Kish procedure can be used to select respondents within a household. Good alternatives for this rather complex procedure are the last or next birthday method (Oldendick, Bishop, Sorenson & Tuchfarber, 1988).

Mail surveys require an explicit sampling frame of names and addresses. Often, telephone directories are used for mail surveys of the general population. Using the telephone directory as a sampling frame has

the drawback that people without a telephone and people with an unlisted telephone cannot be reached. According to Snijkers (1992, p. 60) this type of noncoverage (no telephone or unlisted) is expected to be higher for the unemployed, the young, the elderly, divorcees, people in the low and high income brackets, and people with a low education. The reason for the frequent use of the telephone directory as sampling frame is the relative ease and the low costs associated with this method (Kalfs & Saris, 1991).

A drawback of mail surveys is the limited control the researcher has over the choice of the specific individual within a household who in fact completes the survey. There is no interviewer available to apply elaborate selection techniques and all instructions for respondent selection have to be included in the accompanying letter. As a consequence only simple procedures as the male/female/youngest/oldest alternation (cf. Dillman, 1978, p. 170; Lavrakas, 1987, p. 93-96) or the next birthday method (Oldendick et al., 1988) can be successfully used.

When a complete list of the individual members of the target population is available, which can be the case in surveys of special groups, a random sample of the target population can be drawn regardless of the data collection method used.

Nonresponse

Survey nonresponse is the failure to obtain measurements on sampled units. Nonresponse can be distinguished from another error of nonobservation, coverage error (discussed above), by the fact that nonrespondent units are selected into the sample, but not measured, whereas noncovered units have no chance of being selected in any sample (e.g., no known address, no telephone number), and thus cannot be measured (Groves & Lyberg, 1988).

Response rates can be influenced by many factors: the topic of the questionnaire, the length of the questionnaire, the survey organization, the number of callbacks or the number of reminders, and other design features (cf. Heberlein & Baumgartner, 1978). In this section I will only discuss so called "cold" surveys (i.e., surveys for which a fresh sample is drawn). Surveys that use a panel design or a "respondent pool" of respondents who are willing to participate in on-going research, will in general have a much higher response rate than cold surveys as the hard-core nonrespondents have already been filtered out.

Face to face surveys tend to obtain higher response rates than *comparable* telephone surveys. For instance, in a national comparison of face to face and telephone surveys in the U.S.A. Groves and Kahn (1979, p. 76) report a response rate of 74% for the face to face survey and of 70% for the corresponding telephone survey. Steeh (1981) reports an increase in refusal rate for the Consumer Attitude Survey when the data collection method changed from face to face to telephone interview. In 1975 (last full face to face survey) the response rate was 73% (refusal 15.5%, other nonresponse 11.5%), in 1977 (first full telephone survey) the response rate was 65.5% (refusal 26.9%, other nonresponse 7.6%). Goyder (1987) collected data on 385 mail surveys, 112 face to face surveys and 53 telephone surveys in the U.S.A. and Canada between 1930 and 1980. On average the response rate for the face to face interview was 67.3%, for the telephone interview 60.2%, and for the mailed questionnaire 58.4% (Goyder, 1987, p. 42).

But, nonresponse in face to face surveys appears to increase over the years. For instance, Goyder (1987, p. 67) notes a pronounced increase in nonresponse for the face to face interview, while the nonresponse for mail surveys remains stable. Steeh (1981) also reports an increase in nonresponse over the years on two large-scale American (face to face) surveys. This was mainly caused by an increase in refusal rates: in 1952 the refusal rate for the National Election Study was 6.6% and in 1975 it was 18.2%, the refusal rate for the Consumer Attitude Survey was 5.1% in 1952 and 15.5% in 1975 (Steeh, 1981, Table 1). The same trend is reported by Sugiyama (1992) for Japan.

In the Netherlands a rise in nonresponse has also been noticed. Bethlehem and Kersten (1981, 1986) report nonresponse rates for official government surveys implemented by the Netherlands Central Bureau of Statistics which range from 13% (Labor Force Survey) to 28% (Living Conditions) in the early seventies and from 18% (Labor Force) to 42% (Living Conditions) in 1983. At the Netherlands Bureau of Statistics (CBS) no large differences in overall nonresponse between telephone and face to face surveys have been detected (Kerssemakers, 1985). At present, the response rates for telephone surveys are slightly better than those for face to face surveys. This is attributed to the still increasing nonresponse for face to face surveys conducted by the Netherlands Central Bureau of Statistics, while their telephone surveys as yet do not follow this trend (Snijkers, 1992). Large marketing research firms in the Netherlands report approximately 40% nonresponse for telephone surveys (H. de Bock, personal communication, 11 december 1986). For mail surveys used in Dutch

marketing research nonresponse varied from 60% (car ownership) to 18% (health research) with an average nonresponse of 44.5% (Van Rooy, 1987).

Type of questions and complexity of questionnaire

Face to face interviews are the most flexible form of data collection method. Structured or partly structured questionnaires can be used, respondents can be asked to sort objects or pictures, and highly complex questionnaires can be used. Also, respondents can be presented with all kinds of visual stimuli, ranging from simple response cards with the response categories of a question to advertisement copy or video clips.

Telephone interviews are less flexible. Their major drawback is the absence of visual cues during the interview. Therefore, only questions with a limited number of response categories can be used. This has led to the development of special question formats (e.g., the two step or unfolding procedure) for questions with seven or more response categories, and verbal alternatives for graphically presented questions like the political "thermometer" (cf. Groves & Kahn, 1978, paragraph 5.1; Dillman, 1978, chapter 6). However, as in face to face interviews, the interviewer can assist respondents in understanding complex questions, can administer questionnaires with a large number of screen questions, control the question sequence, and probe for answers on open questions.

The absence of an interviewer makes mail surveys the least flexible data collection technique when complexity of questionnaire is considered. All questions must be presented in a fixed order, only a limited number of simple skips and branches can be used for which written instruction should be provided, and all respondents receive the same instruction without added probing or help in individual cases. But, visual cues can be used, and with well-developed instructions fairly complex questions and attitude scales can be used. The visual presentation of the questions makes it possible to use all types of graphical questions (e.g., ladder, thermometer), and to use questions with seven or more response categories. Also, information booklets or product samples can be sent by mail with an accompanying questionnaire for their evaluation.

Face to face interviews can last longer than either telephone or mail surveys. It takes a highly assertive respondent to end an overly long face to face interview, while this is much easier in a telephone and especially in a mail survey. As a rule, successful telephone surveys can be conducted with an average length of twenty to thirty minutes. Longer interviews will lead

to either a somewhat higher nonresponse rate (Collins, Sykes, Wilson, & Blackshaw, 1988, p. 229) or a higher probability of premature termination of the interview. Still, successful telephone interviews have been reported which took over 50 minutes (Frey, 1983, p. 48). Both Heberlein & Baumgartner (1978) and Goyder (1982) found a small negative effect of length of questionnaire on the response rates of mail surveys. According to Dillman (1978, p. 55) mail questionnaires up to 12 pages, which contain less than 125 items, can be used without adverse effects on the response.

Resources available: Time, organization and personnel

In general, telephone surveys are the fastest to complete, mail surveys are usually locked into a definite time interval of mailing dates with rigidly scheduled follow ups, and geographically dispersed face to face interviews take the longest time to complete. Each data collection technique requires, of course, that certain organizational requirements get met. Dillman (1978, p. 68) gives an example in which a survey unit of 15 telephones can complete roughly 3000 interviews during the 8 weeks it takes to do a complete TDM mail survey. When no permanent telephone survey laboratory or survey research center is available -a fairly common situation at Dutch universities- it takes considerably longer than 8 weeks to recruit and train interviewers, to apply for extra telephone connections, and to implement a telephone survey of 3000 interviews.

The implementation of a successful large scale face to face survey demands most from an organization and its personnel. Interviewers have to be trained, not only in standard interview techniques, but also in how to implement sample and respondent selection rules and solve various problems that can arise when they are alone in the field. In addition, an extensive supervisory network is needed to maintain quality control. Finally, an administrative manager is needed to make sure that new addresses and interview material are mailed to the interviewers on a regular base.

The personnel requirements for a telephone survey are less demanding. Because of the centralized setting, fewer highly trained supervisors are needed. Interviewers should, of course, be well trained in standard interview techniques. But, because of the close supervision the variety of skills needed is less. The majority of the interviewers no longer have to be prepared for every possible emergency and can concentrate on standard, but good quality interviewing. Difficult respondents or problem cases can be

dealt with by the available supervisor or can be allocated to a specially trained interviewer.

Organizational and personnel requirements for a mail survey are even less demanding. Most of the workers are not required to deal directly with respondents, and the necessary skills are mainly generalized clerical skills (e.g., typing, sorting, response administration, and correspondence processing). Of course, a trained person must be available to deal with requests for information, questions, and refusals of respondents. Finally, the number of different persons needed to conduct a mail survey is far less than that required for face to face or telephone surveys with equal sample sizes. For instance, one person can single-handedly successfully complete a TDM mail survey of a sample of 1000 persons in the prescribed 8 week TDM schedule.

1.4. Face to Face, Telephone, and Mail Surveys: Exchangeable alternatives or mutually exclusive choices?

In some cases the decision to use a particular data collection method is made easily because the alternatives are unrealistic or not practical for a particular study. Topic, type of questions, and type of respondent are extremely important factors in the decision process. For example, in a survey of the deaf special forms of self-administered questionnaires are very effective (cf. Breed & Swaans-Joha, 1986). In-depth face to face interviews of experts are necessary for the extraction of knowledge needed for building expert systems (cf. De Greef, Breuker & Wielinga, 1988; Kidd, 1986), while for the continuous monitoring of the media exposure and reading behavior of the Dutch population (Summoscanner) telephone interviews are an optimal choice (cf. De Bock, 1987).

When viable alternatives exist, the choice between modes of data collection is usually guided by factors such as the available organizational infrastructure, the estimated costs, the predicted nonresponse rate, the length of the data collection period, and especially the expected level of measurement error or data quality (Lyberg & Kasprzyk, 1991; Groves, 1989).

Issues of measurement error are not only important when a choice between modes has to be made, but are also extremely important when data, collected with different methods, are combined in one study. "Mixed mode" surveys are being used with increasing frequency throughout the world (Dillman, 1991). Mixed mode survey designs try to take advantages of

the best features of each mode. An example of such a mixed mode strategy is a panel survey in which face to face interviews are used in the first wave and telephone interviews or mail questionnaires in subsequent contacts, thereby lowering survey costs and maintaining an adequate response rate (Kalton, Kasprzyk & McMillen, 1989). Another application of a mixed mode strategy occurs when different modes are used to collect data from different respondents within a sample. Typically, one main data collection mode (e.g., a mail survey) is used to its maximum potential. Then another method (e.g., a face to face or telephone interview) is adopted to increase response rates. An overview of different types of mixed mode surveys is given by Dillman and Tarnai (1988).

The use of mixed mode surveys is stimulated by attempts to reduce costs and to improve response rates. However, combining the data derived by different methods raises the question whether these data are comparable. Do people really respond in the same way to questions posed by means of a different method?

The availability of alternative methods for the rather expensive face to face survey and the growing interest in mixed mode surveys has prompted a long line of comparative research on data collection methods. This book follows in this tradition. It provides both a systematic overview of reported differences between mail, telephone, and face to face surveys, and the results of a controlled field experiment conducted in the Netherlands.

1.5. Outline of this Book

In this book the emphasis is on data quality in mail, telephone and face to face surveys. I concentrate on those cases where the different modes can be viewed as viable alternatives to each other, although each method has its own potential strengths and weaknesses. The purpose of this study is to examine the effects of data collection methods (i.e., mail, telephone, and face to face surveys) on the quality of the resulting data and on substantive conclusions based on those data.

In chapter 2, "Why expect differences," I give an overview of factors that may cause mode effects. This overview is based on general expectations encountered in the literature on survey methods.

In chapter 3, "Empirical evidence of mode effects; a meta-analysis," I present the results of a quantitative literature review of a large number of empirical studies on mode differences.

The results led to the design and implementation of a mode experiment in the Netherlands. In chapter 4, "Design of a field experiment," I describe how the questionnaire was designed and pre-tested; I discuss the results of a pilot study and present the design of a large field experiment.

Chapters 5, 6, and 7 are devoted to in-depth analyses of response differences between the three modes. In chapter 5, "Data quality I: a replication in the Netherlands," I compare the results from the field experiment with the findings from the meta-analysis in chapter 3 and with expectations based on the review in chapter 2.

In chapter 6 and 7 I extend the analyses, using new criteria for data quality that were not available in previous mode comparisons. In chapter 6, "Data quality II: reliability and scalability," I employ psychometric criteria, concentrating on reliability and scalability of multiple item scales. In chapter 7, "Data quality III: a multivariate approach," I investigate the influence of data collection method on the relationships between variables. Two substantive models about the multivariate relationships between variables -one on loneliness and one on subjective well-being- are investigated.

Finally, in chapter 8, "Conclusion," I provide a critical summary of the findings and discuss future directions of survey research.

CHAPTER 2

WHY EXPECT DIFFERENCES?

I think, for example, that it is a law that the irradiation of green plants by sunlight causes carbohydrate synthesis, and I think that it is a law that friction causes heat, but I do not think that it is a law that (either the irradiation of green plants by sunlight or friction) causes (either carbohydrate synthesis or heat).

J.A. Fodor, Representations, 1981, p. 40

2.1. Introduction

In 1944 Deming published one of the first reviews on errors in surveys, which identified thirteen factors threatening the usefulness of surveys. One factor named is "shifting modes of data collection while the study is in progress." In 1965 Kish presented a comprehensive taxonomy for the classification of error within survey statistics in which data collection method is explicitly named as a source of non-sampling error. In his 1989 book on survey errors and survey costs Groves distinguishes four main sources of measurement error: interviewers, respondents, questionnaires and mode of data collection.

For more than forty years the data collection method has been considered a potential source of error and researchers have been concerned about possible differences in answers due to effects of mode of data collection. Why do they expect differences?

In the literature on mode effects several factors are identified as differentiating face to face, telephone, and mail surveys from each other. These factors provide a priori expectations for the existence of mode differences. They can be grouped in three main classes: differences due to media related factors, differences in information transmission, and differences in interviewer impact. An overview of the factors that differentiate the modes of data collection from each other will be presented in this chapter.

Some factors discussed in this chapter are more important for certain indicators of data quality than other factors. Furthermore, as neither the magnitude of the effect of the various factors nor the way they interact is

known, it is difficult to specify a final mode effect. Therefore, no detailed predictions about mode differences on specific indicators will be formulated in this chapter. In specific cases it may be possible to formulate predictions. These will be presented later at their appropriate places.

2.2. Media Related Factors

Face to face, telephone, and mail surveys differ on a number of factors that are inherent to the social conventions associated with the medium of communication.

The *first* media related difference concerns the degree to which people are *acquainted* (i.e., knowledgeable of and familiar) with the media concerned. People are used to all kinds of face to face interactions in which information is being gathered, for example conversations with medical doctors, teachers, and supervisors (Kahn and Cannell, 1957). Face to face contacts in surveys are therefore seen as appropriate and have acquired a place in society.

The first use of the telephone was as an instrument of business for short communications (PTT, 1989). Later, the telephone became an instrument for private conversations with family and friends, enabling people to maintain close contacts over larger distances (Körmendi & Noordhoek, 1989, p. 9). Social customs concerning this private use still differ between cultures. In the United States the telephone is used extensively for both business and friendship contacts (Groves, 1989, p. 510). In Japan the content of the message and the status of the other party determine the choice for a specific means of communication. For instance, for a request face to face talks are preferred for relatives and superiors, while the telephone is used for subordinates (Akuto, 1992). Another example of cultural differences in telephone usage can be found in some Eastern European and third world countries (cf. Zoon, 1992), where the unreliability of the telephone system has led to a specific way of telephone communication (e.g., a tendency to use short messages and to speak in a loud and distinctive tone).

In several countries in Western Europe (e.g., the Netherlands, Germany, France), the more private use of telephones is now being propagated by widespread advertising campaigns, showing happy grandparents phoning their grandchildren, friends discussing their adventures while on holiday, picturing the telephone cable as a "lifeline." Still, telephone calls received at home from strangers are more typically

expected to be for a business purpose than for an exchange of personal information.

The medium for mail surveys is the self-administered form. Most people in our society are familiar with administrative forms, school tests, or tax forms. However, completing these types of self-administered forms is not the most exciting or pleasant thing to do. Also, the completion of self-administered forms demands a relative high level of active command of a language. People feel more compelled to avoid grammatical errors in written communications, which can inhibit the freedom of expression.

The *second* media related factor concerns the *locus of control* during the data collection. In a face to face interview both respondent and interviewer share the locus of control. As initiator of the conversation the initiative is given to the interviewer, but the social rules of good behavior during a personal visit prescribe that the pace of the interview and the communication flow is determined by both parties involved. In a telephone interview the interviewer is more in control. First, the ringing of a telephone immediately creates a sense of obligation to answer it, and people often interrupt a face to face conversation to answer a ringing phone. Second, traditional rules of behavior dictate that the initiator of a telephone conversation, here the interviewer, controls the channel and the regulation of the communication flow (Argyle, 1973; Körmendi & Noordhoek, 1989). In a mail survey the respondent is in total control of the situation and determines when and where the questions are being answered. This gives the respondent the opportunity to complete the form at a considered pace, to look up information at leisure, and consult other members of the household when proxy information about household members is being asked (Lyberg & Kasprzyk, 1991). Furthermore, in a mail survey the respondent and not the interviewer writes down the answer, which gives an extra check on the correctness of the recorded answer and emphasizes the total control of the respondent on the pace of the question-answer sequence (cf. Galtung, 1967).

The *third* media related factor concerns the social conventions regarding the acceptability of *silences* in a conversation. This factor sharply distinguishes the face to face interview from the telephone interview. There is a marked tendency to avoid silences in a telephone conversation, and long silences over the telephone are considered improper and rude.

The *fourth* and last media related factor refers to the ability of the medium to convey *sincerity* of purpose. The personal contact in a face to face situation gives an interviewer far more opportunities to convince a respondent of the legitimacy of the study in question. A telephone

interviewer, without any means of identification, has far less chances to communicate trust and legitimacy. A mail survey can use a logo, a valid return address, and other visual means to emphasize the trustworthiness of the survey. Furthermore, mail surveys do not have to be answered immediately and offer the respondent the possibility to check out the survey organization.

2.3. Information Transmission

Face to face, telephone, and mail surveys differ markedly in the way in which information is transmitted. In this section the emphasis is on the more technical aspects of information transmission and not on social customs as discussed in 2.2.

The *first* difference concerns the *communication channels* used (Sykes & Hoinville, 1985). Three types of communication can be distinguished: verbal communication, nonverbal communication, and paralinguistic communication. Verbal communication is only concerned with the spoken words, non verbal communication is concerned with the meaning of gestures, expressions and body posture, and paralinguistic communication is concerned with (non verbal) auditive signals, like emotional tone, timing, emphasis, and utterances like "mhm-hmm" (cf. Argyle, 1973). In face to face interviews all three channels of communication can be used to transmit information between respondent and interviewer. Telephone interviews have a more limited channel capacity; only verbal and paralinguistic means of communication are available in telephone conversations. The absence of a channel for nonverbal communication makes the transmission of all kinds of information harder for both interviewer and respondent. In mail surveys all information is conveyed by the printed word and the above distinction in three different types of communication is not appropriate. But, it should be noted that the lay-out of a questionnaire and the use of graphical devices and illustrations can partly take over the role of the nonverbal and paralinguistic channels to add extra emphasis to a text or to clarify parts of a text.

The *second* important difference concerns the *presentation of the stimuli* (Schwarz, Strack, Hippler & Bishop, 1991). Stimuli can be presented *visually or auditorily*. In mail surveys the items and response alternatives are visually displayed to the respondent who has to read the questionnaire. In telephone interviews the items and the response alternatives are read aloud to the respondent who has to listen to what is

said. In face to face interviews both types of presentation -visual and auditory- may occur. For instance, response cards can be used when many different response alternatives are presented, thereby making the task easier for both respondent and interviewer.

Another distinction in the presentation of stimuli refers to the *temporal order* in which the material is presented (Schwarz et al., 1991). Face to face and telephone interviews have a sequential organization. The stimuli are presented in temporal succession and respondents cannot go back and forth between the questions. In general, backtracking to a previous question makes the task for interviewers harder, especially with complicated questionnaires that use many different routings, and is therefore not encouraged by interviewers. But, even if respondents are allowed to correct their answers to previous questions, they seldom do so. In face to face and telephone interviews tracking one's previous responses is a difficult memory task indeed. In contrast, keeping track of one's responses and going back and forth between questions is not difficult at all in a mail survey. Furthermore, as mentioned in section 2.2, the locus of control in a mail survey is the respondent, who can use as much time as she or he wishes to work on a questionnaire.

The *third* difference in information transmission between the face to face and telephone survey is the *regulation of the communication flow* between interviewer and respondent. Sykes and Collins (1988) emphasize the importance of nonverbal cues for channel control (to determine turntaking) in face to face interactions. Argyle (1973, p. 72) points out that channel control is an important factor to make verbal exchanges possible. "Interactors have to take it in turns to speak and listen, and speech itself cannot be used to decide who shall speak or for how long . . . channel control is effected by small non-verbal signals, mainly head-nods and eye movements. These signals are presumably learnt." In telephone conversation mainly paralinguistic cues are used to regulate the communication flow. For instance, prolonged silence means "your turn," and mhm-hmm means "continue, I am listening to you." Also, contrary to the custom in face to face interactions, explicit spoken signals are allowed in a telephone conversation. For instance, in a telephone conversation, an explicit "Yes" or "Okay" replaces the nonverbal polite smile or nod. This custom may go back to the early days of telephone communication, when an operator on request made contact with another telephone subscriber. The operator then used a special phrase to indicate that the telephone

conversation could start[1]. In mail surveys no explicit turntaking takes place. The respondent is the locus of control over the information flow and can decide when to stop or to continue the question-answer process.

2.4. Interviewer Impact

The modes of data collection clearly differ in how much they restrict interviewer impact. In mail surveys the interviewer is absent and can not play a role -either positive or negative- in the question-answer process. In telephone interviews, which have a limited channel capacity (see 2.3), interviewers have potentially less impact on respondent behavior than in face to face interviews.

First, the potential *positive* influence of interviewer impact on survey responses will be reviewed. Interviewers have several responsibilities during the interview: they have to motivate respondents, to deliver and when necessary clarify the questions, to answer the respondent's queries, and to probe to clarify answers. In face to face interviews the interviewer could use nonverbal cues to motivate the respondent, and keep the flow of information going. Furthermore, the interviewer could monitor the respondent's nonverbal expressions and react to those. In telephone interviews these tasks are more difficult; nonverbal communication is impossible and interviewers must be alert to attend to paralinguistic information. But, both in telephone and in face to face surveys an interviewer is present to answer questions and give additional information. In mail surveys the respondent is solely dependent on the questions as stated and on the written instructions in the questionnaire and the accompanying letter.

Second, possible *disadvantages* of interviewer impact will be reviewed. The limited impact of the interviewer in telephone surveys can also have a positive influence on the respondent. The interviewer is only a voice over the phone. The respondent is less restricted in his/her "personal space" and can be more relaxed (cf. Argyle & Dean, 1965). In face to face surveys, respondents often fall back on the "receiving a guest script" and their self-imposed role as host will influence their reactions (cf. Groves, 1989, p. 510). The total absence of an interviewer in a mail survey allows the

[1] In Amsterdam around 1881 the telephone operator said the prescribed words 'voorwaarts, mijnheer' (Forward, sir) to indicate that the party that requested the call could start with the telephone conversation (PTT, 1989, p. 82).

respondent even more personal space than a telephone interview and may introduce a greater feeling of anonymity in the respondent (Cannell & Fowler, 1963). The more anonymous and private setting in which mail surveys are completed, reduces the tendency of respondents to present themselves in a favorable light and induces fewer problems of self-presentation (Sudman & Bradburn, 1974).

Interviewer impact may also influence responses through the interviewers themselves. Interviewers affect respondent's answers in a way similar to the clustering effect in sampling (Lyberg & Kasprzyk, 1991). The interviewer effect increases the total variance of the statistics under study (Kish, 1965, 1987; O'Murcheartaigh, 1977) and the measurement of interviewer effects has been given considerable attention over the years (Dijkstra, 1983; Groves & Magilavy, 1986; Kish, 1962). The restricted channel capacity of the telephone interview gives interviewer characteristics less chance to influence respondents. Furthermore, the central setting of telephone interviews allows for a stricter control over interviewers and thereby for a possible reduction of interviewer related error (cf. Fowler, 1991).

2.5. Summary

In this chapter a systematic overview was given of the potential influence of mode related factors on survey measurements. These factors have been ordered in three classes: 1) media related factors, 2) factors influencing information transmission, and 3) interviewer impact. Media related factors are concerned with the social conventions and customs associated with the media utilized in survey methods. Under the heading information transmission more technical aspects of the communication process are described (e.g., channel capacity, regulation of information flow). Interviewer impact is concerned with the degree in which interviewers can -positively or negatively- influence the question-answer process. Figure 2.1 on the next page presents an overview of the factors influencing data quality.

The mode of data collection can by a variety of factors influence survey results. It is, however, difficult to predict how large the final mode effects will be. The magnitude of the effects of the various factors is unknown and certain factors may interact to produce a final mode effect (e.g., channel capacity and interviewer impact) or add up or counteract each other (various aspects of interviewer impact). Without detailed a priori

knowledge, one has to rely on empirical results to supplement the theoretical expectations. Therefore, a meta-analysis was conducted on the existing empirical research on mode differences. The results of this meta-analysis are discussed in chapter 3.

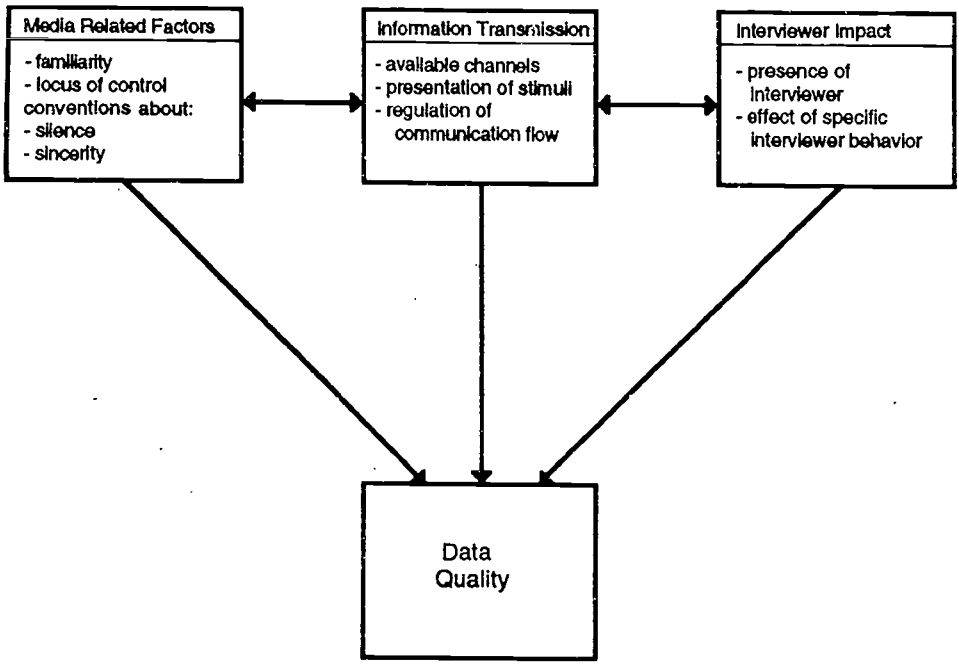


Figure 2.1 Conceptual Model of Data Collection Effects on Data Quality

CHAPTER 3

EMPIRICAL EVIDENCE OF MODE EFFECTS: A META ANALYSIS

I have got the works of all the old masters. I weigh them against each other - balance the disagreements - analyze the conflicting statements - decide which is probably correct - and come to a conclusion. That is the scientific method. At least as I see it.

Cf. Isaac Asimov, Foundation, 1971, p. 53

3.1. Introduction

In the last two decades an increasing number of empirical studies have been published on the influence of survey method on data quality. Most of these studies were prompted by the practical and important question: "What will happen to the quality of the data when we change our major data collection method?" This resulted in mode comparisons in which usually two alternative systems of data collection (e.g., face to face versus telephone survey) were compared on a limited number of quality indicators, which were of direct practical importance for a specific survey or series of surveys.

This chapter discusses the results of previous mode comparison studies. Principles of meta-analysis are used to integrate research and to provide a systematic overview of empirical findings on differences in data quality between mail, telephone, and face to face surveys. This method makes it possible to answer the following two research questions:

1. Do previous studies provide evidence for the existence of mode effects, that is, systematic differences between data collected by means of mail, telephone, and face to face surveys?
2. If mode effects are found, how large are the differences?

In this chapter I will first describe the methods used (section 3.2). In section 3.3 the results are presented for differences in response rate, followed by the results concerning differences in data quality. The chapter ends with a summary of the main results (section 3.4). Appendix A contains a bibliography of mode comparisons.

3.2. Method

On meta-analysis

Though the name meta-analysis deceptively suggests otherwise, meta-analysis is not one method or one type of analysis. Meta-analysis or integrative analysis, as it is often called, is a coherent set of quantitative methods for reviewing research literature (cf. Glass, McGaw & Smith, 1981; Light & Pillemer, 1984; Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1984). The primary aim of meta-analysis is inferring non-causal generalizations about specific substantive issues from a set of studies directly having a bearing on those issues (Jackson, 1980). To achieve this, quantitative study outcomes from known research on a particular, well defined question are statistically combined.

In general, an effect size measure is coded for the dependent variable *study-outcome*. Furthermore, background variables such as year of publication and source of publication are routinely coded, just as age and sex are routinely asked in a survey. Also, several research design characteristics of each study are coded (e.g., sampling method, type of subjects). This coding process results in a data matrix in which the cases (or rows) are the research studies of interest for the meta-analysis. Standard statistical procedures can then be used.

In other words, the basic idea is to apply statistical methods, with the published statistics from previous studies as the data (Walberg & Haertel, 1980). This use of systematic statistical procedures together with a clear description of the retrieval of relevant studies and of the methods used, distinguishes meta-analysis from the more traditional, narrative forms of literature review (Bangert-Drowns, 1986).

Retrieval and selection of studies

First, an on-line computer search was conducted. The abstracting services used were: Psychological Abstracts (1967-1986), Sociological Abstracts (1963-1986), Dissertation Abstracts (1861-1986), and Dialog/SSCI (Social Sciences Citation Index, 1972-1986). The following key-words were used, both singly and in combination: artifact, bias, comparison, data collection method, face to face, interview, mail, personal, postal, response, response bias, response effect, response style, social desirability, survey, telephone.

Most studies found (81%) were conducted in the USA. This could partly be a result of the data bases available for the computer search. To avoid retrieval bias, an appeal for research articles was published in three European newsletters. In addition, and to update the results of the first search, the on-line database of SRM was searched for the period 1979 to 1990. SRM is a documentation service in the field of social research methodology based in the Netherlands, which publishes abstracts (in English) from more than 100 selected American and European journals. The reference lists of the studies found were searched to uncover additional material.

In this review differences of data quality between mail, telephone and face to face surveys are discussed. Therefore, only articles that empirically investigated the influence of these modes of data collection on the quality of the data were included. Studies of only response rates were not included. Also studies that reviewed past literature, or reported a reanalysis of already known data, without presenting any new data, were not included in the meta-analysis.

In total 67 articles and papers were found (for a concise summary, see Appendix A.2). Three articles contained a reanalysis of earlier studies, and one article had very severe design flaws. These articles were excluded (cf. Wortman, 1983; Wortman & Bryant, 1985). Ten articles did not report enough details (e.g., no sample size or no information on experimental groups) for coding and computation of effect size. In cases where studies were (partly) reported in more than one article or paper the information from separate articles was combined to avoid non-independence between the cases in the statistical analyses (Bangert-Drowns, 1986; Rosenthal & Rubin, 1986). This resulted in a total of *52 studies* available for analysis.

Twenty-six different journals in the domains of psychology, sociology, marketing and opinion research, medicine, and criminology provided the relevant literature. The oldest reference was published in 1947, the most recent one in 1990. A variety of topics were covered with health issues the most prominent.

Coding of the studies

An extended version of the coding schedu of Sudman and Bradburn (1974) was used. Included were background variables relating to the research report (e.g., journal, year and country of publication), and the study itself

(e.g., type and size of sample, subject of the research and its saliency for respondents, equivalence of samples and questionnaires used in the study).

For each data collection method in the study the response rate was coded. Response rate was defined as the number of completed interviews divided by the total number of eligible sample units (Groves & Kahn, 1979; Kviz, 1977). Five indicators of data quality were used, reflecting the multivariate nature of this complex concept (cf. Bailar, 1984).

The most direct measure of data quality is response validity. Here the answer of the respondent is checked against the "true" value as found in official records. The use of this indicator is, of course, restricted to those factual questions for which validating information is available (Biemer, 1988). In all other cases, especially in studies of subjective phenomena (attitudes, beliefs or other attributes that cannot be observed directly), there is no direct way to assess the correctness of the answers. In these cases, various proxy variables for the quality of the data have been used (Groves, 1978). As a result, a variety of different indicators of data quality can be found in empirical mode comparison studies.

To make a useful selection of the indicators used in the literature, a content analysis was conducted on a subsample of 20 articles. Only those indicators for data quality used in at least two studies were retained and coded for the meta-analysis.

These indicators are:

(1) response validity, the answer is checked against information from official records;

(2) item nonresponse, also called item missing data rate;

(3) the number of statements made in response to an open-ended question;

(4) social desirability; and

(5) similarity of responses on closed questions, indicated by no difference between the proportion obtained under different modes.

The last two indicators need further explanation. For both indicators responses on a closed question are compared over data collection modes. In other words, for a closed question the response distributions are compared across modes. In that sense, both indicators are measures of the (dis)similarity of the answers between modes. In the case of social desirability, however, *sensitive* questions were explicitly included for comparison by the original authors of the articles, who also made inferences regarding the relative quality of the answer (i.e., "better" or "less socially desirable"). For example, when respondents in a mail survey report more drinking behavior than respondents in a face to face survey, this is often interpreted as a smaller effect of social desirability in the mail survey

condition (cf. De Maio, 1984). What was socially desirable was decided by the original authors of the articles and **not** by the coders for the meta-analysis.

For the indicator "similarity of responses" such assumptions cannot be made. This indicator only expresses whether statistically significant differences between the estimates obtained from different methods do exist, and is as such not an indicator of data quality. However, the absence of statistically significant differences between the estimates from two surveys indicates that both estimates have the same bias, which of course may be zero (cf. Biemer, 1988). A result that is of great practical importance for survey reseachers.

As an estimate for effect size the product moment correlation coefficient r was chosen, for the reasons outlined by Rosenthal (1984, pp. 23-24), which include ease of calculation and simplicity of interpretation. The product moment correlation coefficient provides a convenient gauge of effect size with the square of the correlation indicating the proportion of variance explained by mode. The main reason for preferring the product moment correlation coefficient is analytical; in the original articles a large variety of statistical tests were used, some parametric, some nonparametric. In the literature on meta-analysis, methods are available to convert this variety of test statistics *accurately* to a product moment correlation (cf. Hunter & Schmidt, 1990; Hunter, Schmidt & Jackson, 1982; Rosenthal, 1984; Wolf, 1986). For the indicators "response validity," "item non-response," "number of statements to open questions" and "social desirability" a directional coefficient was coded, indicating which data collection method was best. For the indicator "similarity of response" this was, of course, not possible.

In meta-analysis the unit of analysis is a study. Therefore, whenever an article reported the results for more than one study, each study was coded as a separate case (Bangert-Drowns, 1986; Rosenthal & Rubin, 1986). Likewise, when a particular study used more than one measure of the same indicator, effect sizes were combined by computing the mean correlation prior to the coding (e.g., when in one study item nonresponse was reported for five questions the mean effect size of these five questions was coded to represent the study's item nonresponse). A weighted mean was used in the case where sample sizes differed considerably between questions within the same study as a result of skipping or branching. This procedure results in one effect size estimate for each indicator in a study and the basic assumption of independence is not violated (Hunter & Schmidt, 1990; Wolf, 1986).

The studies were independently coded by two trained coders, using the same detailed coding-book. As a gauge for the intercoder reliability the product moment correlation between the data sets of the two coders was computed. The overall intercoder reliability was .93.

Several of the coded variables concerned facts that could be looked up (e.g., year of publication, type of sample, response rate); other variables to be coded required a more subjective estimate of the coders (e.g., saliency of topic, question threat). It is to be expected that there will be hardly any disagreement between conscientious coders on the factual variables, and this could inflate the estimated intercoder reliability. As a lower boundary for the intercoder reliability, the correlation between the estimates of the two coders for the subjective variables only was computed. The intercoder reliability for these subjective variables was .77.

Analysis

The effect sizes were combined over studies for each dependent variable (i.e., each indicator of the multivariate concept data quality) separately (Bangert-Drowns, 1986; Wolf, 1986). To summarize findings over studies the statistical procedures described in Hedges and Olkin (1985, pp. 223 - 232) were applied. A weighted estimate for the correlation and the corresponding 95-percent confidence interval were computed, using the Fisher z-transformation. The weights were based on sample sizes (Hedges & Olkin, 1985, p. 231). In addition, a homogeneity test (Q) was performed to detect the presence of possible moderator variables (Hedges and Olkin, 1985: pp. 234 - 244). Q indicates whether the weighted effect sizes are sufficiently different from each other to reject the null hypothesis that they are drawn from a common population. If this null hypothesis is rejected, it would be misleading to summarize effect sizes in a single effect size estimate. For example, a significant effect in favor of face to face interviews in half of the studies and a significant effect in favor of telephone interviews in the other half would lead to an overall (nonsignificant) effect size near zero. The homogeneity test is designed to detect this kind of situation in which the underlying population effect sizes are heterogeneous (Hedges & Olkin, 1985, p. 147). Statistical analyses were run for each pairwise comparison (face to face versus telephone, telephone versus mail, and face to face versus mail).

3.3. Results

Response rate

The mean response rates for the three data collection methods differ significantly. Overall, face to face interviews produce the highest response rate and mail survey the lowest. For the face to face interview, a mean response rate of 75 percent is reported in the articles studied, versus a mean response rate of 71 percent for the telephone interview, and 68 percent for the mail survey. Both an overall significance test and all pairwise comparisons were significant at the .01 level.

These differences are in accordance with the review of Goyder (1985), who reports an estimated net response difference between interview and mail surveys of 7.5 percentage points for surveys with response rates between 30% and 70%.

For all three data collection methods the average response rates reported in the mode comparisons are rather high compared with average response rates in general (see chapter 1, section 1.3). This reflects the care taken to reduce nonresponse bias in most mode comparisons. In general, the quality of the studies was high and call-backs and mailed reminders were used to increase the response rates.

The average number of respondents reported in the studies is 1394; the smallest number of respondents reported in a study is 64, the largest number is 6000.

Face to face and telephone surveys compared

Differences in data quality between face to face and telephone surveys are quite small. The largest overall effect found is for similarity of responses on closed questions (r=.05). This falls within Cohen's definition of a small effect size (Cohen, 1969, p. 76).

The indicators "response validity" (record check), and "social desirability" did not show statistically significant differences. Small, but statistically significant differences in data quality between face to face and telephone interviews were detected for the indicators "item nonresponse," "number of statements to open questions," and "similarity of response distributions on closed questions". The face to face interview performed slightly better than the telephone interview. Face to face surveys resulted in slightly less overall item nonresponse (weighted mean correlation: -.02)

and in slightly more statements in response to open questions (weighted mean correlation: -.04).

Table 3.1 summarizes the results. In most comparisons only one or two indicators of data quality were used. As a consequence, the data points for each indicator are limited and differ in number.

Table 3.1 Comparison of Face to Face and Telephone Surveys

Mean weighted product moment correlation (negative means in favor of face to face interviews, positive means in favor of telephone interviews), 95% confidence interval, range (in parentheses), p-value, and number of comparisons in the analysis.

Indicator	Mean r	Confidence interval (Range)	P-value	N
Response validity	+.01	-.02/+.03 (-.04/+.10)	.69	10
Item non-response	-.02	-.03/-.01 (-.08/+.02)	.00	11
# statements to open questions[a]	-.04	-.07/-.02 (-.24/+.09)	.00	4
Social desirability[a]	-.01	-.03/+.01 (-.15/+.08)	.22	14
Similarity of responses	.05	.03/.06 (.03/.08)	.00	6

Note. For response validity, item nonresponse, number of statements to open questions, and social desirability a directional correlation was coded, indicating which data collection method was best. For the indicator similarity of responses on closed-ended questions this was not possible and these results are presented without a sign.

[a] For these indicators the homogeneity test was significant: the null hypothesis that the effect sizes were drawn from a common population was rejected at the .01 level.

The homogeneity test was not significant for the indicators "response validity," "item non-response," and "similarity." Only, for the indicators "social desirability" and "number of statements to open questions" did the homogeneity test Q (Hedges & Olkin, 1985) indicate that the underlying population of effect sizes is not the same for all of the studies, implying the

influence of possible moderator variables. The homogeneity test was significant at the .01 level.

For the indicator "social desirability," further analyses were possible to identify moderator variables. In a previous meta-analysis De Leeuw and Van der Zouwen (1988) found a small but statistically significant effect for social desirability in favor of face to face surveys, which was diminishing over the years. Groves (1989) pointed out that in recent mode comparisons in general no differences were found on sensitive items. When year of publication is incorporated in the analysis, an interesting pattern emerges. The nine studies published before 1980 show a small, significant effect (p=.03), indicating less social desirability in face to face interviews. The weighted mean product moment correlation for these early studies is -.03; the lower limit of the corresponding 95%-confidence interval is -.06, the upper limit is -.00. The five studies published after 1980 showed no difference in social desirability (p=.79). The weighted mean correlation is .00; the 95%-confidence ranges from -.02 to +.03. Although the year of publication did not explain the heterogeneity completely, further analyses with the available independent variables did not reveal any additional moderator variables.

Unfortunately, for the indicator "number of statements to open questions" the limited number of studies (4) available prevents any further detailed analysis.

Sometimes additional indicators for data quality were reported. For instance, Jordan, Marcus and Reeder (1980) compared response styles in telephone and face to face interviews. They found more acquiescence, more evasiveness, and more extremeness in the telephone interview. A tendency for the telephone respondent to choose the more extreme point on a scale was also noted by Groves (1979). This result is partly corroborated by Dillman and Mason (1984, p. 26) who investigated extremeness bias and report that "there is some evidence to support the telephone extremeness response . . . , but it is neither strong nor completely consistent." Aspects of psychometric reliability have been investigated by several authors. Aneshensel, Frerichs, Clark and Yokopenic (1982) found no differences between modes in the internal consistency (Cronbach's alpha) of a multiple-item depression scale. For consistency of an answer over time (test-retest reliability), no differences between telephone and face to face interviews have been uncovered in three separate studies (Herman, 1977; O'Toole, Batistuta, Long & Crouch, 1986; Rogers, 1976).

Mail and interview surveys compared

While the differences between the two interview modes were quite small, those between mail surveys and both types of interview surveys were somewhat larger. None of the studies investigated used "number of statements to open questions" as an indicator for data quality. No statistically significant differences could be detected for "response validity" (record check). For "social desirability" the differences favor the mail survey. Mail surveys resulted in fewer socially desirable answers on sensitive questions than face to face surveys: the mean weighted product moment correlation is +.09. Compared with telephone surveys, mail surveys also resulted in less socially desirable answers. There the mean weighted correlation is +.06. The results are summarized in Table 3.2 and Table 3.3.

Table 3.2 Comparison of Mail and Face to Face Interview Surveys

Mean weighted product moment correlation (negative means in favor of face to face interviews, positive means in favor of mail surveys), 95% confidence interval, range (in parentheses), p-value, and number of comparisons in the analysis.

Indicator	Mean r	Confidence interval (Range)	P-value	N
Response validity	+.03	-.00/+.07 (-.02/+.12)	.08	6
Item non-response[a]	-.03	-.05/-.01 (-.19/+.08)	.01	8
Social desirability	+.09	+.07/+.11 (-.06/+.29)	.00	13
Similarity of responses	.08	.05/.11 (.01/.21)	.00	8

Note. For response validity, item nonresponse, and social desirability a directional correlation was coded, indicating which data collection method was best. For similarity of responses on closed-ended questions this was not possible and these results are presented without a sign.

[a] For this indicator the homogeneity test was significant: the null hypothesis that the effect sizes were drawn from a common population was rejected at the .01 level.

The homogeneity test was only significant for the indicator "item non-response," indicating the presence of possible moderator variables. The test was significant at the .01 level for the comparison of mail and face to face surveys and the comparison of mail and telephone surveys.

Table 3.3 Comparison of Mail and Telephone Interview Surveys

Mean product moment weighted correlation (negative means in favor of telephone interviews, positive means in favor of mail surveys), 95% confidence interval, range (in parentheses), p-value, and number of comparisons in the analysis.

Indicator	Mean r	Confidence interval (Range)	P-value	N
Response validity	+.02	-.03/+.07 (-.02/+.03)	.40	4
Item non-response [a]	-.01	-.03/+.02 -.14/+.09)	.56	5
Social desirability	+.06	+.03/+.09 (+.04/+.17)	.00	5
Similarity of responses	.12	.08/.16 (.09/.28)	.00	3

Note. For response validity, item nonresponse, and social desirability a directional correlation was coded, indicating which data collection method was best. For similarity of responses on closed-ended questions this was not possible and these results are presented without a sign.
[a] For this indicator the homogeneity test was significant: the null hypothesis that the effect sizes were drawn from a common population was rejected at the .01 level.

For "item non-response" the differences favor the face to face interview: face to face interviews resulted in less item nonresponse than mail surveys (r= -.03). The *overall* difference in item nonresponse between telephone and mail surveys did not reach statistical significance. However, the homogeneity hypothesis was rejected for the effect size measures on the indicator "item non-response," indicating the influence of moderator variables.

In the research literature it has been noted that when respondents are asked about sensitive topics like income, self-administered questionnaires produce less item nonresponse, but that the opposite is found when non-

sensitive questions were asked (Nuckols, 1964; Siemiatycki, 1979; Van Sonsbeek & Stronkhorst, 1983). This suggests that sensitivity of topic may serve as a possible moderator variable. When the data on item nonresponse were reanalyzed, excluding the data on income, the resulting weighted mean correlation is decidedly more negative. For comparisons of face to face and mail surveys (7 studies), the weighted mean correlation for item nonresponse is then -.06 (p=.00); the corresponding 95-percent confidence interval ranges from -.08 to -.04. For comparisons of telephone and mail surveys (4 studies), the weighted mean r is also -.06, and does now reach statistical significance (p=.00). The lower limit of the 95-percent confidence interval is -.09, the upper limit -.03. Note that the overall weighted mean correlation for the comparison between telephone and mail surveys for item nonresponse was -.01, which was not statistically significant. Sensitivity of topic acts as a suppressor variable and completely explains the heterogeneity found. The lesser item-nonresponse on income questions in mail surveys obscures the basic finding that in general respondents in **both** (i.e., face to face and telephone) interview modes show less item nonresponse than in mail surveys. When very sensitive questions like income are asked, this relationship no longer exists, and mail surveys can even show less item nonresponse on the income question.

Returning to the individual studies, I note that sometimes additional indicators for data quality have been reported. When I take these into consideration, an interesting pattern emerges. It is harder to have people answer questions in a mail survey. Both the overall nonresponse and the item nonresponse tend to be higher in mail surveys. But when the questions are answered in mail surveys, the resulting data are of higher quality, and well-known response effects are less influential. For instance, Bishop, Hippler, Schwarz and Strack (1988) found in two cross-culturally replicated experiments that order effects are significantly less likely to occur in a mail survey than in a telephone survey; but question wording and question form effects were just as likely to occur in both methods. These results were partly replicated by Ayidiya and McClendon (1990), who with one exception did not find question order effects in mail surveys.

Finally, two of the articles coded provide additional information concerning the extremity of responses. Both studies indicate a higher preference of respondents in both face to face and telephone interviews for the positive end of a response scale. Dillman and Mason (1984) discovered that telephone and face to face respondents are more inclined than mail respondents to use the extreme response category on the positive end of the scale. Van Sonsbeek and Stronkhorst (1983) also found that in face to face

interviews respondents are more likely to use the extreme positive end of a scale than in a mail survey.

3.4. Summary

For years the face to face interview has been considered a highly superior data collection technique. A review of the available empirical research literature only partly corroborates this view. When face to face and telephone surveys are compared only small effects are discovered. Face to face interviews have higher overall response rates and result in data with slightly less item nonresponse and slightly more statements to open questions. No differences were found concerning response validity (record checks) and social desirability. In general, similar conclusions will be drawn from **well-conducted** face to face and telephone interview surveys.

When mail surveys are compared with both telephone and face to face interviews, a clear and interesting picture emerges. It is somewhat harder to have people answer questions in mail surveys: both the overall nonresponse and the item nonresponse are higher in mail surveys. However, when questions are answered, the resulting data tend to be of better quality. In particular, mail surveys perform better with more sensitive questions (e.g., more reporting of drinking behavior, less item nonresponse on income questions). The differences between mail surveys and interview surveys were small but not negligible (the largest effect size found is .12, the smallest is .03).

Finally it should be noted that the studies analyzed in this review all concerned experiments on the influence of the data collection method used. In general, extreme care was taken to optimize both the design and implementation of the surveys (e.g., construction of questionnaires, training of interviewers, supervision), which is reflected in the high overall response rates for all three data collection methods. In the harsh daily world of survey research one sometimes has to make concessions in the design and the implementation procedures. Therefore, it is conceivable that under the constraints of more "normal" field conditions the effects of the data collection method on the data quality are stronger.

On the other hand, mode comparisons are often done with surveys on topics that were a priori expected to produce differences. In this sense, the small differences found in well-conducted surveys are encouraging.

CHAPTER 4

DESIGN OF A FIELD EXPERIMENT

To err is human, to forgive divine, but to include errors in your design is statistical

Leslie Kish, Presidential Address to the American Statistical Association, 1977

4.1. Introduction

As reported in chapter 3, a review of the published research on mode comparisons showed small, but consistent mode effects. In general, comparisons across modes have been restricted to the analysis of univariate distributions. Comparisons involving psychometric indicators of data quality, such as the reliability of multiple item scales, have been scarce. No comparisons were found involving multivariate effects of mode differences. However, minor differences in univariate measures could produce more dramatic differences between the modes in the estimated multivariate relationships. This potential mode effect should be a source of worry, especially in academic research, where multivariate relationships between the measures are commonly analyzed. Therefore, a field experiment was designed which focused on these underexamined areas.

The modes of data collection investigated are the mail questionnaire, the telephone interview and the face to face interview. In planning the design of this mode comparison, care was taken to optimize the internal validity of the experiment without jeopardizing the external validity (cf. Cook and Campbell, 1979, p. 37). In other words, the influence of error variance and extraneous variables was controlled as far as possible, but the implementation of the survey procedures remained realistic in terms of general survey practice (cf. Biemer, 1988, p. 274; Groves, 1989, p. 506). To fulfill this goal detailed decisions had to be made concerning the construction of the questionnaire, the sample used, and the allocation of respondents to interviewers. These decisions will be reported in the next sections.

This chapter is organized as follows. First, the questionnaire construction is described. In the next sections a description is given of the

sampling procedure and the procedures for the selection and training of the interviewers. Next, the implementation of the data collection methods is described, followed by a report on the pilot study. Then follows a description of the design and the fieldwork of the mode comparison. In the final section, information on the sample is given and the response rate is examined. Examples of the questions asked are given in Appendix B.

4.2. Questionnaire Construction

In criticizing alternatives to the face to face interview it is often noted that only very restricted surveys have been compared and that mail surveys and telephone interviews are limited regarding the type, format and number of the questions asked. To realize a meaningful and fair comparison, a questionnaire was constructed in which I tried to push the mail and telephone survey to their limits. It was decided to use potentially "sensitive" questions regarding subjective phenomena like loneliness, happiness, and well-being in combination with more factual questions on objective attributes like financial situation, labor force participation, and extension of the social network. Also, standard biographical information on the respondents would be collected.

Psychometric indicators of data quality are of particular interest in this mode comparison, therefore several multiple item scales had to be included in the questionnaire. Furthermore, specific questions on respondent attributes had to be included to be able to investigate potential mode effects on multivariate relationships and models. Well-documented conceptual models have been published for well-being and loneliness (see Burt, Wiley, Minor & Murray 1978; De Jong-Gierveld, 1987). In these research domains, several reliable multiple item scales have been applied successfully. These scales formed the core of the questionnaire.

A first version of the questionnaire was drafted following the rules for writing questions as formulated by, among others, Dillman (1978, chapter 3) and Sudman and Bradburn (1982). Different question formats were included: checklists, open questions, and closed questions. The latter differed in number of answer categories (varying from two to seven categories). The topic was the well-being and the financial situation of Dutch citizens. The questions varied in question threat and saliency. Three well-known multiple item scales were used to measure well-being: a balanced extension of Bradburn's Affect Balance Scale, measuring positive and negative affect (Bradburn, 1969; Hox, 1986), De Jong-Gierveld's

loneliness scale (De Jong-Gierveld & Kamphuis, 1985), and a condensed form of Brinkman's self-evaluation scale (Brinkman, 1977; Dykstra, forthcoming). Several questions about the extension of the social network and the types of relationships constituting the network were added. The financial situation was estimated by asking the net family income, and several questions concerning the family's budget and balance. In addition questions on survey experience, labor force participation, and on biographical attributes were added. This resulted in a questionnaire with 82 questions.

This draft version was first discussed with a group of experts in the field of conceptualization and measurement. An updated version was then pretested, using cognitive interview methods (Belson, 1981; Willis, Royston & Bercini, 1991). An analytic sample of 12 persons was used, varying in age and education. Special attention was given to the understanding of the questions and of the terms used. As a result several questions were adapted by adding a clarification. For instance, a more precise definition of the term "social contacts" was added to a question on satisfaction.

Based on the resulting basic questionnaire three equivalent versions of the questionnaire were developed, one for each of the three data collection methods. An iterative procedure was used in which an expert in mail surveys, an expert in telephone surveys and an expert in face to face interviews optimized the questionnaire for each method, taking care that question formats remained comparable and that no method was given extra advantages. At each step of the iteration changes were discussed; the process stopped when consensus was reached among these experts. It was decided that response cards (i.e., visual aids to present the response categories) should be used in the face to face interview for all checklists and for closed questions with five or more answer categories. Interviewer instructions were added to the questionnaires for both the face to face and the telephone interview. These instructions were printed in a special letter type, clearly distinguishing them from all material that is read to the respondent. The major difference in the printed interviewer instructions concerned the use of response cards. In the face to face mode, interviewers were simply instructed to hand the card to the respondent. At the same point in the telephone mode, interviewers were instructed to repeat the answer categories when necessary. This could be followed up by repeating the *total* question including all answer categories.

The equivalent versions of the questionnaire were field tested during a feasibility study. This study was a complete pilot study, that is, all procedures necessary for conducting a mail, a telephone and a face to face

survey were followed through on a smaller scale. For examples of the final questions used see Appendix B[2].

4.3. Sampling Procedures

Effects found in mode comparisons are often confounded because different types of respondents are selected in each mode. To control for this possible source of error, the same sampling frame and the same sampling procedures were used for each data collection mode.

The sampling frame was the total telephone directory of the Netherlands. Five municipalities were selected, stratified according to urbanization (cf. CBS, 1988). These municipalities were Schermer (a very rural region, more than 20% is farmer), Barneveld (a small municipality in a rural setting), Zeist (a medium municipality with many commuters to a nearby large city), Alkmaar (a large municipality), and Amsterdam (a very large municipality). For each municipality the local government provided a list of towns constituting the municipality. Based on these lists a computer program was written, that randomly selected a sample of addresses from the telephone directory. Whenever a typical business address was selected it was replaced by a new, randomly selected, address. In this way, a stratified random sample was taken for each data collection mode.

On each address a respondent aged 18 years or older was selected with the next birthday method (i.e., ask for the person within the sampling unit who is 18 years or older and will have the next birthday). The birthday method is nonintrusive, does not take much time, and is fairly effective (cf. Oldendick, Bishop, Sorenson & Tuchfarber, 1988; Salmon & Nichols, 1983). Therefore, the birthday method can be implemented without difficulties in both mail surveys and face to face and telephone interviews. For an overview of respondent selection techniques, see Lavrakas (1987, chapter 4).

4.4. Procedures for Selection and Training of Interviewers

Interviewers were recruited via newspaper advertisements in the selected municipalities and via advertisements at the newspapers and bulletin boards of the universities in Amsterdam. Important selection criteria were

[2] The complete (Dutch) text of the final equivalent versions of the questionnaire is available on request (see also De Leeuw, 1991).

clarity of voice over the telephone, legible handwriting and higher education.

All interviewers were extensively trained during three training sessions. A standardized interviewer training was given based on the SRC-manual (1976) and the VOI-manual (De Bie & Dijkstra, 1989). An interviewer manual and field guide was sent to the interviewers before the training started with the request to study certain chapters[3]. Basic interviewer rules were discussed and illustrated with video-examples[4]; role-play was used to practice these skills. An additional training was given in telephone interviewing techniques. In this session special attention was given to the different channels of communication used in face to face and telephone contacts. The discussion centered on ways to use paralinguistic and explicit verbal communication to compensate for the absence of nonverbal communication in a telephone conversation.

Previous to the training, all interviewers had completed a self-administered version of the questionnaire. They were asked to send an inventory of perceived "problem" questions and situations to the trainer. Special attention was given to these interviewer comments during training and supervision.

The same interviewers were used in both the face to face and the telephone condition. The interviewers were randomly divided in two groups. The first group started with telephone interviews and then conducted face to face interviews, the remaining interviewers started with face to face interviews. Respondents were randomly assigned to interviewers within geographical units.

4.5. Implementation of Data Collection Procedures

In the *mail survey condition* Dillman's Total Design Method (TDM) was followed completely, including a third and last reminder by certified mail. Important features of Dillman's TDM are: a personalized cover letter, an attractive questionnaire, and follow-up mailings. One week after the initial

---

[3] A separate field guide was developed for telephone interviewing (De Leeuw & Hox, 1989a) and for face to face interviewing (De Leeuw & Hox, 1989b). The (Dutch) text of these field guides is available both in hard copy and on a floppy disc.

[4] The videotapes used were: 'Een vraag en een weet', developed by the Erasmus University, Rotterdam, and 'Verantwoord vragen' developed by the Vrije Universiteit, Amsterdam.

mailing, the entire sample (respondents and non-respondents) receives a postcard serving as a thank you or as a reminder. Three weeks after the initial mailing all non-respondents receive a new questionnaire and cover letter. Seven weeks after the initial mailing this procedure is repeated, but this time preferably by certified mail (Dillman, 1978; De Leeuw & Hox, 1988). In addition, a short letter notifying the respondents of the mail survey was mailed one week in advance. In the cover letter we requested a specific member of the household (i.e., 18 years or older and first birthday) to complete the questionnaire. No incentives were offered, besides a summary of the major results.

In the *face to face condition* all sample units received a letter one or two days before they were contacted by the interviewers. This letter incorporated all the information of both the mail advance letter and the mail cover letter. Interviewers contacted respondents by phone to make an appointment for an interview, using the next birthday method to select an eligible respondent. Interviewers were instructed to make at least seven calls, and phone at different times at night and during the day time and in weekends. Scripts were used to persuade eligible respondents. No attempt was made to convert definite refusers, meaning that refusers were **not** called back by selected interviewers specialized in refusal conversion.

Response cards were used with checklists and with questions offering five or more alternatives. To optimize interviewer supervision in the field, all interviews were tape recorded and spotchecks of the quality of the interviews were held by listening to parts of the audiotapes.

In the *telephone survey condition* again all sample units received an advance letter. The interviews were conducted at a centralized setting. A paper and pencil procedure was used for the majority of the interviews. A supervisor was present all the time. Tape recordings were made of the interviews. At the beginning of an interview session additional instructions or feedback was given to the interviewers if necessary.

Telephone interviews were conducted weekdays from 7 p.m. until 10 p.m. and on weekends from 10 a.m. until 2 p.m. Eligible respondents were selected using the next birthday method. Parallel to the situation in the face to face condition, interviewers were provided with scripts for the selection and persuasion of respondents. When necessary, appointments for telephone interviews were made. At least seven call backs were made, but further attempts to interview non-contacts were made till the end of the data collection period. Again we did not use refusal conversion for definite refusers.

4.6. Pilot Study

A pilot study was conducted in the autumn of 1989. In this pilot all procedures necessary for conducting a mail, a telephone and a face to face survey were followed through on a small scale. The objective of the pilot was twofold: (1) to pretest the equivalent versions of the questionnaire, and (2) to field test the administrative design and the logistics of the main experiment in a realistic setting.

Nine interviewers were selected and trained as described in section 4.4. Three stratified random samples of addresses were drawn according to the procedures outlined in section 4.3. A total of 100 addresses were contacted for the mail survey of which 69 (69%) completed the questionnaire. For the telephone survey 60 addresses were contacted, resulting in 38 (63%) completed telephone interviews. For the face to face survey 42 addresses were contacted, resulting in 22 (52%) completed face to face interviews.

The three equivalent versions of the questionnaire performed well. One extra instruction to the interviewers was added in both the face to face and the telephone questionnaire: the interviewers were asked to field code the precision with which respondents answered a question on family income. In the mail questionnaire this coding was done by a coder immediately after the questionnaire was returned. No further changes were necessary. The entire data collection process went very smoothly, and again no changes were required.

4.7. Field Experiment

One modification was made to the design of the field experiment. A small CATI experiment was added to investigate a specific hypothesis concerning the reliability of multiple item scales. For more detail on this subject, see chapter 6. The paper and pencil telephone questionnaire was implemented straightforwardly, including the appropriate skippings and branchings. The program used for the CATI-application was THIS (The Interview System)[5]. The questions of the four major multi-item scales (positive affect, negative affect, loneliness, and self-evaluation) were randomized within each scale. This was the only important difference with the paper and pencil telephone questionnaire.

[5] This part of the experiment was done in collaboration with J.J. Hox, Department of Education, University of Amsterdam.

Twenty interviewers were selected and trained as described in section 4.4. Six of them had already worked for this project during the pilot study. The data collection started on 4 September 1989 and the last interview was completed by 30 November 1989. The procedures are described in section 4.5. All twenty interviewers conducted both face to face and paper-and-pencil telephone interviews. Ten randomly selected interviewers started with telephone interviews and then conducted face to face interviews, the other ten started with face to face interviews. A subgroup of seven interviewers received a special one evening training session in CATI-procedures at the end of the data collection period and conducted a series of computer assisted telephone interviews. The procedures were the same as in the paper and pencil telephone interview.

During the fieldwork the interviewers were closely supervised (see section 4.5). Spotchecks of the quality of the face to face interviews were held by listening to parts of the audiotapes. The telephone interviews were checked by listening to the interviews in progress. The training and supervision of the interviewers were successful. Only small interviewer effects were found in both the face to face and telephone interviews. Furthermore, the effects did not differ between the two modes. For a detailed description see Hox, De Leeuw and Kreft (1991).

4.8. Sample and Nonresponse

Response rate

Four stratified random samples of households were taken from the telephone directory of the Netherlands as described in section 4.3. Within households respondents of 18 year and older were selected according to the next birthday method. Sample sizes were: 400 (mail survey), 530 (face to face survey), 450 (paper-and-pencil telephone survey) and 120 (computer assisted telephone survey). In the interview conditions at least seven call-backs were made trying to contact respondents, but no attempt was made to convert explicit refusals by special call-back methods. In the mail survey condition Dillman's TDM was followed completely, including a third and last reminder by certified mail (see also 4.5).

The response rate was calculated as the percentage of completed interviews or questionnaires to all *eligible* cases (including noncontacts). The mail survey resulted in a final response rate of 68%. The face to face interview had a response rate of 51%, the paper-and-pencil telephone

interview had a response rate of 66%, and CATI resulted in a response rate of 71%. The results are summarized in Table 4.1. The face to face interview resulted in a significantly lower response rate than either the mail survey or both types of telephone interview (p=.00). Pairwise comparisons did not reveal any statistically significant difference in response rate between the mail and telephone surveys. The difference in response rate is almost entirely due to a higher proportion of explicit refusals in the face to face condition. For instance, 40% of the eligible face to face respondents refused cooperation, as did only 28% of the eligible paper and pencil telephone respondents.

Table 4.1 Response and Nonresponse by Type of Data Collection Method

	Mail	Face to Face	Telephone	
			P&P	CATI
Total	400	530	450	120
%	100%	100%	100%	100%
Completed	254	243	266	77
%	64%	46%	59%	64%
Refusals	44	191	114	23
%	11%	36%	25%	19%
Ineligible	27	50	47	12
%	7%	9%	10%	10%
Noncontact	75	46	23	8
%	19%	9%	5%	7%

Note. Very strict criteria for ineligibility were used. For instance: business number/address, telephone not working and no new number known at telephone company, household/family unknown, did not speak Dutch at all. When a potential respondent answered too old, sick, someone in family is sick/died, it was recorded as refusal.

Selectivity of nonresponse

Nonresponse, especially the relatively large nonresponse of the face to face interview, could be a potential source of error. Fortunately, external information was available on both respondents and nonrespondents, and

could be used in further analysis of the nonresponse. The additional information is based on the Dutch zip code system (Geo-marktprofiel) and consists of aggregated information for 373.000 zip codes, with on average a density of *15 households* per zip code. Linked with the zip code, the following information was available for the sample units: type of dwelling, value of property (i.e., rent or buying price), building year, family income, family stage (i.e., young - old), and urbanization.

First a homogeneity analysis (Gifi, 1990, chap. 3 ; Van de Geer, 1985) was performed on the zip code information for the total sample (respondents and nonrespondents). This resulted in three dimensions. The first main dimension can be described as "affluence." Type of dwelling, value of property, and urbanization have a high discrimination measure on this dimension. The second dimension can be described as "starting house owners"; mainly characterized by the year the house was built, the neighborhood and the urbanization. The third dimension merely indicates that little is known about the households on the key (i.e., zip code based) variables. Object scores for the three dimensions were calculated and added to each sample unit. Differences between respondents and nonrespondents were then analyzed using the auxiliary zip-code information.

To investigate whether the modes differed in selective nonresponse I used analysis of variance with mode of data collection and response (yes/no) as factors and the three homogeneity dimensions as dependent variables. No significant differences were observed for the dimensions "starting house-owner" and "no information" at the 05 level. Respondents and nonrespondents did differ significantly on the dimension "affluence" (p=.02). However, *no* significant interaction with mode of data collection was found; in other words, there was no difference in selective nonresponse between the data collection methods.

Further analysis of the difference in affluence between respondents and nonrespondents showed that the nonrespondents more often lived in big cities, in rented houses, and had a lower income. Respondents on the other hand lived more often in rural areas, owned their homes and belonged to the middle and higher income classes. These trends were very small. When the type of nonresponse is incorporated in the analysis, an interesting pattern emerges. Respondents and refusers do not differ from each other, but they do differ from the noncontacts and the "unreachables" (i.e., sick, senile, language problem) (p= 00). In general, these groups were less affluent, did not own a house and were more often found in urban areas. Also, less was known about them concerning the zip-code information as a

whole. Again, no significant interaction was found with mode of data collection.

Socio-demographic characteristics of respondents

I investigated whether respondents in the four modes differed in important background variables like gender, age, education, marital status, having children, and previous interview experience (see also Appendix B). Chi-square tests were employed at the .05 level. The only statistically significant differences observed over modes concerned gender (p=.02) and marital status (p=.00). Pairwise comparison of the methods showed that this overall difference was caused by differences between the face to face and the mail survey.

In the mail condition relatively more respondents were men, in the face to face condition relatively more respondents were women (p=.01). When subsequently the distribution of the respondents on the variable gender is compared with figures on the general population (CBS, 1990), no statistically significant difference is found for the telephone respondents. Among the face to face respondents women are indeed overrepresented (p=.03), and there is a nonsignificant (p=.07) tendency of an overrepresentation of men in the mail survey.

In the mail condition more married persons and in the face to face condition slightly more divorcees and widowers were present (p=.00). Also, in the telephone survey relatively more widowers and unmarried were present, while more married people responded to the mail survey. Respondents on the telephone survey (both paper-and-pencil and CATI) did not differ from respondents on the face to face survey, neither did respondents to the (paper and pencil) telephone survey differ from respondents to CATI (smallest p-value .11). When population data on official marital status (CBS, 1990) are considered it is found that there is a general overrepresentation of unmarried individuals for all four data collection methods (p=.01), and of divorcees for the face to face mode (p=.00).

It is interesting that the respondents in the four modes did not differ in age (p=.68) or education (p=.34) as is often presumed. Across the four modes the only statistically significant differences concerned the variables gender and marital status. These differences can confound substantial conclusions on mode differences. To statistically correct for this, the variables gender and marital status will be included in the subsequent mode comparisons.

Furthermore, it should be noted that the finding that respondents hardly differ across modes does **not** mean that the respondents are **completely** representative for the Dutch population. In fact, respondents and nonrespondents did differ slightly in "affluence"(see above). But, there was no interaction with data collection method; the selectivity of response was the same for all modes. The same is true concerning education: the respondents in the four modes do not differ on educational level. But, when these figures are compared with data on the educational level of the Dutch population in general[6], individuals with a high educational level (college or university) turn out to be overrepresented, while individuals with only elementary (primary school) education are overall underrepresented (p=.00). No clear differences were found concerning age.

4.9. Summary

Four well-known potential sources of error are: the mode of data collection, the questionnaire, the interviewers, and the respondents (Groves, 1989). Effects found in mode comparisons are often confounded, for instance when different question types are used, or different types of respondents are selected and interviewed during different periods of the year. To be able to investigate the influence of the data collection technique itself it is necessary to control for other possible sources of error. In this chapter I described the design of a mode comparison experiment. Special care was taken to optimize the internal validity of this field experiment without jeopardizing the external validity. Equivalent versions of the same questionnaire were used in which a variety of question types were applied, the topic being the well-being and the financial situation of Dutch citizens. The same trained interviewers were used in both the face to face and the telephone modes, and random samples from the same sampling frame were taken for each mode using the same sampling procedure.

Also in this chapter figures on the (non)response were presented, and the potential threat of selective nonresponse was further investigated. There was a significant difference in response rate between the methods. The face to face survey resulted in the lowest response rate (51%). There was no statistical difference in the response between the mail survey (68%)

[6] The sources for comparison were for educational level 'Sociaal en Cultureel Rapport 1988' (Social and Cultural Report: SCP. 1988, p. 315) and for age, gender and marital status 'Statistisch Jaarboek 1990' (Statistical Yearbook: CBS, 1990).

and the paper and pencil telephone survey (66%), and the added (small) CATI-survey (71%). For all sample units (respondents and nonrespondents) additional information was available on the household and the neighborhood. When respondents and nonrespondents were compared on this background information, small differences in affluence were found. This difference can be mainly attributed to those nonrespondents that could not be reached; respondents and refusers did not differ strongly from each other. Although the data collection methods do differ in response rate, no difference in selective nonresponse could be detected for these background variables: the pattern was the same for all three data collection methods.

In addition, the respondents were compared on available background characteristics across modes. A statistically significant difference was detected for the variables gender and marital status. To control for this confounding, it was decided to include the variables gender and marital status in all subsequent statistical analyses. It is very interesting to note that the respondents in the modes did **not** differ in age or education, as is often presumed. All modes did as well (or as badly) in sampling the elderly and the poorly educated. A comparison with published statistics (CBS, 1990; SCP, 1988) showed that respondents with a college or university education were overrepresented, while respondents with only a primary education were underrepresented in all four surveys. No clear age differences were found.

CHAPTER 5

DATA QUALITY I: A REPLICATION IN THE NETHERLANDS

'Data! data! data!' he cried impatiently. 'I can't make bricks without clay.'
Sir Arthur Conan Doyle, The copper beeches; The adventures of Sherlock Holmes, 1981, p. 268

5.1. Introduction

This chapter presents the results of a first comparison of the data gathered in the field experiment. The data of the mail survey, the face to face interview, and the paper and pencil telephone interview are examined for mode effects. The ultimate dependent variables in the analyses are the differences between the answers to specific questions received in the three modes. Since there is no direct way to check the information on the *subjective* phenomena under study, record checks to estimate the data quality are impossible (cf. Groves, 1989, p. 304). Instead, the following aspects of data quality are investigated: number of responses to open questions, item missing data (item nonresponse), differences in response distributions on sensitive topics, acquiescence and preference for extreme answer categories (extremity). Furthermore, respondents' evaluation of the survey is compared over modes.

Mode differences concerning these aspects are discussed in the sections 5.3 to 5.7. Each section starts with a concise overview of a priori expectations; these expectations are based on the theoretical discussion in chapter 2 and the results of the meta-analysis as presented in chapter 3. Next the results of the statistical analyses are presented and discussed.

A short overview of the methods of data analysis is given in section 5.2; a summary of the main results is given in section 5.8.

5.2. Data analysis

The following general strategy was used throughout this chapter: First an overall statistical test was performed. If the overall test indicated a statistically significant difference between the modes, it was followed up by a series of pairwise comparisons. A significance level of .05 was adopted in all tests.

In cases with more than one dependent variable multiple tests were done (e.g., the data on four open questions were analyzed to investigate mode influences on the number of responses to open questions). To avoid chance capitalization I used Holm's sequentially rejective Bonferroni test. This is a simple procedure in which n tests are ordered according to their exact p-value (the smallest first). For the first test the significance level (.05/n) is employed, for the second test the significance level used is (.05/(n-1)), etcetera (Holm, 1979).

The final strategy employed was slightly more complicated than the one described above. Recall, that the respondents in the modes differed on two background variables. In the mail condition slightly more men and married persons were present, while in the face to face condition slightly more respondents were women and slightly more respondents were divorced (see section 4.8). These differences in gender and marital status can influence the conclusions. Mode differences detected could be the result of the different processes taking place in the data collection modes, but could also be partly attributed to the differences in gender and marital status. Therefore, a two-step procedure was used. First, an overall test (e.g., an analysis of variance) was done, thereby answering the practical question whether the data collection methods each would get the same results. Second, the data were reanalyzed while taking into account the observed differences in gender and marital status (e.g., an analysis of covariance with gender and three dummy codes for marital status as covariates). This reanalysis makes it possible to decide whether a "pure" mode effect is present (cf. Biemer, 1988, p. 274; Groves, 1989, p. 502). Unless stated differently, pairwise tests were always conducted in the second step, taking into account the differences in gender and marital status.

5.3. Responses to Open Questions

Open questions allow the respondent to formulate her/his own answer to a question. The number of different responses that a person gives to an open

question is a useful proxy for the extent to which the answer fully characterizes the respondent's thoughts (Groves, 1978). In general, the more effort a respondent invests in the task of answering, the more complete will be the answer.

A well-trained interviewer can motivate respondents during the interview process and probe for additional answers (cf. chapter 2). In telephone interviews, however, the channel capacity is limited to verbal and paralinguistic means of communication. Since nonverbal communication plays a function in both motivating respondents (indicating that attention is being paid) and in feedback (cf. Argyle, 1973), it is expected that respondents in face to face interviews will give more responses to open questions than respondents in telephone interviews.

In mail surveys no interviewer is present to stimulate more detailed answers. Besides, a specific medium related factor hampers the performance of the mail respondent even further: writing down a full answer demands a relatively high active command of a language compared to verbalizing it to an attentive listener. People feel often compelled to avoid grammatical errors in written communications and are more apprehensive about their capacities to write something down than about their capacities to tell a story (see also Lévy-Leblond, 1990). This can inhibit their motivation to fully answer an open question in writing (cf. chapter 2).

Mail surveys are therefore supposed to be poor performers when open questions are being used. Surprisingly, I could not find a study comparing mail surveys and interview surveys on this criterion in the meta-analysis. Comparisons between face to face and telephone interviews showed that in face to face interviews open ended responses are indeed longer and contain more units of information (cf. chapter 3).

To compare the performance of mail surveys with interviews I analyzed four open questions. Three questions asked the respondents to elucidate their responses. The first question asked for an inventory of items that were perceived by the respondent as important, but could not be afforded financially at that time. The second question asked for reasons why the respondent had refused previous surveys, if applicable. The third question asked respondents to explain their previously stated preference for a data collection method. The fourth question asked the respondents at the end of the interview or the questionnaire if they had any comments, questions etc. about this survey. This last question is common to (TDM) mail surveys, but is less often asked in face to face interviews. For each question the total number of different statements was coded.

Table 5.1 An(c)ova on Number of Statements to Open Questions: P-values

P-values for the main effect of mode, for the total effect of the covariates (gender and marital status) and for the main effect adjusted for differences in covariates among the modes. Percentage of variance explained by mode of data collection; the percentage adjusted for differences in covariates is given in parentheses.

Dep. Var.	Main Effect p-value	Covariates p-value	Adj. Main p-value	% Var. Expl. unadj. & adj.
Inventory	.458	.042	.560	0.60% (0.43%)
Reasons I (refusal)	.345	.152	.321	0.71% (0.75%)
Reasons II (preference)	.006	.000	.006	1.65% (1.61%)
Comments	.020	.760	.017	1.03% (1.08%)

Analysis of variance did not detect differences between the modes for the first two questions. The third question did show differences. Subsequent pairwise tests showed that respondents on the mail survey gave fewer reasons for their preference for a particular mode. Contrary to expectation, no significant difference in number of reasons was detected between telephone and face to face surveys. A marginally significant difference was observed for the fourth question[7]. Respondents in the mail survey condition made slightly more comments at the end of the questionnaire than respondents in either face to face or telephone interviews. Again no differences were found between the face to face and the telephone condition. All differences found were very small. These differences can be the result of the different processes taking place in the three modes, but can also be partly attributed to the self-selection of respondents and the differences in gender and marital status as reported in section 4.8. In addition to a simple analysis of variance on the number of statements, I reanalyzed the data

[7] To avoid capitalization on chance I used the sequentially rejective Bonferroni test as proposed by Holm (1979).

using analysis of covariance. Gender and three dummy codes for marital status were used as covariates. The same conclusion holds when I corrected for self-selection of respondents. The only significant covariate was gender: women make slightly more statements to open questions. The results are summarized in Table 5.1 and Table 5.2.

Table 5.2 An(c)ova on Number of Statements to Open Questions: Means

Mean number of statements for each data collection method. Means adjusted for the covariates are given in parentheses. Methods that differ significantly (p=.05) on an additional pairwise test are reported.

Dep. Var.	Mail	Face to face	Telephone	Pairwise	Ntot
Inventory	1.67 (1.68)	1.87 (1.84)	1.82 (1.83)	n.a.[a]	263
Reasons I (refusal)	1.50 (1.50)	1.53 (1.52)	1.63 (1.63)	n.a.[a]	302
Reasons II (preference)	1.68 (1.67)	1.95 (1.94)	1.84 (1.86)	M-F, M-T	617
Comments	1.00 (1.01)	0.76 (0.75)	0.62 (0.62)	M-T	762

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

It should be noted that the open questions asked in this field study were short and dealt with well-defined topics. Asking for attitudes on vague concepts could produce other and perhaps stronger effects. Nevertheless, in this study open-ended questions do seem to perform reasonably well in mail surveys.

No statistically significant differences were detected between face to face and telephone interviews. However, the meta-analysis revealed a small, but statistically significant overall effect in favor of the face to face interview. It was also found that the effect sizes were heterogeneous over the studies, which indicates the influence of possible moderator variables. As only four studies on open questions were available for the meta-analysis,

further detailed statistical analysis of the heterogeneity was not possible. Groves (1978) points out that for some open questions the differences found between face to face and telephone interviews are negligible, but that the difference is rather large for other questions, such as abstract or generic open questions about the most important problems facing the country. Both Jordan et al. (1980) and Herman (1977) did not find a statistically significant effect with more concrete questions about medication used or important issues raised in a past union campaign. The questions analyzed in this field study were also short and dealt with relatively well-defined topics. This indicates that on concrete and short open questions both telephone and face to face interviews perform equally well.

5.4. Item Missing Data

Missing data can pose serious problems in statistical analysis. As a consequence, item missing data rate or item nonresponse has received considerable attention in empirical mode comparisons. In general, it is expected that interviews produce less missing data than mail surveys. An interviewer can repeat questions and probe to get an answer. In a face to face situation an interviewer can use more communication channels than in telephone interviews, which could lead to better communication and fewer missing data. A review of the empirical literature did indeed show that face to face interviews resulted in the lowest proportion item nonresponse, telephone interviews produced a somewhat higher proportion, and mail surveys had the highest proportion item nonresponse (chapter 3). But, the differences between methods were small and the differences between face to face and telephone interviews tend to diminish over time (Groves, 1989. p. 514). Also, there is some evidence that mail surveys perform better when sensitive questions are asked. For instance, income questions in mail surveys result in less item nonresponse (Nuckols, 1964; Siemiatycki, 1979; Van Sonsbeek & Stronkhorst, 1983; see also chapter 3). It is therefore conceivable that a differential pattern of item nonresponse will be found, depending on the topic of the questions asked. To investigate this expectation, I computed both a global and several topic-specific indicators of item nonresponse.

As a global indicator the proportion of item nonresponse was computed over all 82 questions. Four topic-specific missing data indicators were constructed: measuring the proportion item nonresponse on questions about loneliness and availability of social support, on questions about happiness

and affect, on financial questions, and on biographical questions. Questions about finances are generally viewed to be among the most threatening ones (Sudman & Bradburn, 1974; Körmendi & Noordhoek, 1989). In accordance with this view I expect less item nonresponse for the mail survey on this topic compared to both interview modes.

The results only partially corroborate this hypothesis. An analysis of variance detected a statistically significant but small effect for the global indicator. The largest difference was between face to face interviews and mail questionnaires, and was in favor of the face to face interview. The telephone survey did not differ much from either method. The topic-specific indicators followed the same pattern, except the financial questions, which led to no difference in item nonresponse between methods (See Table 5.3).

Mode differences can be the result of the different processes taking place in the three modes, but can also be partly attributed to the self-selection of respondents and the differences in gender and marital status as reported in section 4.8. Therefore, I reanalyzed the data using analysis of covariance. Gender and three dummy codes for marital status were used as covariates. In all cases gender was not significant. Marital status had some influence, but the pattern found earlier remains the same. Table 5.3 gives a summary of the results.

Table 5.3 An(c)ova on Item Missing Data Indicators: P-values

P-values for the main effect of mode, for the total effect of the covariates (gender and marital status), and for the main effect adjusted for differences in covariates among the modes. Percentage of variance explained by mode of data collection; the percentage adjusted for differences in covariates is given in parentheses.

Dep. Var.	Main Effect p-value	Covariates p-value	Adj. Main p-value	% Expl. Var unadj. & adj.
Global	.019	.003	.013	1.04% (1.12%)
Social support	.000	.000	.000	2.28% (2.53%)
Happiness	.052	.128	.490	0.77% (0.79%)
Finances	.102	.038	.117	0.60% (0.56%)
Biographical	.037	.463	.012	0.86% (1.16%)

With the exception of the financial questions, statistically significant differences between the modes were observed. In the case of happiness this was very marginal[8]. Further analysis, using pairwise tests, showed that the overall statistical difference was caused by more missing data in the mail survey (see Table 5.4). Pairwise tests did *not* detect significant differences between face to face and telephone surveys (lowest p-value=.061). The differences detected were extremely small as is indicated by the percentage of explained variance. The largest effect (for questions on social support and loneliness) attributed only 2.5% of the variance to mode effects (see Table 5.3). This is further illustrated by the size of the mean proportion item nonresponse for each mode, as given in Table 5.4. Differences between modes are small indeed[9].

Table 5.4 An(c)ova on Item Missing Data Indicators: Means

Mean proportion item nonresponse for each data collection method. Means adjusted for the covariates are given in parentheses. Methods that differ significantly (p=.05) on an additional pairwise test are reported.

Dep. Var.	Mail	Face to face	Telephone	Pairwise	Ntot
Global	.02	.01	.01	M-F	762
	(.02)	(.01)	(.01)		
Social support	.04	.01	.02	M-F, M-T	762
	(.04)	(.01)	(.02)		
Happiness	.01	.00	.01	M-F	762
	(.01)	(.00)	(.01)		
Finances	.05	.06	.07	n.a.[a]	762
	(.05)	(.06)	(.07)		
Biographical	.00	.00	.00	M-F, M-T	762
	(.00)	(.00)	(.00)		

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

[8] To avoid capitalization on chance I used the sequentially rejective Bonferroni test as proposed by Holm (1979).

[9] Since the distributions of the indicator for item missing data are highly skewed, I also analyzed the data using a logit transformation for the dependent variables. This did not substantially change the conclusions.

Overall, the mail survey resulted in slightly more missing data than the face to face and the telephone interviews. This confirms the results of the meta-analysis. Contrary to expectation, no differences were detected between the face to face and the telephone mode. But, both Groves & Kahn (1979) and Jordan et al. (1980) noted that the differences in item missing data rate between face to face and telephone interviews gradually decreased when more experience was gained with the telephone mode. It should be noted that this field study has profited from the large experience gained in telephone survey methodology (e.g., Groves et al., 1988), and it is assumed that later studies will replicate this finding.

Although the mail survey produced the fewest number of missing data on the financial questions, the differences were not statistically significant. A further analysis of the data on sensitive questions will be presented in the next section.

5.5. Sensitive Topics

Data collection methods are supposed to differ especially on sensitive questions. The physical absence or presence of the interviewer is generally believed to be important. However, contradictory hypotheses are formulated in the literature. For instance, the physical presence of a skilled interviewer may motivate respondents and create a feeling of trust (Galtung, 1967). Others argue that self-administered questionnaires and telephone surveys present fewer problems of self-presentation and introduce a greater feeling of anonymity (Cannell & Fowler, 1963; Sudman & Bradburn, 1974), provided that the legitimacy of the survey was clear (De Leeuw & Van der Zouwen, 1988; Dillman, 1978; Groves, 1989).

The results of the meta-analysis indicate that mail surveys perform slightly better than both face to face and telephone interviews. Also it was found that the differences between the two interview modes on the indicator "social desirability" were heterogeneous. In recent comparisons between face to face and telephone surveys no differences were detected on sensitive questions, but in older comparisons differences were found to be statistically significant (see also Groves, 1989, p. 520).

In this field experiment we focused on the more emotionally difficult subjects for social surveys. Therefore, questions on sensitive topics and with a potential high risk for social desirability bias were included. In the next part I will discuss the results of the mode comparison on questions about

income, loneliness, self-evaluation, and well-being, assuming that acknowledgment of negative feelings is a socially undesirable action.

Income

In all three modes an open-ended question on net family income was asked. In the western world questions on income are generally seen as threatening (Sudman & Bradburn, 1982). Both cognitive and emotional factors could influence the answers given (Körmendi, 1988; Körmendi & Noordhoek, 1989). For instance, memory and knowledge can play an important role in the precision of the answers. In mail and face to face surveys respondents have far more opportunities to look up the net income and/or check it with other household members than during a telephone survey. This is especially true in the mail survey where the respondent is the sole locus of control. Issues of privacy and perception of social acceptability of high or low incomes can influence the willingness to respond.

However, no significant differences in item nonresponse and in reported income were observed across the modes (Table 5.5), indicating an unexpected absence of mode effects. It should be noted that the proportion item nonresponse for the income question was by far the highest in this survey (mail: .14, telephone: .18, and face to face: .17). Compared with for instance the item nonresponse on personal questions like "I have a low opinion of myself" (respectively: .00, .00, .00) or "I really miss a close friend" (respectively: .02, .00, .00) this is high.

Table 5.5 An(c)ova on Monthly Net Family Income

Proportion missing data and reported net income. Reported are means and p-values for the main effect of mode, p-values for the total effect of the covariates and for the main effect adjusted for differences in covariates among the modes. As an effect size indicator the percentage of variance explained by mode of data collection is given. Estimates adjusted for differences in covariates are given in parentheses.

	Prop. Missing income quest.	Reported income
Mean Main Effect		
Mail	.14 (.14)	2953.65 (2865.83)
F-t-f	.17 (.17)	2628.43 (2712.18)
Tel.	.18 (.18)	2758.89 (2766.60)
% Var. Expl.	0.22% (0.23%)	0.70% (0.16%)
P-value Main Eff.	.426	.108
P-value Covars.	.305	.000
P-value adj. Main	.423	.572
N-tot	762	635

Finally, the precision of an answer was determined by a simple code indicating whether the respondents reported their family income in guilders and cents, reported it in rounded guilders, or whether respondents spontaneously added words like approximately to their answer. Significant differences were found between the three modes (p=.00). In the mail survey condition more often a precise amount in guilders and cents was reported, while in the face to face interview more often the qualifier "approximate" was added by the respondent (see also Table 5.6).

Table 5.6 Mode and Precision of Reported Income

Cell counts, column percentages and adjusted (standardized) residuals.

	Mail	Face to face	Telephone	N
Guilders & Cents	32	8	5	45
	16%	4%	2%	
	5.7	-2.2	-3.5	
Rounded Guilders	128	71	105	304
	61%	35%	49%	
	4.5	-4.6	0.1	
Approximate	48	123	105	276
	23%	61%	49%	
	-7.5	5.8	1.7	
N	208	202	215	625

Chi-square=78.93, df=4, p=.00, likelihood ratio chi-square=80.55, p=.00

The respondents in the three survey conditions were found to differ on the variables gender and marital status. These differences can be (partly) responsible for the observed differences in precision. To investigate this alternative hypothesis I employed a loglinear model (cf. Fienberg, 1978). A significant effect of marital status on precision was detected (Likelihood ratio chi-square=18.77, df=6, p=.00). After correcting for this effect the interaction between precision and data collection method remained significant (Likelihood ratio chi-square=85.47, df=4, p=.00). Inspection of the parameter estimates for the interaction of precision by data collection method confirmed the conclusions based on the data in Table 5.6.

In short: no differences between the three data collection methods were observed on magnitude of reported income and on item nonresponse. The only differences discovered were in reported precision. This last finding suggests a greater tendency of respondents in mail surveys to look up or check their responses.

Loneliness and well-being

One of the main advantages of self-administered questionnaires is that the absence of the interviewer may introduce a greater feeling of anonymity in

the respondent (Cannell & Fowler, 1963). The more anonymous and private setting in which self-administered questionnaires are completed, reduces the tendency of respondents to present themselves in a favorable light (Ellis, 1947; Sudman & Bradburn, 1974). Telephone interviews are somewhere in between self-administered questionnaires and face to face interviews as to their degree of impersonality (Bradburn, 1983). Respondents have more "personal space" in a telephone interview; the proximity of an interviewer in a face to face contact and the opportunities for eye contact may be detrimental to the discussion of intimate subjects (Argyle & Dean, 1965). Thus face to face interviews may present more problems of self-presentation than telephone interviews, which in turn may present more problems than mail surveys; resulting in greater self-disclosure and acknowledgment of feelings of loneliness, low self-evaluation and unhappiness in the mail survey (cf. Hochstim, 1967; Wiseman, 1972; Siemiatycki, 1979). The greatest advantage of face to face interviews -the physical presence of the interviewer- may at times be its greatest drawback (Dillman, 1978).

For the eleven-item loneliness scale both the total score and proportion of item nonresponse were computed. There was a small but statistically significant difference between the modes. The mean loneliness score in the mail condition was slightly higher, supporting the hypothesis that the more anonymous mail survey leads to more self-disclosure. The only significant covariate was marital status; correcting for this self-selection of respondents increases the effects found. The overall statistical significance was caused by more reported feelings of loneliness in the mail condition. Pairwise tests did not find a difference between the face to face and the telephone condition. Furthermore, the mail survey resulted in somewhat more missing data on the loneliness items; this difference was only marginally significant[10]. Perhaps the social pressure to answer an interviewer produces less missing data while it inhibits self-disclosure at the same time (Groves, 1989; Sigelman, 1982). Scott (1968, p. 236) takes this argument even further and points out that a desire to appear cooperative may confound test scores in the direction of fewer don't knows. It should be kept in mind that the effects found are small (see also Table 5.7).

[10] To avoid capitalization on chance I used the sequentially rejective Bonferroni test as proposed by Holm (1979).

Table 5.7 An(c)ova on Loneliness Scale

Proportion missing data and total score on an eleven-item scale. Reported are means and p-values for the main effect of mode, p-values for total effect of the covariates and for the main effect adjusted for differences in covariates among the modes. As an effect size indicator percentage of variance explained by mode of data collection is given. Estimates adjusted for differences in the covariates are given in parentheses.

	Prop. Missing Loneliness-scale	Total score on 11 items
Mean Main Effect		
Mail	.01 (.01)	3.30 (3.36)
F-t-f	.00 (.00)	2.67 (2.61)
Tel.	.00 (.00)	2.67 (2.66)
% Var. Expl.	0.99% (1.06%)	1.06% (1.37%)
Pairwise test (p=.05)	M-F, M-T	M-F, M-T
P-value Main Eff.	.023	.019
P-value Covars.	.488	.000
P-value adj. Main	.018	.005
N-tot	762	749

The eight-item self-evaluation scale shows a similar pattern, confirming the hypothesis on self-disclosure. The mail survey resulted in a slightly lower score for self-evaluation. Again, pairwise tests did not find a difference between the face to face and the telephone condition. Significant covariates are gender and marital status: women and widow(er)s report a lower self-evaluation (Table 5.8). No differences were found concerning item missing data.

Table 5.8 An(c)ova on Self-evaluation Scale

Proportion missing data and total score on an eight-item scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates and for the main effect adjusted for differences in covariates, and of variance explained by mode. Estimates adjusted for differences in covariates are given in parentheses.

	Prop. Missing self-evaluation scale	Total score on 8 items
Mean Main Effect		
Mail	.00 (.00)	5.16 (5.17)
F-t-f	.00 (.00)	5.66 (5.69)
Tel.	.00 (.00)	5.70 (5.67)
Pairwise test (p=.05)	n.a.[a]	M-F, M-T
% Var. Expl.	0.14% (0.12%)	1.32% (1.26%)
P-value Main Eff.	.592	.007
P-value Covars.	.273	.000
P-value adj. Main	.626	.007
N-tot	762	750

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

The two happiness-scales reveal no clear differences between the modes. See also Table 5.9 and Table 5.10.

Table 5.9 An(c)ova on Negative Affect (Unhappiness) Scale

Proportion missing data and total score on a nine-item scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates and for the main effect adjusted for differences in covariates, and of variance explained by mode. Estimates adjusted for differences in covariates are given in parentheses.

	Prop. Missing Neg. Affect-scale	Total score on 9 items
Mean Main Effect		
Mail	.01 (.01)	2.40 (2.46)
F-t-f	.00 (.00)	2.94 (2.87)
Tel.	.00 (.00)	2.70 (2.70)
Pairwise test (p=.05)	M-F, M-T	n.a.[a]
% Var. Expl.	0.92% (0.91%)	1.03% (0.58%)
P-value Main Eff.	.030	.022
P-value Covars.	.185	.000
P-value adj. Main	.031	.099
N-tot	762	743

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

Negative affect (unhappiness) initially shows a significant difference between the data collection methods. However, this effect can be completely explained by differences in gender and marital status between respondents in the three modes. Women and divorcees rapport slightly more feelings of negative affect, while married people report less negative feelings. No significant effects were found for positive affect (happiness).

Table 5.10 An(c)ova on Positive Affect (Happiness) Scale

Proportion missing data and total score on a nine-item scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates and for the main effect adjusted for differences in covariates, and of variance explained by mode. Estimates adjusted for differences in covariates are given in parentheses.

	Prop. Missing Pos. Affect-scale	Total score on 9 items
Mean Main Effect		
Mail	.01 (.01)	6.35 (6.36)
F-t-f	.01 (.01)	6.44 (6.43)
Tel.	.01 (.01)	6.44 (6.44)
Pairwise test (p=.05)	n.a.[a]	n.a.[a]
% Var. Expl.	0.00% (0.01%)	0.05% (0.02%)
P-value Main Eff.	.953	.830
P-value Covars.	.767	.001
P-value adj. Main	.969	.912
N-tot	762	729

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

Summing up, some support is found for the hypothesis that the more anonymous setting in mail surveys leads to more self-disclosure. A slight tendency for more acknowledgment of *negative* feelings in mail surveys is revealed, no differences were found concerning positive feelings. This indicates a clear influence of degree of perceived sensitivity of the topic (cf. Bradburn, 1983; Sudman & Bradburn, 1974).

5.6. Response Styles

Two types of response style have been investigated: acquiescence and extremity.

Acquiescence

Acquiescence is defined as the tendency to answer affirmatively (say yes) with apparent disregard of the content of the question (Couch & Keniston, 1960). Some investigators regard acquiescence as a subject trait (cf. Bentler, Jackson, & Messick, 1971), but the tendency to agree is not consistently correlated from one type of test content or one type of question to another (Block, 1971, McClendon, 1991; Schuman & Presser, 1981). These findings support the classification of acquiescence as primarily an instrument or methods factor instead of as a trait factor (Rorer, 1965; Scott, 1968), and acquiescence might be more a characteristic of the question and the way or mode by which it is asked than of the respondent (cf. Groves, 1989). Especially the telephone interview, which is characterized by a limited channel capacity and a faster pacing, may induce respondents to use simplified cognitive representations and to resort to a simpler answering scheme. Acquiescence can be the result of applying such a simplified cognitive representation in producing an answer to a specific question (Krosnick & Alwin, 1987; McClendon, 1991). Especially the amount of time a respondent has to consider the question and the answer categories should have a pronounced effect on the complexity of the cognitive processing that produces the answer. In mail surveys where the respondent is in total control of the processing time, acquiescence should be smaller than in either the telephone or the face to face interview mode. It follows that most acquiescence is expected in the telephone condition, less in the face to face condition, and the least in the mail condition. In the literature there is indeed some evidence for the existence of a mode effect on acquiescence; Jordan et al. (1980) detected more acquiescence in a telephone survey than in a face to face survey.

In the Dutch version of the Affect Balance Scale, used in the field experiment, all positively formulated items were balanced by negatively formulated items (Hox, 1986). All 18 items had a two-point no/yes response scale; response cards were not used in the face to face condition. In a balanced scale with an even number of positively and negatively formulated questions, acquiescence or Yeah-saying can be estimated by counting the

number of agree answers. Therefore, for each respondent the number of yes-answers on the Affect Balance Scale was counted, with disregard of the content of the questions. Initially a significant difference between methods was detected, suggesting less acquiescence in the mail survey. However, when differences in self-selection of respondents were taken into account the differences between methods disappear. See Table 5.11.

Table 5.11 An(c)ova on Acquiescence

Total number of yes-answers on a balanced 18-item scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates and for the main effect adjusted for differences in covariates, and of variance explained by mode. Estimates adjusted for differences in covariates are given in parentheses.

	Acquiescence based on 18 items	
Mean Main Effect		
Mail	8.76	(8.83)
F-t-f	9.39	(9.30)
Tel.	9.24	(9.26)
% Var. Expl.	0.90%	(0.55%)
P-value Main Eff.	.040	
P-value Covars.	.000	
P-value adj. Main	.118	
N-tot	717	

The small difference in acquiescence observed can be attributed to the slightly higher number of male and of married respondents in the mail survey. It is interesting that acquiescence is not influenced by differences in data collection procedures as such, and that telephone interviews are not at a disadvantage as was hypothesized. However, from a practical point of view, we should conclude that differences between methods do exist in self-selection of respondents, and therefore also in acquiescence.

Extremity

Extremity is the tendency to check extreme answer categories (e.g., "strongly agree" or "strongly disagree") or to check the extremes of a numerical scale (e.g., the numbers 1 or 5 on a five-point scale) (Scott, 1968).

The data collection method used may influence this tendency through the following two factors. The limited channel capacity and faster pacing of the telephone interview may again lead to a simplified answer scheme. The available processing time should have an effect on the complexity of the cognitive processing that produces the answer. Therefore, in mail surveys where the respondent is in total control of the processing time, potential extremity effects should be the smallest. When only the auditory channel is used the last response category presented is more likely to be recalled than the first one, provided that this answer category is plausible to the respondent. This results in a recency effect or higher endorsement of categories last in the list (see Schwarz et al., 1991).

Mode comparison experiments investigating extremity bias are scarce, but there is some evidence of mode effects (cf. chapter 3). Jordan et al. (1980) found more extremeness in a telephone survey than in a face to face survey. In their comparison they did not distinguish between recency and primacy effects. Groves (1979) also reports a tendency for telephone respondents to choose the more extreme (positive) part of a scale. However, there is no indication for a specific recency effect in telephone surveys as in his comparison the more positive alternative was offered first. This is corroborated by Dillman & Mason (1984) who report a slight tendency in telephone respondents to choose the more extreme positive category, independent of whether it was offered first or last in the list. Their main finding is that both face to face and telephone interviews appear to exhibit more extremeness of response in relation to the mail method (Dillman & Mason, 1984, p. 26), giving some support to the effect of available processing time mentioned above (see also Tarnai & Dillman, 1992). This is also supported by Bishop et al. (1988) who found that response order effects were less likely in mail than in telephone surveys.

The questionnaire used in the field experiment contained five questions on different domains of well-being. Answers could be given on a five-point scale, ranging from "very dissatisfied" to "very satisfied." "Very dissatisfied" was always presented as the first response alternative, "very satisfied" was always presented as the fifth and last alternative. In the face to face condition a response card containing the five possible answers was handed to the respondent while simultanously these response alternatives were

read aloud by the interviewer. In the telephone condition the response alternatives were read aloud and when necessary all five response alternatives were repeated completely.

To measure extremity two indices were constructed: a primacy index and a recency index. For the primacy index the number of "very dissatisfied"-answers on the five well-being questions were counted for each respondent. For the recency index the number of "very satisfied"-answers were counted. Recall, that the same answer categories were used for the five well-being questions, and that in all questions the first response alternative is "very dissatisfied" and the last response alternative is "very satisfied." Therefore, the primacy and the recency index can be confounded by the "real" state of well-being of a respondent. A respondent can answer "very satisfied" because she/he is in fact very satisfied with a certain aspect of life, but can also answer "very satisfied" because she/he has a preference for extreme answers. To control for this confounding, the score on the positive affect scale was used as a covariate. Positive affect was *independently* measured with nine yes/no balanced questions on several domains of happiness and well-being. A high score on this positive affect scale indicates that someone has a general feeling of well-being.

A statistically significant difference between the modes was detected for the recency index in the predicted direction, although the effect was small. Pairwise comparison showed that the telephone condition differed significantly from the mail condition. No statistically significant difference was detected between the face to face and telephone survey, nor between the face to face and the mail survey on the recency index. No statistically significant differences between the three modes were detected for the primacy index. See also Table 5.12.

Table 5.12 An(c)ova on Extremity

Primacy index (total number of response 1) and recency index (total number of response 5) based on five well-being questions each with a five-point answering scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates (gender, marital status and happiness score) and for the main effect adjusted for differences in covariates, and of variance explained by mode. Estimates adjusted for differences in the covariates are given in parentheses.

	Primacy index	Recency index
Mean Main Effect		
Mail	.09 (.09)	1.18 (1.19)
F-t-f	.11 (.11)	1.30 (1.30)
Tel.	.09 (.10)	1.50 (1.49)
Pairwise test (p=.05)	n.a.[a]	T-M
% Var. Expl.	0.01% (0.02%)	1.11% (0.96%)
P-value Main Eff.	.97	.02
P-value Covars.	.00	.00
P-value adj. Main	.91	.03
N-tot	724	724

[a] Not applicable. Pairwise tests were only performed when the overall ANOVA showed significant differences between methods.

The recency-effect found can be the result of the different processes taking place in the three modes, but can also be partly attributed to the self-selection of respondents and the differences in gender and marital status as reported in section 4.8. Furthermore, as stated above respondents can choose the extreme answer "very satisfied" on a well-being question because they *really* feel satisfied or happy. To control for these effects, I reanalyzed the data using analysis of covariance. Gender, three dummy codes for marital status and the score on the independently measured positive affect (happiness) scale were used as covariates. Positive affect and marital status were both statistically significant; happy and married people more often answer "very satisfied." However, correcting for the covariates did not change the conclusion stated. In comparison to respondents in the

mail condition, respondents in the telephone condition still choose the last - extreme positive - response category more often (see Table 5.12).

In sum: no mode differences were detected for acquiescence, but a small recency effect was found. Telephone respondents more often chose the last response category. Because of the limited channel capacity and the faster pacing of the telephone interview both more acquiescence and more extremity were expected in the telephone mode. A possible explanation for the conflicting findings can be the complexity of the questions on which the indices were based. Acquiescence was based on the answers on yes/no questions; the extremity indices were based on the answers to questions with five response categories. When a yes/no question is verbally presented to a respondent it is not too difficult to remember these two answer categories, and there is no necessity to use a simplified cognitive representation and to resort to a simpler answering scheme or algorithm. When more response categories are presented without any visual aid, it is more difficult to keep all categories in mind. As a result, respondents have to fall back on a simplified representation and a response effect under auditory presentation emerges. However, experimental research in which the number of response categories and the general complexity of the questions is manipulated is necessary to decide whether this ad hoc explanation is correct.

5.7. Respondents' Evaluation of Data Collection Method

At the end of the questionnaire the respondents were asked which method they preferred if they were given the choice, how they evaluated the procedure in terms of enjoyment, and whether they experienced the questions asked as threatening.

In all three modes respondents had a marked preference for the method they had just experienced. This effect was stronger for the mail survey (76%) and the face to face survey (68%) than for the telephone survey (44%). Relatively more respondents in the telephone condition as compared to the face to face condition preferred a mail survey. No large differences were found for the no-preference group. See also Table 5.13.

I used loglinear analyses to correct for the differences on gender and marital status between the conditions. Neither gender nor marital status had a significant effect on preference; furthermore the interaction between preference and data collection remained significant (Likelihood ratio chi-square=482.93, df=6, p=.00). Inspection of the parameter estimates for

the interaction term preference by data collection method confirmed the conclusions based on the data in Table 5.13.

Table 5.13 Mode and Preference for Data Collection Method

Cell counts, column percentage and adjusted standardized residuals.

	Mail	Face to face	Telephone	N
Preference:				
Mail	186	27	65	278
	76%	11%	24%	
	15.5	-10.2	-5.2	
Face to face	14	167	45	226
	6%	68%	17%	
	-10.0	16.0	-5.8	
Telephone	3	14	117	134
	1%	6%	44%	
	-8.2	-6.0	13.9	
No Preference	41	37	40	118
	17%	15%	15%	
	0.6	-0.2	-0.4	
N	244	245	267	756

Chi-square=502.50, df=6, p=.00, likelihood ratio chi-square=494.91, p=.00

Respondents did express a very strong preference for the data collection method just experienced. If we ignore these cells, we may find that the remaining cells are independent and that there is no difference in preference for a specific data collection method other than the one just experienced. However, this hypothesis had to be rejected; the quasi-independence model did not fit well (Likelihood ratio chi-square=21.59, df=3, p=.00). Inspection of the residuals of the quasi-independence model showed that respondents in the telephone condition about equally preferred a mail survey or a face to face survey, in the face to face condition more respondents preferred a telephone survey and less respondents chose a mail survey, and in the mail survey more respondents expressed an explicit no preference.

When asked to evaluate the past experience in terms of enjoyment far more respondents in the face to face condition reported that they enjoyed the experience very much, while respondents in the mail survey more often chose the neutral category. See also Table 5.14.

Table 5.14 Mode and Evaluation of Experience

Cell counts, column percentages and adjusted standardized residuals.

	Mail	Face to face	Telephone	N
Very Pleasant	12	29	6	47
	5%	12%	2%	
	-1.1	4.4	-3.2	
Pleasant	72	148	124	344
	29%	61%	47%	
	-6.2	5.8	0.4	
Neutral	148	68	132	348
	60%	28%	50%	
	5.5	-7.0	1.5	
Unpleasant	12	0	4	16
	5%	0%	1.3%	
	3.7	-2.7	-1.0	
Very Unpleasant	2	0	1	3
	1%	0%	0%	
	1.4	-1.2	-0.2	
N	246	245	267	758

Chi-square=92.21, df=8, p=.00, likelihood ratio chi-square=97.62, p=.00

Again, I used loglinear analyses to correct for the differences on gender and marital status between the conditions. As can be seen in Table 5.14 the extreme response categories very pleasant and very unpleasant were rarely chosen. To avoid statistical problems in the loglinear analyses, adjoining categories were joined, which resulted in a three-point scale with the categories pleasant, neutral and unpleasant.

A significant effect of marital status on enjoyment was observed (Likelihood ratio chi-square=16.68, df=6, p=.01), but the interaction between expressed enjoyment and data collection remained significant (Likelihood ratio chi-square=85.30, df=4, p=.00). Inspection of the parameter estimates for the interaction term of enjoyment by mode confirmed the conclusion that far more enjoyment was expressed at the end of the face to face interview, while at the end of the mail survey respondents evaluated the experience more often as neutral or slightly unpleasant.

Interestingly, no differences in experienced questionnaire threat were observed across methods (see also Table 5.15). Although respondents do not differ between the modes in experienced questionnaire threat, they do report differences in enjoyment. A possible explanation of this phenomenon can be sought in the differences in self-disclosure between the methods. Respondents in the mail situation reported more feelings of extreme loneliness than in either the face to face or telephone condition. According to the mood induction theory a negative affective state could be induced by reporting feelings of loneliness. This will influence the responses on the more general evaluative question on enjoyment of the whole question-answer process (cf. Gouaux, 1971). In accordance with this assumption I did observe a negative correlation between expressed enjoyment and reported loneliness (r=-0.13, p=.00). However, this effect was not large enough to explain away the differences in reported enjoyment between the methods. When avowed loneliness is used as a covariate in a loglinear analysis the independence model had to be rejected (Likelihood ratio chi-square=83.41, df=3, p=.00). Further inspection of the residuals showed that more respondents in the mail condition gave a neutral or negative evaluation than could be expected under independence and far more respondents in the face to face condition gave a positive evaluation.

Table 5.15 An(c)ova on Questionnaire Threat Scale

Proportion missing data and total score on a five-item questionnaire threat scale. Reported are means and p-values for the main effect, p-values for the total effect of the covariates (gender and marital status) and for the main effect adjusted for differences in covariates, and the variance explained by mode. Estimates adjusted for differences in covariates are given in parentheses.

	Prop. Missing Quest. threat scale	Total score on 5 items
Mean Main Effect		
Mail	.04 (.04)	1.37 (1.37)
F-t-f	.06 (.06)	1.33 (1.33)
Tel.	.07 (.07)	1.56 (1.56)
% Var. Expl.	0.57% (0.50%)	0.38% (0.37%)
P-value Main Eff.	.113	.295
P-value Covars.	.171	.654
P-value adj. Main	.147	.307
N-tot	762	649

5.8. Summary

To assess the data quality five indicators were used: the number of responses to open questions, item missing data (item nonresponse), differences in response distributions on sensitive topics (income, loneliness, and well-being), acquiescence and preference for extreme answer categories (extremity). Furthermore, the way respondents evaluated their experience is compared over modes. Small differences were observed between the methods. A concise summary of the main results is presented in Table 5.16.

Table 5.16 Concise Summary of Main Results: Univariate Mode Effects

A Mail (M), Telephone (T) and Face to face (F) survey are evaluated on several criteria. For each criterion a prediction and the result of the statistical test are given in the first and second column. ">" indicates a higher score on the criterion and "<" indicates a lower score. For example M>F on the indicator precision means *more* precision (i.e., better performance) in the mail survey, but F<M on the indicator item missing data means more missing data (i.e., worse performance) in the mail survey. A reference to the appropriate section of this chapter is given in the last column.

Criterion	Prediction	Result Ancova	Section
Open questions	F > T > M	F=T, F>M, T>M (interview best)	5.3
Item miss. data:			
Overall	F < T < M	F=T, F<M, T<M (mail most missing)	5.4
Income question:			
Willingness	M > F,T	M = F = T	5.5
Precision	M > F > T	M > F, T (mail more precise)	5.5
Sensitive topics:			
Self-disclosure	M > T > F	F=T, M>F, M>T (mail more open)	5.5
Acquiescence	M < F < T	M = F = T	5.6
Extremity:			
Primacy	M < F < T	M = F = T	5.6
Recency	M < F < T	M<T, F=T, M=F (mail least recency)	5.6

Note. This is a concise summary of the results of the statistical tests. When the modes did *not* differ on a significance level of 0.05 this is indicated in the table by "=". The equal sign does mean that there are no statistical differences between the modes, not that the results are completely identical. For a more detailed discussion of the results see the appropriate section in this chapter.

The mail survey resulted in more item nonresponse, but also in more self-disclosure on sensitive topics and a tendency to report income more precisely (i.e., in guilders and cents). No differences between the face to face and telephone surveys were detected on these point No consistent differences between modes were found on open questions. Also, no clear mode differences were detected for acquiescence, but a small recency effect was found. Respondents in the telephone condition had a tendency to choose

the extreme positive answer more often than respondents in the mail condition.

In general, no consistent differences between the telephone and the face to face survey were detected. These findings are in accordance with results from other *recent* mode comparisons, since the earlier differences between face to face and telephone surveys have become smaller over time (cf. De Leeuw & Van der Zouwen, 1988, also chapter 3). These results support Groves' conclusion that the most consistent finding in studies comparing responses in face to face and telephone interviews is the lack of differences in results obtained through these two modes (Groves, 1989, p. 551).

The main differences detected in this study were between the mail survey on the one hand and the two interview surveys on the other hand. In general, it is somewhat harder to have people answer questions in the mail survey as the higher item missing data rate indicates, but when the questions are answered, the resulting data are of better quality (more self-disclosure, more precision). The differences between all three methods were very small and the findings suggest a dichotomy between self-administered questionnaires and interview strategies (both telephone and face to face), confirming the main conclusions of the meta-analysis reported in chapter 3.

The presence of an interviewer, either in person or over the telephone, seems to be an important factor. The interviewer can motivate a respondent and probe for additional answers. At the same time, the presence of an interviewer may lead to problems of self-presentation, especially with sensitive questions. The greater recency effect detected in telephone surveys, suggests the influence of a second factor: the way the information is transmitted. Visual presentation of the information, in a self-administered questionnaire or with special response cards during an interview, may relieve the cognitive burden of the respondent and may lead to fewer response effects.

When asked about their preferences a majority of respondents chose the method they had just experienced. However, relatively few respondents in the telephone condition, compared to the other two data collection conditions, preferred the experienced method. Similar results have been found by Groves and Kahn (1979). Groves (1989) suggests that the physical presence of the interviewer in the face to face interview magnifies the reported preference for the method experienced. However, in the mail survey this effect could not be observed; a remarkably high number preferred mail surveys. It seems safe to assume that preferences for a

specific survey method are multidimensional concepts. For instance, although all methods scored equally on experienced questionnaire threat, their were differences in reported enjoyment. Different subgroups can prefer a method for different reasons; while some prefer a face to face interview for the pleasant social contact, others might prefer a mail survey for the absence of contact. To disentangle these effects a more refined method than a single preference question is required.

Furthermore, respondents in the mail survey condition gave more comments when asked for any comments at the end of the questionnaire. Together with the higher reported pleasure in the interview condition, this indicates the better and positively valued opportunity for respondents to elucidate their responses in an interview situation. When using a mail survey it is wise to give respondents opportunities to react or comment either in writing on the questionnaire or by telephone to the researcher in charge (see also Dillman, 1978).

CHAPTER 6

DATA QUALITY II: RELIABILITY AND SCALABILITY

Wondering in idle moments whether an increased precision might perhaps be rather better . . . . .

Maurice G. Kendall, Hiawatha designs an experiment, American Statistician, 1959, 13, 23-24

6.1. Introduction

Little attention has been given in mode comparisons to psychometric indicators of data quality. For example, 67 articles and papers were reviewed in the meta-analysis in chapter 3; of these 67 only four articles reported comparisons on some indicator of psychometric reliability. In a health community survey, Aneshensel, Frerichs, Clark and Yokopenic (1982) observed no significant difference between face to face and telephone interviews concerning the reliability of a multiple item depression scale (coefficient alpha was 0.91 in the face to face condition and 0.90 in the telephone condition). The other three studies (Herman, 1977; O'Toole et al, 1986; Rogers, 1976) all focus on the consistency over time of answers on specific questions and did not investigate multiple item scales (see also chapter 3).

Mode effects on both psychometric reliability and scale properties were investigated by Van Tilburg and De Leeuw (1991). They did a secondary analysis on the data of a multiple item loneliness scale collected in six Dutch surveys. Different interview modes were used for the data collection: three surveys used a self-administered paper questionnaires, two surveys employed face to face interviews, and one survey collected the data with a computer assisted self-administered questionnaire (a "telepanel"). In this study, both the internal consistency and the scalability tend to be higher in the self-administered surveys.

Little is known about the influence of the data collection method on the psychometric properties of multiple item scales. This is surprising, because the importance of well-operationalized and reliably measured concepts has been strongly emphasized in social sciences. (For an overview see Hox and

De Jong-Gierveld, 1990). Multiple item scales have traditionally been extensively used in psychological and educational research. Also in social sciences in general, multiple questions or indicators are frequently used to measure one underlying concept. As a result, in surveys on such different topics as mental health, well-being and social change, short multiple item scales are used (for example, see De Jong-Gierveld, 1987; Dykstra, 1990; Andrews & Withey, 1978). Therefore, it is important to know how robust multi-item scales are against data collection effects.

In the following sections the influence of mail, telephone and face to face survey methods on several psychometric properties of multiple item scales is investigated. First, a short description is given of the scales used. This is followed by a discussion of expected mode differences. In the subsequent part the influence of data collection method on psychometric reliability is described, using classical test theory. Next, the effects on scalability are investigated, using non-parametric item response theory. Finally, the potential influence of data collection method on the occurrence of aberrant or unexpected individual response patterns is explored.

6.2. The Multiple Item Scales

To investigate the influence of data collection method on scale properties of multiple question scales, four well-known scales were used in the questionnaire: De Jong-Gierveld's Loneliness scale (De Jong-Gierveld & Kamphuis, 1985), a condensed form of Brinkman's Self-evaluation scale (Brinkman, 1977; Dykstra, forthcoming), and balanced extensions of Bradburn's Affect Balance Scale measuring respectively Positive and Negative Affect (Bradburn, 1969; Hox, 1986).

The 11-item loneliness scale consists of both negative and positive items. Each item has three response categories (i.e, "yes," "more or less," and "no"). The self-evaluation scale in its condensed form has eight items, again with three response categories. The extended affect balance scale has a total of 18 dichotomous yes/no items. Each negatively formulated item is balanced by a positively formulated one. The affect balance scale consists of two subscales: one measuring "positive affect" or "happiness" (nine items) and one measuring "negative affect" or "unhappiness" (nine items). A score of 1 was assigned when the answer on an item indicated the concept measured by the scale, otherwise a score of 0 was assigned. For instance, a score of 1 on a positive affect item indicates happiness, and a score of 1 on a negative affect item indicates unhappiness. "No-answers" and "do-not

knows" were assigned a missing value. The items on the loneliness scale and the self-evaluation scale were dichotomized; the "more-or-less" responses were not viewed as neutral responses, but as indicators of loneliness or a positive self-evaluation (see also Van Tilburg & De Leeuw, 1991). Examples of items of these four scales are given in Appendix B.

All four scales were used in the mail survey condition, the paper-and-pencil telephone condition, the CATI condition, and the face to face interview condition. No response cards were used during the face to face interviews. The paper-and-pencil telephone interviews and the computer assisted telephone interviews differed on one major point. In the CATI condition it was possible to randomize the questions within a multiple item scale for **each** interview. By randomizing questions within scales systematic context effects are avoided, making it possible to investigate how far respondents use the immediately preceding questions as a cognitive clue to produce consistent answers. This prospect was the main reason for including a small number of computer assisted telephone interviews.

6.3. The Potential Impact of Mode on Psychometric Properties

The specific data collection mode used in a survey, can influence the reliability and scalability of the measurement instruments. It can also influence the individual response patterns on a multiple item scale. Mail, telephone, and face to face surveys differ in their impact on the cognitive and communicative processes that underlay question answering.

An important difference between self-administered procedures and interviews is the recording process (see also the discussion on media related factors in section 2.2). In self-administered questionnaires the respondent, and not the interviewer, writes down the answer. This provides the respondent with an extra check on the correctness of the answer (Galtung, 1967), and gives the respondent total control over the pace of a question-answer sequence. In interview situations the pace is determined by both respondent and interviewer. However, traditional rules of behavior dictate that in a telephone conversation the initiator (which is the interviewer) controls the channel (cf. Argyle, 1973), while in a face to face conversation a more balanced situation is created. This could be one reason for the often noted faster pace in telephone interviews (cf. Groves, 1989, Groves & Kahn, 1979; Körmendi & Noordhoek, 1989; Sykes & Collins, 1988).

The faster pace of the telephone interview was also observed in this data set[11]. The average actual interview time (i.e., time from first question to last answer) for the face to face interview was 31 minutes, while for the paper-and-pencil telephone interview the average time was 24 minutes and for CATI 25 minutes. For CATI the interview-time was also registered by the computer system; the average interview-time according to the system was again 25 minutes. The correlation between the time as recorded by the interviewer and by the system was 0.90.

Time pressure has been shown to increase "top of the head phenomena": respondents just answer with the first thing that comes to mind (cf. Schwarz et al., 1991). A slower pace will give respondents more time to give deliberate consideration to the meaning of a question and to evaluate or edit their provisional answer, resulting in less random error in the answers. A mail survey provides a respondent with total control over the pace of the question-answer sequence, a telephone survey provides a respondent with the least control. Therefore, I expect the highest reliability and scalability in the mail survey, and the lowest in the telephone survey. Likewise, an effect of data collection method on the individual response patterns is expected, resulting in respondents with more aberrant response patterns in the interview conditions than in the mail survey condition.

A second factor that can influence the quality of a multiple item scale is the opportunity the respondents have to relate different questions to each other, and the opportunity they have to relate their answers to these questions to one another (see also the discussion on information transmission in section 2.3). A self-administered questionnaire allows a respondent to go back and forth between the questions. The respondent, therefore, sees the context in which an item fits and sees that a certain item is one in a series of items on the same topic. In an interview the sequential presentation of the questions gives the respondent less opportunity to relate their answers to different questions. If respondents have a tendency to deliberately relate questions and make their answers consistent this would lead to respondents with less aberrant response patterns in a mail survey than in face to face and telephone interview surveys. Furthermore, it should also result in a higher reliability and scalability of multiple item scales in a mail survey.

---

[11] Because the distributions of the variable "interview-time" were highly skewed I performed a normalizing transformation and reanalyzed the data using analysis of variance on the transformed data (cf. Kirk, 1968). The difference in pace between the methods remains highly significant (p=.00).

Summarizing, two factors -pace of interview and opportunity to deliberately relate different questions- can influence the consistency of response patterns on related questions and the psychometric quality of multiple question scales. To disentangle the influence of these two factors a small CATI experiment was conducted, in which questions were randomized within scales for each CATI-interview. Recall, that the duration of the paper-and-pencil telephone interviews (on average 24 minutes) did not differ significantly from the duration of the computer assisted telephone interviews (on average 25 minutes). However, while the question order was the same for all respondents in the paper-and-pencil condition, the question order was different for respondents in the CATI-condition, making it possible to investigate how far respondents use the immediately preceding questions as a cognitive clue to produce consistent answers.

6.4. Psychometric Reliability

In this section models and procedures, which are based on classical psychometric test theory (cf. Lord and Novick, 1968), are used to investigate the influence of data collection method on the quality of multiple item scales. For the four multiple item scales "Loneliness," "Self-evaluation," "Positive Affect" and "Negative Affect" Cronbach's coefficient alpha was computed as an indicator for scale reliability. The results are shown in Table 6.1.

Coefficient alpha, proposed by Cronbach (1951), gives a lower bound for the reliability (i.e., the squared correlation between observed scores and "true" scores) on a multiple item scale. Coefficient alpha can be interpreted as the proportion "true" score variance in the observed scores. Nunnally (1967, p. 226) recommends values for coefficient alpha of 0.70 and higher as an acceptable value for research; lower values with a minimum of 0.50 are only to be tolerated in early stages of test construction. When important decisions are based on individual test scores (e.g., in psychological testing) a minimum value of 0.90 is mandatory.

Table 6.1 Psychometric Properties by Method

Reliability (Cronbach's coefficient alpha) for the loneliness-scale (11 items), the self-evaluation scale (8 items), the positive affect scale (9 items), and the negative affect scale (9 items).

Scale	Mail alpha (n)	FtF alpha (n)	Tel. alpha (n)	CATI alpha (n)
Loneliness	.84 (248)	.83 (239)	.81 (263)	.79 (75)
Self Eval.	.78 (251)	.76 (236)	.72 (263)	.78 (76)
Pos. Aff.	.74 (246)	.65 (230)	.58 (252)	.57 (75)
Neg. Aff.	.73 (246)	.71 (240)	.68 (258)	.64 (77)

In Table 6.1. the reliability values are depicted. There are small differences in coefficient alpha across the methods. The differences are generally in the expected direction with the highest internal consistency for scales in the mail condition and the lowest in the telephone condition. A multiple group significance test according to Hakstian and Whalen (1976) showed that only for the Positive Affect Scale the observed mode differences were statistically significant at the .05-level (p=.00)[12]. Subsequent pairwise tests (Feldt, 1969) revealed that the mail survey resulted in a higher reliability coefficient than the face to face survey (p=.03), the paper-and-pencil telephone survey (p=.00), and the CATI survey (p=.02). No statistically significant differences were observed between the face to face interviews and both forms of telephone interviews, nor between the paper-and-pencil and the computer assisted telephone interviews (smallest p=.18).

Differences in reliability between groups can be the result of group differences on one or two items. To assess the quality of the individual items the corrected item-total correlation (r_{it}), and the contribution (f_i) of an individual item to the signal-noise ratio were estimated for each group separately. The corrected item-total correlation or item rest correlation is the correlation between a specific question that belongs to a multiple item scale and the total score for that scale computed without that particular question. This index indicates how strongly a specific question measures the concept measured by the total multiple item scale. The signal-noise ratio is closely related to the reliability and is defined as the ratio between the

[12] To avoid capitalization on chance I used the sequentially rejective Bonferroni test as proposed by Holm (1979).

"true-score" variance and the "error-score" variance (Nunnally, 1967). The index f_i indicates how much a specific individual item contributes to the signal-noise ratio of the total multiple item scale (cf. De Groot & Van Naerssen, 1969).

Table 6.2 Reliability Analysis: Summary Statistics by Method

Mean, standard deviation, minimum, and maximum for the corrected item-total correlation (r_{it}) and item signal-noise ratio (f_i) over all 37 items (loneliness, self-evaluation, positive affect and negative affect) by method.

Corrected item-total correlation (r_{it})

	Mail	FtF	Tel.	CATI
Mean	.46	.43	.38	.38
St. deviation	.10	.12	.12	.13
Minimum	.23	.14	-.01	.17
Maximum	.60	.67	.57	.64

Item signal-noise ratio (f_i)

	Mail	FtF	Tel.	CATI
Mean	.39	.33	.28	.28
St. deviation	.22	.24	.19	.23
Minimum	-.13	-.14	-.31	-.04
Maximum	.78	.90	.59	.86

Inspection of these indices showed that items that are well-behaved from a psychometric point of view are generally well-behaved in all conditions. For instance, items that have a high corrected item-total correlation in the mail condition, also have a relatively high corrected item-total correlation in the face to face and telephone conditions. The Spearman rank correlations between conditions for the corrected item-total correlations vary from a minimum of .62 to a maximum of .83. The Spearman rank correlations between modes for f_i, an item's contribution to the signal-noise ratio, vary between .58 and .76. It should be noted however, that there is a slight tendency for items to have a higher corrected item-total correlation and a higher contribution to the signal-noise ratio in the mail condition and lower ones in the telephone condition. This can be more easily seen in Table 6.2,

which presents the summary statistics for these indices over all 37 items. Note that the corrected item-total correlation and the contribution to the signal-noise ratio in the CATI-condition are only based on 75 persons, and are therefore less stable than the same indicators for the other conditions, which are based on a minimum of 230 persons per condition.

Also, as some differences in self-disclosure between respondents on the mail survey and the interview surveys were detected (see section 5.5), it is conceivable that the more extreme items are subject to differential self-disclosure and so cause group differences in reliability. To investigate this possibility I computed the proportion affirmative answers or item p-value (p) of all scale items for each data collection condition separately. These revealed a slight overall tendency of more acknowledgment of negative feelings and attributes in the mail survey as can be concluded from the proportion affirmative answers p, but this tendency is the same for all items (cf. De Leeuw, 1991).

In sum: small differences were found between the methods in the expected direction: the mail survey showed the best results, while the telephone survey was the least satisfactory. The explicit randomization of the items in the CATI-condition did not have a clear influence on the reliability; no differences were found between the paper-and-pencil and the CATI-condition.

6.5. Scalability

Item response theory

Classical psychometric test theory is mainly concerned with the detection of measurement error. A high reliability of the total test score is therefore an important quality criterion. Modern psychometric test theory emphasizes the explanation of test behavior through the development of latent trait models. Latent trait models assume that a person's responses can be explained by a number of traits (e.g., loneliness). These traits are called latent because they are unobservable and conclusions about them have to be reached by referring to the observable consequences of the model (e.g., answers to questions on a multi-item scale).

Modern psychometric measurement theory is often referred to as "Item Response Theory" or IRT. Wright and Stone (1979) characterize item response theory as a theory that describes what happens when a person

encounters an item. Sijtsma (1988) gives an even more daring description and states that item response theory is not only a (psychometric) test theory. Item response theory is also a formalized psychological theory, which explains answering behavior by taking into account attributes of both persons and questions. Person attributes are usually the traits, attitudes or abilities measured by means of the multiple item scale. Question attributes are, for instance, the "item difficulty," which in classical test theory is defined as the proportion persons who receive the score 1 on a dichotomous scored 0/1 item. Together these person and question attributes determine the probability of the selection of a specific answer from a set of possible answer categories. Important concepts in the Item Response Theory are the Item Characteristic Curve (ICC) and the Person Characteristic Curve (PCC). For dichotomously coded questions the ICC provides the probability of persons answering the question affirmatively or correctly (i.e., coded 1) as a function of the person attribute or person characteristic (i.e., the latent trait). In a similar way the PCC provides the probability of items answered correctly by a person as a function of the item difficulty.

Two IRT-models that have been given much attention in applied research during the last decade are the Rasch model and the Mokken model (cf. Meijer, Sijtsma & Smid, 1990). The Rasch model and the Mokken model are both unidimensional cumulative models: both models assume that there is only one latent trait underlying the answers and that the probability of a positive or a correct answer for each item is a non-decreasing function of this latent trait value. That is, the Item Characteristic Curve (ICC) is non-decreasing. The two models differ mainly in the assumptions they make about the shape of the functions relating the response probabilities to the person and the question characteristics. It should be kept in mind that both models are probabilistic models: a person may produce a correct or positive answer to a "difficult" question and a negative answer to an "easier" question.

In addition to the reliability analysis I performed both a Rasch- and a Mokken analysis. The very restrictive Rasch model did not fit in most cases. For the results of these analyses, see De Leeuw (1991). In the remaining part of this chapter I concentrate on the Mokken model.

Scalability according to the Mokken model

The Mokken model is a nonparametric probabilistic model in the Item Response Theory, developed by Mokken (1971), and elaborated by Mokken

and Lewis (1982), Molenaar (1982) and Sijtsma (1988). The Mokken model is a nonparametric approach to latent trait theory because the Item Characteristic Curves are not parametrically defined. Also, no assumptions are made concerning the distribution of the latent trait. But, unidimensionality and local stochastic independence are assumed. The other assumption concerns the Item Characteristic Curves: it is assumed that there is monotonicity in the latent trait (a higher value implies a non-decreasing probability of answering positively to a question). This is known as the Mokken model of monotone homogeneity. When the assumption is added that there is monotonicity in the item difficulties, this results in the Mokken model of double monotonicity. Together the two assumptions of monotonicity imply that the ICC's do not intersect. The Mokken model of double monotonicity makes no other assumptions for the ICC's; they may coincide or touch and may all have a different shape, as long as they do not intersect.

The nonparametric Mokken model does not produce numerical estimates of person and item parameters. Therefore, the total or sum score is used as an estimator for rank ordering persons. Also, the items can be ordered according to their difficulty, that is, the proportion of persons giving a "positive" or "correct" answer to a question (Meijer, Sijtsma & Smid, 1990).

As an overall indicator of Mokken scalability Loevinger's H was computed for each of the four multi-item scales[13]. This overall scalability coefficient should be nonnegative, but Mokken (1971) recommends the value H=.30 as a practical lower bound. In addition to the scalability index H, its standard error (SE) was computed (Mokken, 1971). The results are summarized in Table 6.3.

[13]Actually this only constitutes a necessary condition for monotone homogeneity. Additional visual inspection of the P-matrix did not reveal many severe violations of double monotonicity. Clear violations were only detected for the positive and negative affect scales in the telephone condition.

Table 6.3 Mokken Scalability Analysis by Data Collection Method

Mokken Scalability: Loevinger's H for the total scale and the standard error (SE) for H.

Scale	Mail		FtF		Tel.		CATI	
	H	S.E	H	S.E.	H	S.E.	H	S.E.
Loneliness	.44	.03	.40	.04	.36	.03	.34	.07
Self Eval.	.45	.03	.45	.04	.37	.04	.49	.06
Pos. Aff.	.36	.03	.27	.04	.22	.03	.22	.06
Neg. Aff.	.36	.04	.34	.03	.30	.03	.24	.06

There are small differences in the overall H across the methods. The differences are generally in the expected direction with the highest values in the mail condition and the lowest in the telephone condition. A multiple group comparison (Marascuilo, 1966) showed that again only for the Positive Affect Scale the observed differences were statistically significant at the .05-level (p=.00)[14]. Subsequent pairwise tests revealed that the mail survey resulted in a higher overall scalability index than the face to face survey (p=.04), the paper-and-pencil telephone survey (p=.00), and the CATI survey (p=.03). No statistically significant differences were observed between the face to face and the telephone interviews (paper & pencil and CATI), nor between the paper-and-pencil and the computer assisted telephone interviews (smallest p=.35).

Also, for each question in a scale the item value H_i was computed; this H_i for individual questions should be non-negative. Again, items that are well-behaved from a psychometric point of view, are well-behaved in all conditions: items that have a high value for H_i in the mail condition, also have a relatively high H_i in the face to face and telephone conditions. The Spearman rank correlations between survey conditions varied for H_i from a minimum of 0.68 to a maximum of 0.84. It should be noted however, that there is a slight tendency for items to have a higher scalability index H_i in the mail condition and lower ones in the telephone condition. This can be more easily seen in Table 6.4, which presents the summary statistics for the individual item H_i over all 37 questions.

[14] To avoid capitalization on chance I used the sequentially rejective Bonferroni test as proposed by Holm (1979).

Table 6.4 Mokken Analysis: Summary Statistics by Method

Mean, standard deviation, minimum, and maximum for item H_i over all 37 questions (loneliness, self-evaluation, positive affect and negative affect) by method.

	Mail	FtF	Tel.	CATI
Mean	.41	.37	.31	.32
St. Dev.	.09	.10	.10	.13
Minimum	.22	.14	-.01	.13
Maximum	.63	.55	.47	.66

Table 6.4 shows that only in the telephone condition the lowest H_i-value was negative. It concerned one single question from the Negative Affect Scale; the H_i-values for all other questions are non-negative. For a detailed overview see De Leeuw (1991).

Besides the Mokken scalability, the precision of measurement under the Mokken model (rho) was also examined for each data collection method (Sijtsma & Molenaar, 1987). The results are presented in Table 6.5.

Table 6.5 Mokken Reliability Analysis by Data Collection Method

Reliability under the Mokken model; rho and number of respondents for each scale

	Mail		FtF		Tel.		CATI	
Scale	Rho	N	Rho	N	Rho	N	Rho	N
Loneliness	.86	248	.84	239	.81	263	.81	75
Self Eval.	.80	251	.77	236	.72	263	.80	76
Pos. Aff.	.76	246	.66	230	.61	252	.57	75
Neg. Aff.	.74	246	.72	240	.70	258	.65	77

Again, the same pattern emerges: the highest values for rho are found in the mail condition, the lowest in the telephone condition.

In sum: the results of the Mokken analyses are in accordance with the results derived from the classical psychometric test theory discussed in section 6.4. When differences between methods were discovered, these

differences were small. All survey methods performed moderately well with the mail survey showing the best results, while the telephone survey was the least satisfactory. No clear differences were found between the paper-and-pencil telephone interviews and CATI.

6.6. Person Fit

In this section procedures based on person fit research are used to investigate the influence of data collection method on the quality of four multi-item scales.

Person fit indices

Person fit research, which originated in the field of psychological and educational testing, is concerned with the investigation of individual response patterns. In person fit research persons with unexpected or aberrant response patterns with respect to a test model or with respect to other response patterns in the sample are identified and further examined. For example, if a student answers 8 out of a total of 10 items correctly, one expects that s/he will have missed the two most difficult ones. If, instead, the two easiest questions are answered incorrectly, the item response pattern is totally unexpected. Between these two extremes, there is a wide range of possible item response patterns. Several indices of person fit have been developed to indicate the degree of aberrance of an individual response pattern.

Two groups of person fit indices can be distinguished. The first group consists of indices that are based on the assumptions of parametric IRT-models, such as the Rasch model. For an overview, see Kogut (1986); see also Molenaar and Hoijtink (1990). The second group consists of indices that evaluate a response pattern given the assumptions of a nonparametric IRT model (Sijtsma, 1988; Van der Flier, 1982), or by means of statistics based on the group to which a person belongs (Harnisch & Linn, 1981; Tatsuoka & Tatsuoka, 1982). For a detailed overview, see Meijer (1990). The strict assumptions of the Rasch-model were not met in this data set (see paragraph 6.5), and person fit indices based on these assumptions could not be used. Among the remaining indices, the U3-index (Van der Flier 1980) is one of the best documented and tested. Therefore, the U3-index is used in the final analyses in the next section.

Person fit and data collection method

According to Van der Flier (1980, 1982) a response pattern of a person on a multiple item scale is called aberrant when it has a low probability of occurrence in comparison with the other response patterns of persons with the *same* total score. To decide whether an individual response pattern is aberrant Van der Flier proposed the U3-index. U3 equals zero (its minimum value) when a response pattern equals the perfect Guttman pattern. U3 equals one (its maximum value) when a response pattern equals a reversed Guttman pattern. A relative high value of U3 indicates that a response pattern deviates from the other response patterns. Furthermore, Van de Flier (1980) showed that U3 is approximately normally distributed, given the null hypothesis that the response behavior fits the order of the item difficulties in the total score group the individual respondent is compared to.

The scores on the person fit index U3 were computed for the respondents within each data collection separately[15] (see also Meijer & De Leeuw, 1992). This was done for each of the four scales (i.e., the loneliness scale, the self evaluation scale, the positive affect scale, and the negative affect scale). When respondents had either the minimal total score of zero or the maximum total score possible on a multiple item scale, a missing value was assigned. In those cases the response pattern is totally predictable, and U3 is undefined.

An analysis of variance was performed with the scores on Van der Flier's U3-index as dependent variable and data collection method as independent variable. The results are summarized in Table 6.6. As the correction for differences in gender and marital status of the respondents in the four conditions did not influence the results, the uncorrected figures are given.

[15] The U3-score was computed with a program for the computation of person fit scores developed by Rob Meijer of the Department of Industrial and Organizational Psychology, Vrije Universiteit, Amsterdam.

Table 6.6 Anova on Person Fit Index U3

Four scales are investigated: loneliness (11 items), self-evaluation (8 items), positive affect (9 items), and negative affect (9 items). Reported are means and p-values for the main effect of data collection mode. As an effect size indicator percentage of variance explained by mode of data collection is given.

	Loneliness	Self-eval.	Pos.Af.	Neg.Af.
Mean Main Effect				
Mail	.27	.16	.22	.24
F-t-f	.31	.16	.22	.23
Tel.	.36	.19	.23	.25
CATI	.34	.16	.19	.29
% Var. Expl.	1.92%	0.66%	0.27%	0.66%
P-value Main Eff.	.01	.25	.61	.22
N-tot	606	632	673	674

For the loneliness scale the mean value of U3 in the mail survey condition is lower than in the other interview conditions, indicating less extreme aberrant patterns in the mail survey as was expected. No statistically significant differences between the data collection methods could be detected for the self-evaluation scale, the positive affect scale, and the negative affect scale. Subsequent pairwise tests for the loneliness scale showed that only the difference between the mail survey condition and the telephone interviews (p=.01) reached statistical significance at the 5%-level.

6.7. Summary

The four data collection procedures were compared on psychometric reliability and Mokken scalability. Four multiple item scales were used in this investigation: an eleven-item loneliness scale, an eight-item self-esteem scale, a nine-item positive affect scale, and a nine-item negative affect scale. Small differences were observed between the methods. A concise summary of the main results is given in Table 6.7.

Table 6.7 Concise Summary of Main Results: Psychometric Mode Effects

A Mail (M), Telephone (T), CATI (C) and Face to face (F) survey are evaluated on several criteria. For each criterion a prediction and the result of the statistical test are given in the first and second column. The sign ">" indicates a higher score on the criterion (e.g., better performance) and "<" indicates a lower score (e.g., worse performance). For example M>F on the indicator reliability means higher reliability (i.e., better performance). A reference to the appropriate section of this chapter is given in the last column.

Criterion	Prediction	Result	Section
Psychometric reliability (alpha)	M > F > T >C	M > F,T,C F=T=C (positive affect only)	6.4
Mokken scalability (Loevinger's H)	M > F > T >C	M > F,T,C F=T=C (positive affect only)	6.5
Person Fit (U3)	M > F > T > C	M > T,C F=T=C, M=F (loneliness only)	6.6

Note. This is a concise summary of the results of the statistical tests. When the modes did not differ on a significance level of 0.05 this is indicated in the table by "=". The equal sign does mean that there are no statistical differences between the modes, not that the results are completely identical. For a more detailed discussion of the results see the appropriate section in this chapter.

Only in a limited number of cases did I detect statistically significant differences at the .05-level. When a difference between modes was significant it always indicated a (small) difference between the mail survey condition and the other three conditions. However, a small (not significant) trend could be noticed in the predicted direction. All survey methods performed moderately well on the reliability and scaling criteria: the mail survey showed the best results, while the telephone survey was the least satisfactory. From a strictly psychometric view the mail survey should be considered as slightly better. Also, from a psychometric point of view, the performance of the four scales was only moderately good for all four modes.

When individual response patterns were investigated, a small mode effects could be distinguished. Respondents had a slight tendency to have less extreme aberrant response patterns in the mail survey.

From a practical point of view these results are reassuring: only very small effects were found. From a theoretical point of view, these results are slightly disappointing. Two important factors were distinguished which could influence the psychometric data quality: time pressure and opportunity to relate different questions to each other. The mail survey, in which the time pressure is the least and the opportunity to relate responses to different questions the greatest, did show better results. The CATI-condition in which the average time pressure equaled the telephone condition, but in which the questions were randomized within scales, did not give statistically different results. There was a slight trend for the not randomized paper-and-pencil telephone interview to produce slightly better data, indicating that the opportunity to relate different questions has some influence. Further experimentation seems necessary. Recent developments in computer assisted interviewing, and especially in computer assisted self-administered testing makes it possible to design strictly controlled experiments in which time pressure and question order can be independently manipulated at several levels.

CHAPTER 7

DATA QUALITY III: A MULTIVARIATE APPROACH

. . . they had 27 8×10 colored glossy pictures with circles and arrows and a paragraph on the back of each one, explaining what each one was, to be used as evidence . . .
Arlo Guthrie, Alice's Restaurant

7.1. Introduction

Although the influence of data collection method on the quality of the data has received considerable attention in survey research, published mode comparisons were mainly restricted to the analysis of univariate distributions (for an overview see chapter 3). Only a few studies investigated psychometric indicators of data quality (cf. chapter 6), and hardly any attention has been given to the potential effect of the mode of data collection on the empirical estimates of the relationships between variables.

In the social and behavioral sciences the multivariate analysis of relationships between variables (e.g., path analysis, factor analysis) is an important and often used research tool. A potential influence of the data collection method on the estimated coefficients representing relationships between variables and corresponding model parameters, would threaten the comparability of research conclusions and would have severe consequences for mixed-mode research (i.e., a research project in which more than one data collection method is used). Therefore, there is a limit to the growth of the acceptance of mail and telephone surveys as alternatives for the face to face interview and to the growth of the acceptance of mixed mode research, pending further demonstrations of the robustness of multivariate statistics against mode effects.

Two rival hypotheses can be formulated about the effect of the data collection method on the estimated relationships between variables.

The first one states that, even if mode effects may exist when univariate statistics are compared, this does not necessarily imply an effect on multivariate statistics, such as covariances. The reasoning is that the observed differences between the marginals of the univariate distributions

just reflect a shift of position of a specific variable on the x- or y-axis, but that the shape of the bivariate distribution of any two variables -as reflected in the bivariate scatterplot- will not be altered. This is sometimes called the "form-resistant correlation hypothesis" (cf. Krosnick & Alwin, 1987). This reasoning leads to the hypothesis that, even if mode effects are detected in marginal distributions, multivariate statistics will remain fairly stable.

The second hypothesis derives from statistical distribution theory. This theory states that, in general, higher order moments are less stable than first order moments. This implies that rather small differences in the responses can cause a dramatic change in statistics based on higher order moments such as covariances and correlations. This reasoning leads to the hypothesis that, if mode effects are detected in marginal distributions, multivariate statistics are expected to show larger effects.

Which hypothesis is the most likely, remains to be seen. A survey among 85 experts in the field of data collection methods and experts in the field of multivariate analysis revealed some support for the first hypothesis stating that multivariate mode effects are smaller. The experts were asked to indicate their a priori conviction on a line with endpoints -10 (hypothesis 1 is most likely) and +10 (hypothesis 2 is most likely); zero indicating that both hypotheses are seen as equally likely. The mean score is -1.6, and the median is -2; no difference could be detected between the answers of experts in data collection methods and experts in multivariate statistics. On average, the experts are slightly in favor of hypothesis 1. However, the standard deviation of 4.9 indicates that there are large differences in the expressed opinions. When the scores are trichotomized, 43 experts (51%) favor hypothesis 1, 17 experts (20%) think that both hypotheses are equally likely, and 25 experts (29%) favor hypothesis 2.

In this chapter I investigate the potential influence of data collection method on the parameter estimates of two substantive structural models: a model about experienced loneliness and a model about subjective well-being. Two different aspects of structural modeling are investigated: the loneliness model is a causal model of the determinants of loneliness, the subjective well-being model is a factor analysis (measurement) model of the structure of well-being. In section 7.2 a short description of these models is given, followed by an outline of the statistical search strategy. In section 7.3 the results are presented for the loneliness model and the well-being model. A summary of the main results is given in 7.4.

7.2. Method

Two different substantive structural models will be used to investigate the effect of data collection method on the estimated relationships: a model of loneliness and a model of well-being.

The loneliness model

The first model -a causal structural equation model about the determinants of loneliness- is derived from De Jong-Gierveld (1987). This model has four exogenous variables (living alone, extension of social network, self-evaluation, and age) and two endogenous variables (evaluation of social network and loneliness).

The exogenous variable living alone (X_1) indicates the degree in which people live together with important others. This variable is based on responses to questions about the living arrangements of the respondents. The scale values range from 1 (living together with more than one important other) to 3 (living completely alone). The extension of the social network (X_2) is measured by asking respondents to state the number of persons who are very important to them. This variable has a minimum value of 0. Self-evaluation (X_3) is measured using an eight-item scale. The minimum score is 0, the maximum score (very positive self-evaluation) is 8. Age (X_4) is measured in years.

The endogenous variable evaluation of social network (Y_1) is measured with a closed question about the degree of satisfaction with social relationships; the response categories range from 1 to 5: the value 1 indicates that the respondent is very dissatisfied, the value 5 means very satisfied. Loneliness (Y_2) is measured on an 11-item scale; the minimum score is 0, the maximum score (extreme loneliness) is 11.

In this model loneliness is negatively affected by the extension of the social network (number of important relationships), the amount of satisfaction with the social network, and a positive self-evaluation. Loneliness is positively influenced by living alone and age (see also Figure 7.1 on the next page). The loneliness model is a path model with observed variables only.

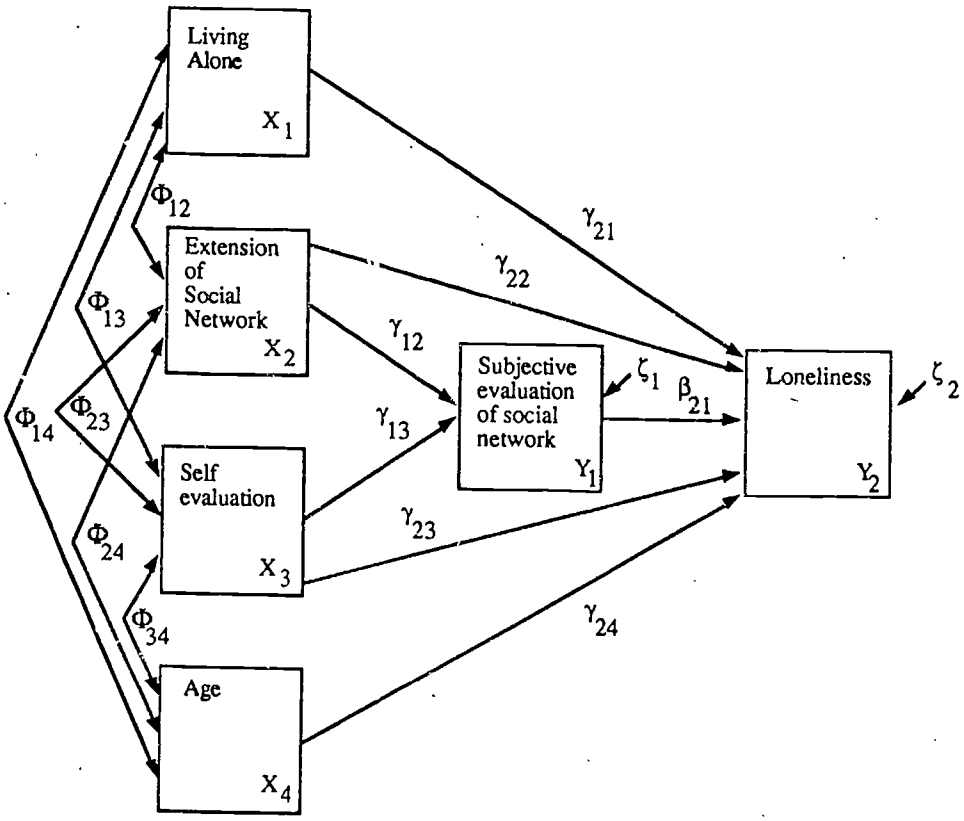


Figure 7.1. Loneliness Model

The following search strategy was used. First, I examined whether the covariance matrices differed for the three data collection methods. This was followed by a series of multi-group analyses to investigate whether the models have the same parameter values for the mail survey, the telephone survey, and the face to face survey (Bollen, 1989, chap. 8; Jöreskog & Sörbom, 1989, chap. 9). I started with the strictest model (model 1) in which each parameter, specified in the loneliness model, is assumed to be invariant over the three groups (i.e., the mail, the telephone, and the face to face survey). In this model the measurement error variances are fixed at zero.

The next model (model 2) includes information about the reliability of the measurement of the multiple item scales loneliness and self-evaluation. Preliminary analyses had indicated that the reliability of multiple item scales differed across data collection methods: the mail survey showed the most reliable results, while the telephone survey was the least satisfactory in this respect (cf. chapter 6). Therefore, in the next step I allowed for differences in variances of measurement errors between the groups. The reliability estimates under the congeneric test model are available for the two multiple item scales loneliness and self-evaluation. The variance of the measurement error epsilon for the variable loneliness and the variance of the measurement error delta for the variable self-evaluation is set according to the different reliabilities for these two variables in the three survey groups (Bollen, 1989, p. 168).

In the next step (model 3), invariance restrictions between groups were only imposed on parameter estimates for the two interview modes (face to face and telephone). The model for the self-administered mail survey group was only restricted to have the same pattern as the two interview groups; the loadings in the mail survey group were allowed to differ from the interview groups. Finally, for all three groups the only restrictions concerned the form (i.e., same dimensions and patterns); all parameter estimates were allowed to differ in the three groups (model 4).

To compare subsequent models the overall Chi-square and the overall root mean squared error were calculated. Furthermore, the normed incremental fit index Delta was calculated (Bentler & Bonett, 1980). Delta measures the proportionate reduction in the chi-square values when moving from a baseline model to the maintained model (Bollen, 1989, p. 270). As a baseline model the most restrictive model (model 1: all parameter estimates invariant in the three groups) is used. Furthermore, in most cases the subsequent models are nested within each other. For two nested models the difference in chi-squares is again chi-square distributed with degrees of

freedom equal to the difference in degrees of freedom for the two models. This makes it possible to test whether the improvement of fit is substantial.

The well-being model

The second model -a measurement model of the structure of well-being- is derived from Burt et al. (Burt, Wiley, Minor, & Murray, 1978; Burt, Fischer, & Christman, 1979). Four dimensions are distinguished: "general satisfaction," "satisfaction with specific domains," "positive affect" and "negative affect" (see also Figure 7.2 below).

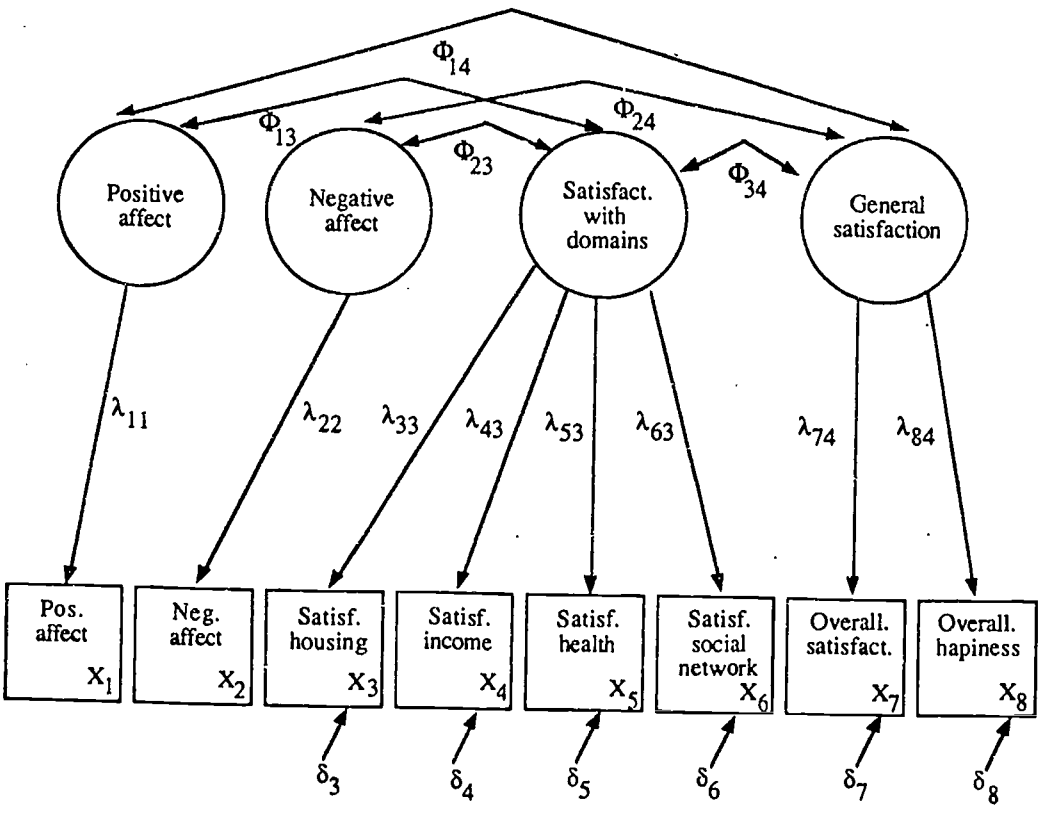


Figure 7.2. Well-being Model

The general satisfaction dimension is measured by two global variables: (X_8) overall happiness as indicated on a seven-step ladder (1: worst that could happen - 7: best) and (X_7) overall satisfaction with life in general as indicated on a single five-point scale (1: very dissatisfied - 5: very satisfied). The satisfaction with specific domains dimension is measured by four variables regarding satisfaction with certain domains of life (i.e., (X_3) housing, (X_4) income, (X_5) health, and (X_6) social network). Again, answers were given on a five-point scale, ranging from 1 (very dissatisfied) to 5 (very satisfied). The positive affect dimension is measured by a nine-item positive affect scale (X_1); the minimum score is 0, indicating the absence of any feelings of positive affect (happiness), the maximum score is 9 (extremely happy). The negative affect dimension is measured by a nine-item negative affect scale (X_2); the minimum score is 0, indicating the absence of any feelings of negative affect, the maximum score is 9. The positive and negative affect dimensions are assumed to be uncorrelated (cf. Bradburn, 1969; Hox, 1986).

The original well-being model, as published by Burt et al. (1978), is not identified. For a discussion of restrictions to make the well-being model identifiable, see Burt et al. (1979). In my version of the well-being model, the variance of the factors is fixed at 1.00. The measurement error variance of the two observed variables positive and negative affect is fixed at zero.

A related search strategy was used as in the loneliness example. First, I examined whether the covariance matrices differed for the three data collection methods. This was followed by a series of multi-group analyses to investigate whether the model has the same parameter values for the mail survey, the telephone survey, and the face to face survey (Bollen, 1989, chap. 8; Jöreskog & Sörbom, 1989, chap. 9). I started with the strictest model (model 1) in which each parameter, specified in the well-being model, is assumed to be invariant over the three groups (i.e., the mail, the telephone, and the face to face survey). The measures for positive and negative affect were treated as error free (i.e., error variance fixed at 0).

In the well-being model, multiple observed variables were available for the dimensions "general satisfaction" and "satisfaction with specific domains." This made it possible to allow the estimated variances of the measurement errors delta for these variables to differ across groups (model 2). Next, information about the reliability of measurement of the multiple item scales positive affect and negative affect is also included (model 3). Here I allowed differences in variances of measurement errors between the groups. The reliability estimates under the congeneric test model are available for positive affect and negative affect. The variance of the

measurement errors delta for these two variables is set according to the different reliabilities for the two scales in the three survey groups (Bollen, 1989, p. 168).

In the next step (model 4), invariance restrictions between groups were only imposed on parameter estimates for the two interview modes (face to face and telephone). The model for the self-administered mail survey group was restricted to have the same pattern as the two interview groups; but the loadings in the mail survey group were allowed to differ from the two interview survey groups. Subsequently, it was investigated if allowing for different measurement errors in the two interview modes improved the fit further (model 5 and model 6). Finally, for all three groups the only restrictions concerned the form (same dimensions and patterns); all parameter estimates were allowed to differ between the three groups (model 7).

The overall Chi-square, the overall root mean squared error, and the incremental fit index Delta were calculated. For nested models the difference in chi-squares was calculated to investigate whether the improvement of fit is substantial.

7.3. Results

The loneliness model

The loneliness model analyzed in this study is a causal (path) model with six observed variables. The four exogenous variables are living alone, extension of social network, self-evaluation, and age; the two endogenous variables are evaluation of social network and loneliness (see Figure 7.1 on page 100).

For each data collection method (mail, telephone and face to face survey) a covariance matrix was computed. The covariance matrices were significantly different in the three data collection groups (p=.00). Therefore, it is not surprising that the strictest model (model 1) did not fit. This model constrains all parameter estimates to be equal across the three groups.

In model 1 the measurement error variances were all fixed at zero. In the next model (model 2) estimates of the measurement error variance of the multiple item scales (loneliness and self-evaluation) were set in the error-variance matrices; for each data collection group different values were used based on the reliability estimates under the congeneric test model.

This did not improve the fit of the model, and the next models do not include these estimates of the measurement errors.

In the next step all parameters are constrained to be invariant for the face to face and the telephone interview group. In the mail survey group the parameter matrices are only constrained to have the same dimensions and patterns as in the two interview groups (model 3). This model has a reasonable fit (see Table 7.1). Since model 3 is nested in model 1 the difference in chi-squares can be used to test whether the increase in fit is statistically significant. Although the value of the incremental fit index is substantial (Delta=.39), the difference in chi-squares between model 1 and model 3 turns out to be not significant (p=.08).

In the final step (model 4), the restrictions are freed even further. In model 4 the only constraints are on the pattern of the parameter matrices. The same dimension and pattern are demanded, without restricting any of the non-fixed parameters to have the same value across groups. Model 4 shows a good fit. Compared to model 1 the fit is significantly better (p=.02). Also, compared to model 3 the fit of model 4 is better (p=.04). For an overview of the model fit see Table 7.1.

Table 7.1 Three Group Path Model Loneliness: Overall Fit

A three group model (Mail, FtF, Tel) was fitted with several restrictions. For each model the overall Chi-square, degrees of freedom (DF) and p-value and the overall root mean squared residual (RMSR) are presented. Delta gives the value of the normed incremental fit index (against model 1, the strictest model).

Model Restriction	CHI^2	DF	P-VALUE	RMSR	DELTA
(1) Mail=FtF=Tel	39.8	24	.03	1.12	--
(2) Mail=FtF=Tel/α	39.4	24	.02	1.06	.01
(3) Mail≈FtF=Tel	24.3	15	.06	1.10	.39
(4) Mail≈FtF≈Tel	6.4	6	.38	0.46	.84

Note. "=" indicates that the parameters in this model are invariant over groups; "≈" indicates the weaker same pattern restriction. "/α" that in this model the measurement error variance for the variables loneliness and self-evaluation is set according to their reliability.

In Table 7.2 the root mean squared residual and goodness of fit index are presented for each survey condition under all four models. Inspection of this

table suggests that model fit problems are most serious in the face to face condition.

Table 7.2 Three Group Path Model Loneliness: Group Fit

A three group model (Mail, FtF, Tel) was fitted with several restrictions. For each group in a model the goodness of fit index (GFI) and the root mean squared residual (RMSR) are presented.

	MAIL		FACE TO FACE		TELEPHONE	
Model Restriction	GFI	RMSR	GFI	RMSR	GFI	RMSR
(1) Mail=FtF=T	.98	0.28	.98	1.73	.99	0.84
(2) Mail=FtF=Tel/α	.98	0.24	.98	1.66	.99	0.75
(3) Mail≈FtF=Tel	1.00	0.38	.98	1.60	.99	0.96
(4) Mail≈FtF≈Tel	1.00	0.38	1.00	0.70	1.00	0.10

Note. "=" indicates that the parameters in this model are invariant over groups; "≈" indicates the weaker same pattern restriction. "/α" that in this model the measurement error variance for the variables loneliness and self-evaluation is set in accordance with their reliability.

When comparing over groups, the unstandardized parameter estimates are preferred (Bollen, 1989, p. 126). For the least restrictive model (model 4) the unstandardized parameter estimates are given in Table 7.3.

To interpret the relative importance of the parameter estimates correctly, it is essential to keep in mind the scale on which the variables are measured. For loneliness the minimum score is 0 and the maximum score is 11; the self-evaluation score ranges from 0 to 8. The variable living alone ranges from 1 to 3. Extension of the social network is a count of the number of important relations with a minimum of 0. Age is measured in years. Satisfaction with social network is measured on a single five-point scale.

The following (conservative) decision rule was adopted: a difference in parameter estimates between modes is seen as substantial if that difference is larger than twice the largest standard error for that specific parameter. Inspection of Table 7.3 shows that the major differences between data collection methods occur for the parameters $Beta_{21}$ (effect of subjective evaluation of social network on loneliness), $Gamma_{12}$ (effect of extension of social network on the subjective evaluation of social network), $Gamma_{23}$

(effect of self-evaluation on loneliness), and $Gamma_{24}$ (effect of age on loneliness).

Table 7.3 Three Group Same Pattern Model (Mail=FtF=Tel) Loneliness: Parameter Estimates

Unstandardized ML estimates for the mail, face to face, and telephone condition. Standard errors are given in parentheses. The squared multiple correlations for the endogenous variables evaluation of social network $[R_{y1}]^2$ and loneliness $[R_{y2}]^2$ are presented for each group.

Parameter	MAIL	FACE TO FACE	TELEPHONE
$Beta_{21}$	-2.11 (0.17)	-1.29 (0.16)	-1.37 (0.19)
$Gamma_{21}$	0.55 (0.33)	0.51 (0.30)	0.76 (0.30)
$Gamma_{22}$	-0.29 (0.10)	-0.30 (0.11)	-0.23 (0.12)
$Gamma_{12}$	0.08 (0.04)	0.15 (0.04)	0.05 (0.04)
$Gamma_{13}$	0.09 (0.03)	0.10 (0.03)	0.05 (0.03)
$Gamma_{23}$	-0.18 (0.07)	-0.28 (0.07)	-0.37 (0.07)
$Gamma_{24}$	0.00 (0.01)	0.03 (0.01)	-0.00 (0.01)
Psi_{11}	0.75 (0.07)	0.83 (0.08)	0.62 (0.06)
Psi_{22}	4.58 (0.44)	4.58 (0.43)	5.33 (0.48)
$[R_{y1}]^2$ .08	.11	.02	
$[R_{y2}]^2$ .52	.41	.29	

These differences can have a major influence on the interpretation of social science results. An illustration is given in Figure 7.3 on the next page. This figure contains the graphical representation and the parameter estimates for model 4 (same pattern for each data collection method). Parameter estimates are often standardized when interpreting results. Figure 7.3 presents the same parameter estimates as Table 7.3, but now standardized to a common metric for the three groups. This preserves across groups comparability (Jöreskog & Sörbom, 1989, p. 238).

It should be noted that the respondents in the three data collection modes differed slightly on two important background variables: gender and marital status. In the mail condition slightly more men and married persons were present, while in the face to face condition slightly more respondents were women and slightly more respondents were divorced (see chapter 4, section 4.8).

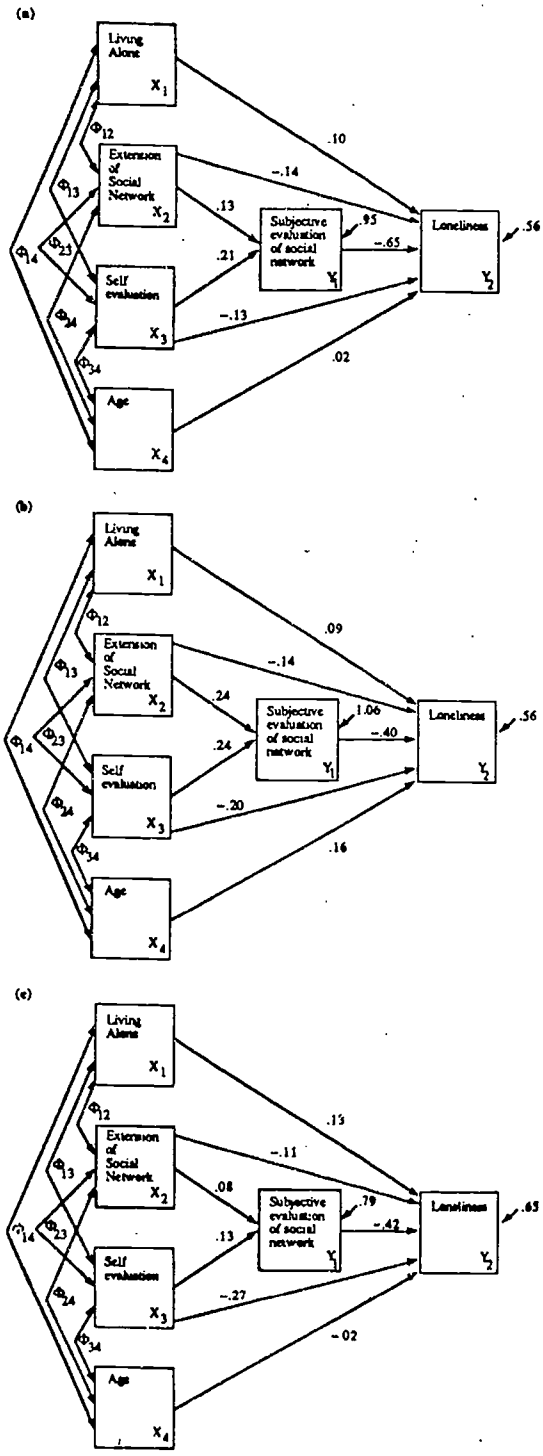


Figure 7.2 Standardized parameter estimates for loneliness model (model 4) for (a) Mail Survey, (b) Face to Face Interview, and (c) Telephone Interview

To investigate the potential confounding influence of these differences between the groups, I repeated all analyses using weighted covariance matrices. These weighted covariance matrices were adjusted for the differences in gender and marital status between the three groups. The reanalyses did not result in different conclusions.

In sum: the least restrictive statistical model had a good fit. This model assumes the same dimension and pattern across groups without restricting any of the non-fixed parameters. The differences appear large enough to influence the substantive interpretation of the results, and give cause for some concern about the robustness against data collection method of substantive interpretations of empirical models.

The well-being model

The well-being model analyzed, is a confirmative factor analysis model with four dimensions (positive affect, negative affect, domain satisfaction, and general satisfaction) measured by eight observed variables. See Figure 7.2 on page 102. The variance of the factors is fixed at 1.00, and the measurement error variances of the two observed variables positive affect and negative affect are fixed.

I started with the computation of a separate covariance matrix for each data collection method (mail, telephone and face to face survey). The covariance matrices were significantly different in the three groups (p=.00). Given this result, it is not surprising that the strictest model (model 1), which constrains all parameter estimates to be equal across the three data collection groups, did not fit. In model 1 the measurement error variances for the two observed variables positive affect and negative affect were fixed at zero, all other measurement error variances were constrained to be equal across the three groups. In the next model (model 2), the measurement error variances of the observed variables for the factors "domain satisfaction" and "general satisfaction" were estimated separately in the three groups. Remember, that more than one observed variable was available for each dimension. This results in a model that fits much better than the first model (p=.00), although the overall fit is still not good (see also Table 7.4).

Table 7.4 Three Group Factor Model Well-being: Overall Fit

A three group model (Mail, FtF, Tel) was fitted with several restrictions. For each model the overall Chi-square, degrees of freedom (DF) and p-value and the overall root mean squared residual (RMSR) are presented. Delta gives the value of the normed incremental fit index (against the strictest model 1).

Model Restriction	CHI²	DF	P-VALUE	RMSR	DELTA
(1) Mail=FtF=Tel	220.1	89	.00	0.21	--
(2) Mail=FtF=Tel/δ	149.1	77	.00	0.21	.32
(3) Mail=FtF=Tel/δ+α	148.6	77	.00	0.21	.32
(4) Mail≈FtF=Tel	131.1	70	.00	0.14	.40
(5) Mail≈FtF=Tel/δ	117.6	64	.00	0.14	.47
(6) Mail≈FtF=Tel/δ+α	117.2	64	.00	0.13	.47
(7) Mail≈FtF≈Tel	93.0	51	.00	0.10	.58

Note. "=" indicates that the parameters in this model are invariant over groups; "≈" indicates the weaker same pattern restriction. "/δ" indicates that in this model measurement error variances are estimated separately in the three groups. "δ+α" indicates that in addition the measurement error variance for the variables positive and negative affect is set according to their reliability.

The next model (model 3) sets the error variances for the two remaining observed variables (positive affect and negative affect) according to the reliability estimates under the congeneric test model. This results in a slightly better fit. In the subsequent model (model 4) all parameters are constrained to be invariant for the face to face and the telephone interview group only. In the mail survey group the parameter matrices are only constrained to have the same dimensions and patterns as in the two interview groups. This model fits better than model 2 and 3, which constrain the factor loadings and correlations, but allow the measurement errors to differ across all groups (see Table 7.4).

In the next two steps, I again allowed differences in measurement errors. In model 5 I allowed differences in the variances of the measurement errors delta of the observed variables for domain satisfaction and general satisfaction. This resulted in a slightly better fit than model 4 (p=.04). Model 6 also estimates the fixed error variances of observed positive and negative affect using reliability estimates. This again results in a slightly better fit than model 4 (p=.03). Furthermore, model 6 can be compared statistically with model 3, which allows for different

measurement errors across groups, but constrains all other parameter estimates to be equal. Model 6 fits significantly better than model 3 (p=.00).

In the final step (model 7), the restrictions are freed even further. In model 7 the only constraints are on the pattern of the parameter matrices. The same dimension and pattern are assumed, without restricting any of the nonfixed parameters to have the same value across groups. Compared to model 2 (identical loadings and correlations, different measurement errors) the fit is significantly better (p=.00). Also, compared to model 4 (restrictions across face-to-face and telephone conditions) the fit of model 7 is better (p=.00). Compared to model 5 (restrictions across face-to-face and telephone conditions, different measurement errors) the fit of model 7 is also better (p=.03), but the overall fit of model 7 is still not quite satisfactory. However, the value of the root mean squared residuals (.10) and the relative size of the chi-square and the degrees of freedom ($chi^2/df=1.82$) suggest that this model is acceptable.

For an overview of the fit statistics of the models see Table 7.4. In addition, the root mean squared residual and goodness of fit index for each survey condition under all four models are presented in Table 7.5.

Table 7.5 Three Group Factor Model Well-being: Group Fit

A three group model (Mail, FtF, Tel) was fitted with several restrictions. For each group in a model the goodness of fit index (GFI) and the root mean squared residual (RMSR) are presented.

	MAIL		FACE TO FACE		TELEPHONE	
Model Restriction	GFI	RMSR	GFI	RMSR	GFI	RMSR
(1) Mail=FtF=Tel	.92	0.26	.94	0.15	.93	0.21
(2) Mail=FtF=Tel/δ	.95	0.26	.95	0.15	.95	0.20
(3) Mail=FtF=Tel/δ+α	.95	0.26	.95	0.14	.95	0.20
(4) Mail≈FtF=Tel	.97	0.13	.94	0.16	.95	0.14
(5) Mail≈FtF=Tel/δ	.97	0.13	.95	0.16	.96	0.13
(6) Mail≈FtF=Tel/δ+α	.97	0.13	.95	0.15	.96	0.12
(7) Mail≈FtF≈Tel	.97	0.13	.96	0.09	.97	0.07

Note. "=" indicates that the parameters in this model are invariant over groups; "≈" indicates the weaker same pattern restriction. "/δ" indicates that in this model measurement error variances are estimated separately in the three groups. "δ+α" indicates that in addition the measurement error variance for the variables positive and negative affect is set accordancing to their reliability.

When comparing over groups, unstandardized parameter estimates are preferred (Bollen, 1989, p. 126). For the least restrictive model (model 7) the unstandardized parameter estimates are given in Table 7.6. To interpret the relative importance of the parameter estimates, it is important to know the scale on which the variables are measured. Positive and negative affect are measured by two 9-item scales, with a range from 0 (lowest score) to 9 (highest score). The domain satisfactions and global satisfaction variables are measured by single five-point questions. Global happiness is measured on a single seven-point scale.

Table 7.6 Three Group Same Pattern Model (Mail=FtF=Tel) Well-being: Parameter Estimates

Unstandardized ML estimates for the mail, face to face, and telephone condition. Standard errors are given in parentheses.

Parameter	MAIL	FACE TO FACE	TELEPHONE
$Lambda_{11}$	2.29 (0.11)	2.01 (0.10)	1.81 (0.09)
$Lambda_{22}$	2.14 (0.10)	2.25 (0.11)	2.07 (0.10)
$Lambda_{33}$	0.33 (0.07)	0.23 (0.07)	0.09 (0.07)
$Lambda_{43}$	0.42 (0.07)	0.28 (0.08)	0.34 (0.09)
$Lambda_{53}$	0.27 (0.06)	0.27 (0.08)	0.25 (0.08)
$Lambda_{63}$	0.41 (0.06)	0.65 (0.10)	0.21 (0.07)
$Lambda_{74}$	0.60 (0.04)	0.54 (0.06)	0.47 (0.05)
$Lambda_{84}$	1.01 (0.07)	0.83 (0.10)	0.91 (0.11)
Phi_{13}	0.56 (0.09)	0.39 (0.09)	0.35 (0.15)
Phi_{23} -0.62 (0.09)	-0.41 (0.09)	-0.40 (0.15)	
Phi_{14}	0.45 (0.05)	0.39 (0.07)	0.42 (0.07)
Phi_{24}	-0.46 (0.05)	-0.52 (0.07)	-0.40 (0.08)
Phi_{34}	1.13 (0.09)	0.68 (0.11)	1.21 (0.25)
Theta-$delta_3$	0.92 (0.09)	0.69 (0.07)	0.95 (0.09)
Theta-$delta_4$	0.88 (0.09)	0.92 (0.09)	0.83 (0.09)
Theta-$delta_5$	0.69 (0.06)	0.78 (0.08)	0.91 (0.09)
Theta-$delta_6$	0.64 (0.06)	0.54 (0.11)	0.54 (0.05)
Theta-$delta_7$	0.12 (0.02)	0.23 (0.05)	0.28 (0.04)
Theta-$delta_8$	0.53 (0.08)	1.23 (0.15)	1.22 (0.16)

Relatively large differences between the groups are found for the loadings of the observed variables housing-satisfaction and social network-satisfaction ($lambda_{33}$ and $lambda_{63}$) on the domain satisfaction factor. Smaller, but still substantial differences (twice the largest standard error) are found for the loadings of the positive affect scale on the positive affect factor ($lambda_{11}$), and for the variable overall satisfaction on the general satisfaction factor ($lambda_{74}$). Furthermore, it should be noted that the correlations of the satisfaction with domains factor (factor 3) with the other factors show some differences over the groups (phi_{13}, phi_{23}, phi_{34}). The latter even shows two values outside the permitted range, which again indicates that there are problems with the overall model.

In the well-being model, the variances of the factors have been fixed at 1.00. To facilitate the interpretation of the factor loadings, the observed variables' parameters are often standardized too. Figure 7.4 on the next page contains the graphical representation of model 7, and presents the same factor loadings as Table 7.6 The difference is that now the observed variables are standardized to a common metric for the three groups. This standardization is based on the pooled variance estimates for the observed variables under the fitted model, and preserves the comparability across groups (cf. Jöreskog & Sörbom, 1989, p. 238).

Again, all analyses were repeated employing weighted covariance matrices to adjust for the differences in gender and marital status between the three groups. Once more, the reanalyses did not result in different conclusions.

In sum: the least restrictive statistical model was more appropriate. This model assumes the same dimension and pattern across groups without restricting any of the non-fixed parameters. The relative importance of some estimated parameters varied considerably across data collection modes. This gives cause for concern, because the differences appear large enough to influence the substantive interpretation of the results, and may lead to different substantive interpretations under different data collection modes.

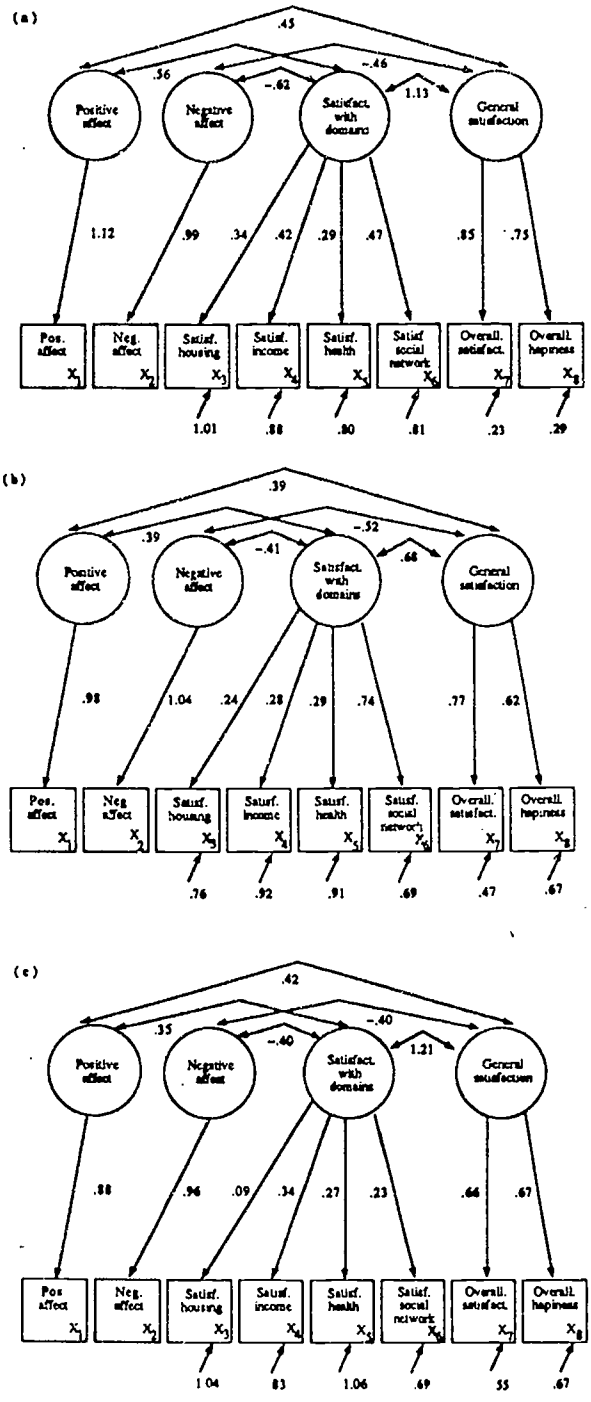


Figure 7.4 Standardized parameter estimates well - being model (model 7) for (a) Mail Survey, (b) Face to Face Interview, and (c) Telephone Interview

7.4. Summary

To investigate the potential influence of data collection method on the estimates of relationships between variables I compared two substantive structural-equation models across different data collection methods: a loneliness model and a well-being model. The loneliness model analyzed in this study is a causal model with four exogenous variables (living alone, extension of social network, self-evaluation, and age) and two endogenous variables (evaluation of social network and loneliness). The loneliness model is a path model with observed variables. The well-being model analyzed is a confirmatory factor analysis model with four factors (positive affect, negative affect, domain satisfaction, and general satisfaction) measured by eight observed variables.

Two rival hypotheses were investigated. The first hypothesis states that, although small mode effects are in general found on marginal distributions of variables, the multivariate estimates will remain stable (form resistant correlation hypothesis). The second hypothesis states that if (small) mode effects are found in marginal distributions, multivariate statistics will show even larger effects (instability of higher order moments hypothesis).

A small survey among experts in the field of data collection and experts in the field of multivariate analysis disclosed that a slight majority (51%) favored hypothesis 1, 20% thought that both hypotheses were equally likely, and 29% favored hypothesis 2. The results of a Lisrel multi-group analysis lend support to the second hypothesis.

For both the loneliness-model and the well-being model the strictest statistical model was rejected; this model assumes invariance of all parameters over the three groups (i.e., the mail, the telephone, and the face to face survey). A less strict model was more appropriate. This model assumes the same dimension and pattern across groups without restricting any of the non-fixed parameters. Comparison of the estimates under this model for the two substantive models gives cause for some concern.

For the loneliness model, the least restrictive (same pattern) model had a good statistical fit. The loneliness model is a path-model in which the score on a loneliness scale is the major dependent variable. In both the mail survey and the face to face interview group the proportion variance explained was relatively high (.52 and .41), in the telephone condition this figure was only 0.29 (cf. Table 7.3). The same variables explain far less variance in the telephone survey condition. Also, the relative importance of the individual predictors varies considerably across data collection method

(cf. Figure 7.3 on page 108). In the mail survey condition the influence of subjective evaluation of the social network on feelings of loneliness is considerably larger than in either the face to face or the telephone condition (the standardized parameter estimates are mail: -.65, face to face: -.40, telephone: -.42). However, in all three groups evaluation of social network is the most important determinant of feelings of loneliness. A striking difference is found when the variable age is considered. Only in the face to face condition age is a relatively important determinant of feelings of loneliness.

The well-being model (a factor model with four dimensions or factors) showed a less satisfactory overall statistical fit for the least restrictive (same pattern) model specification (p=.00). However, the value of the root mean squared residuals (.10) and the relative size of the chi-square and the degrees of freedom ($chi^2/df=1.82$) suggest that this model is acceptable.

The standardized parameter estimates under this model reveal a marked difference in the relative importance of the variables. In the mail survey condition the observed variable (satisfaction with) social network is the most important variable for the domain satisfaction dimension (lambda=.47), immediately followed by income. Housing and health are less important. In the face to face interview condition the most important variable is social network (lambda=.73); the variables health, income and housing hardly differ in relative importance. In the telephone condition income is the most important variable for the domain satisfaction dimension (.34), while social network is the third important variable (.24). See also figure 7.4 on page 114, which contains the parameter estimates standardized to a common metric for the three groups.

As mentioned above the statistical fit for even the least restrictive (same pattern) model was not quite satisfactory. Exploratory analyses in which restrictions between groups were freed based on the modification indices resulted in a fitting model. In this model the structure of well-being diverges even more across groups, because several factor loadings in the parameter matrix lambda had to be freed. This model specifies a different pattern of additional factor loadings for each of the three data collection methods (De Leeuw & Hox, forthcoming).

In sum: a clear influence of data collection method on estimated relationships between variables has been detected. The same pattern and the same dimension were discovered under each data collection method, but the relative importance of some estimated variables varied considerably across modes.

CHAPTER 8

CONCLUSION

. . . and go on until you come to the end: then stop.
Lewis Carroll, The annotated Alice, 1976, p. 158

8.1. The Major Results

Prior to the 1970's, the face to face interview was the dominant and accepted method for conducting surveys. Since then there has been a dramatic change in data collection techniques. Mail and especially telephone surveys have become increasingly popular in the last decade. Also, mixed mode surveys (e.g., surveys that combine the use of more than one data collection method to gather data for a single survey project) are occurring more and more. These changes give rise to questions such as: Is one mode as good as the other? May we combine data that are collected by different modes? How valid are these modes?

One of the most important questions for both survey researchers *and* for consumers of survey research is whether the data obtained by one survey mode differ from the data obtained by another. This question forms the central problem in this study. To provide an answer, I compared three major modes of survey research, that is, face to face interviews, telephone interviews, and mail questionnaires. I started with a comprehensive literature review based on a meta-analysis of experimental comparisons of these data collection methods. The meta-analysis was followed up by a controlled field experiment, in which a face to face interview, a telephone interview, and a mail survey were compared. Three different types of possible mode effects were investigated. First, I analyzed univariate mode effects. Next, I compared how items scale in different modes (psychometric mode effects), and finally I compared the behavior of Lisrel models (multivariate mode effects).

The meta-analysis detected small differences in data quality, suggesting a dichotomy of survey modes: modes with and modes without an interviewer. None of the modes was superior on all criteria (response validity, item nonresponse, number of statements made in response to an

open question, social desirability, and similarity of response distributions across modes). The modes with an interviewer resulted in higher response rates and lower item nonresponse, but also produced more socially desirable answers (cf. chapter 3).

The field experiment showed a significant difference in response rates between the methods (cf. chapter 4). The face to face survey resulted in the lowest response rate, which is contrary to the results of the meta-analysis. However, recent surveys in the Netherlands corroborate this unexpected finding: at the Netherlands Central Bureau of Statistics the response to telephone surveys tend to be higher than the response to face to face surveys (De Heer, Akkerboom & Israëls, 1990; Snijkers, 1992).

The univariate analyses replicated the main conclusions of the meta-analysis. The mail survey resulted in more item nonresponse, but also in more self-disclosure on sensitive topics. No consistent differences between face to face and telephone interviews were discovered on these points. Additional analyses detected no differences in acquiescence between the modes, but a small recency effect was found. In the telephone condition respondents more often chose an extreme positive answer (cf. chapter 5).

The psychometric mode comparisons involved both reliability and scalability. Again, small differences were found: the mail survey performed slightly better when reliability and item scalability were investigated. Psychometric analysis of the individual response patterns on multiple item scales revealed slightly more respondents with unexpected or aberrant response patterns in the two interview conditions (cf. chapter 6).

The empirical comparisons until this point supported Groves' conclusion that the most consistent finding in studies comparing face to face and telephone interviews is the lack of differences (Groves, 1989, p. 551). The main differences found were between the mail survey on the one hand and the two interview surveys on the other hand. It was somewhat harder to have people answer questions in the mail survey as the higher item missing data rates indicate, but when questions were answered, the resulting data seem to be of better quality (more self-disclosure, more reliable and consistent responses). However, the differences are relatively minor and survey researchers might feel justified in ignoring them.

The pleasant picture painted above is shaken by the results of the covariance structure analyses. Two substantive models (a path model and a factor analysis model) were compared over modes. The results give some ground for optimism: the same pattern and the same dimensionality were confirmed under each data collection method. On this point all three modes led to the same structure. There is also a reason to be pessimistic: the

relative importance of some estimated parameter values varied considerably across data collection methods. This could lead to different conclusions concerning the importance and strength of the influence of one variable on another, when different data collection methods are used. However, the conclusion that there is some influence of that specific variable on a second specific variable will still be drawn under each of the data collection modes (cf. chapter 7).

8.2. Some Critical Comments

Comparisons between data collection methods are of course only possible on that common middle ground on which these modes are comparable. A telephone interview of the deaf would not really be a good idea, and a certain level of literacy is necessary to understand a self-administered questionnaire. But, the shared, common ground on which mode comparisons can be made is much larger than many realize: For instance, in this mode comparison checklists and open questions were used as well as closed questions, and a variety of response categories were employed. A total of 82 questions was asked; including standard biographical information, but also potentially sensitive questions. The average interview time (i.e., time from first to last question, excluding introduction and conclusion of the interview) was 31 minutes for the face to face interview and 24 minutes for the telephone interview.

The approach chosen was a controlled field study in which I tried to optimize the internal validity of the experiment without jeopardizing the external validity: error variance was controlled as far as possible, but the implementation of the survey procedures remained realistic in terms of general survey practice. Many different aspects of survey measurement error were studied, and a variety of statistical techniques were employed on global indicators of data quality. A completely different approach is the laboratory experiment in which successive series of tightly controlled small experiments are conducted, focusing on one specific (mode) effect at the time (cf. Schwarz, Strack, Hippler & Bishop, 1991; Hippler & Schwarz, 1992). Also, in my approach I focused on the end product of the survey process. The question-answer process itself (cf. Cannell, Miller, & Oksenberg, 1981; Dijkstra & Van der Zouwen, 1977; Strack & Martin, 1987) was not studied, and no attempts were made to study the potential influence of respondent-interviewer interaction (cf. Schaeffer, 1991; Van der

Zouwen, Dijkstra & Smit, 1991) or the thought processes that respondents use to interpret and answer survey questions (cf. Forsyth & Lessler, 1991).

The topic of mode effects and measurement error is complex, and different approaches have been used in studying it. At the current stage of the scientific inquiry a diversity of approaches is a positive contribution to the progress of science, adding beautifully colored stones to the interdisciplinary mosaic of our knowledge (cf. Cronbach, 1957; Kruskall, 1991). Each approach uses different but valid methods; each approach answers questions that the other does not. Sometimes a question answered in one approach gives rise to new questions, which can be answered only by switching to another research strategy. The approach I followed in this study is optimal for discovering which differences between modes actually exist. To find out which processes explain these differences, other approaches such as laboratory experiments or cognitive interviews are needed. For instance, one of the most striking findings in my study was the apparent dichotomy between self-administered questionnaires and interview strategies (both telephone and face to face). To answer the very simple "why?," a successive series of detailed and highly controlled experiments should be conducted focusing on differences in the offered stimuli and the subsequent responses.

Finally, it should be noted that the results discussed here are based on studies in the USA and Western Europe, and are not necessarily valid in other countries and cultures.

8.3. Computer Aided Data Collection Methods

At the moment a technological change is going on in the field of data collection. Computers have been used for data analysis for several decades, and microcomputers have become standard tools for word processing. Computers have recently become popular as data collection tools too. Computer assisted telephone interviewing (CATI) has been developed in the USA in the seventies and is now widely used. In the Netherlands CATI-systems are used at the Netherlands Central Bureau of Statistics, at the major marketing research institutes, and at some universities. Also the traditional face to face interview is gradually being replaced by computer assisted personal interviewing (CAPI). Even computer aided procedures for self-administered questionnaires (CASAQ) have been developed. For an overview, see Hox, De Bie, and De Leeuw (1990) and Saris (1991). Direct comparisons of computer aided data collection methods (CADAC) are very

rare, most of the literature concerns comparisons between a paper and pencil and a computer assisted form of the *same* data collection mode (cf. Snijkers, 1992). In the next paragraphs I will extrapolate my main conclusions to the computer aided forms of data collection methods.

For respondents in a telephone interview nothing changes when a research institute switches from paper and pencil telephone surveys to CATI. For the interviewers the task becomes less complex, because administrative duties have been taken over by the computer. As a result, the differences, if any, point toward a slight advantage for CATI, for instance fewer routing errors (cf. Nicholls & Groves, 1986; Groves & Nicholls, 1986). Contrary to what might be expected, CATI does not lead to a faster interviewing pace (Hox, 1992). In CAPI the computer is visible to the respondent, who might react to its presence. However, very few adverse reactions and no reduction in response rates have been reported (Van Bastelaar, Kerssemakers & Sikkel, 1987; Sikkel, 1988; Martin & O'Muircheartaigh, 1991). No evidence of differences in responses could be detected.

It seems safe to assume that the main findings concerning mode differences between telephone and face to face surveys are also valid for the computer aided versions of these survey techniques. This means that with well-trained interviewers and the *same* well-constructed structured questionnaire, both CAPI and CATI will perform well and differences in data quality will be extremely small. Of course, it should be noted that CAPI has a greater potential than CATI, just as paper and pencil face to face interviews have a greater potential than paper and pencil telephone interviews (cf. chapter 1). Unfortunately these potentials have hardly been challenged.

There are several forms of computer aided self-administered questionnaires. Existing computer networks or bulletin boards can be used to distribute a questionnaire, or diskettes whith a self-contained questionnaire program can be sent to respondents, who then answer the questions on a personal computer (e.g., business surveys, school surveys). A special form of CASAQ is computer assisted panel research (CAPAR). This is a panel survey where a small home computer and a modem are placed in the respondents home (Saris, 1989). Finally, during a CAPI-session an interviewer can hand over the computer to the respondent, who can then answer some questions in privacy. This is equivalent to handing over a questionnaire to a respondent during a paper and pencil face to face interview.

All these variations have in common that the question is read from a screen and the answer is entered into the computer by the *respondent*. Just as in paper and pencil self-administered questionnaires the respondents answer the questions in a private setting, which reduces a tendency to present themselves in a favorable light. There is some evidence (Waterton, 1984) that CASAQ produces less socially desirable answers than CAPI, when sensitive questions are asked. Furthermore, in a CASAQ-session the respondent and not the interviewer paces the questions. However, the respondent is not the only locus of control (cf. chapter 2). The computer program controls the order of the questions, either by presenting one question at the time or by presenting a screen with several questions. The respondent is, in general, not allowed to go back and forth unlimited as can be done in a paper and pencil questionnaire. In this sense a CASAQ-session resembles more an interview-session than a self-administered questionnaire.

When I extrapolate the main findings concerning mode differences between interview surveys and mail surveys, I have to consider the similarities and dissimilarities between CASAQ and self-administered mail surveys discussed above. When sensitive questions are used CASAQ should provide more "valid" and less socially desirable answers than either CATI or CAPI. In a CASAQ-session the respondent has more opportunities to control the pace of the interview than in a CATI- or CAPI-session, but the opportunity to deliberately relate different questions is almost the same. I therefore, expect that on psychometric data quality criteria the differences will be smaller for the computer-aided versions than for the paper and pencil versions. One of the first empirical comparisons between a computer assisted telephone interview and a computer assisted self-administered questionnaire is now in progress at the University of Amsterdam (cf. Kalfs & Saris, 1991).

8.4. Future Directions in Survey Research

In 1956 the British "Astronomer Royal" predicted that space travel would be technologically impossible for a long time. A year later the first Sputnik was successfully launched, and in 1968 the first man walked on the moon. Predicting the future is hazardous. Still, there are some clearly discernible trends in survey methodology that need mentioning.

The telephone interview is emerging as the heir apparent to the face to face interview, at least for large surveys with structured questionnaires (cf.

Dillman, 1992). The expensive face to face interview will be saved for those special cases that really need the flexibility and high potential of this method. Telephone surveys are less costly than face to face surveys, and differences in data quality between *well-conducted* telephone and face to face surveys are small. Although the differences are small, it seems wise to run two parallel surveys before switching methods in long running (annual) surveys. This procedure makes it possible to calibrate the new method.

Mail surveys will remain popular. Compared to face to face and telephone surveys, mail surveys are the least expensive and perform better when sensitive questions are asked. The recent developments and progress in word processing and desk top publishing bring new possibilities to mail surveys (cf. Tufte, 1991). Highly individualized mail surveys, a sophisticated lay-out, and intricate graphical question formats are now within reach of every survey research institute.

Mail and telephone surveys are here to stay, in its pure form or as part of a mixed mode survey design. Mixed mode surveys take place with an increasing frequency, and are used for major governmental surveys in the U.S. and Europe (Dillman & Tarnai, 1988). Mixed mode surveys involve combining data from several sources into a single data set. This is done on the assumption that these data are exchangeable. In the past, only small response differences have been found between methods. More worrisome is the influence of data collection method on covariance structure models reported in chapter 7. One rather conservative solution would be not to mix methods at all, when statistical modeling is aimed at. However, mixed mode surveys have many positive points (cf. Dillman & Tarnai, 1988). A far more constructive solution is to include mode of data collection as an explanatory variable in statistical modeling, and only collapse data over modes if the preliminary analyses do not reveal a significant mode influence.

Computer aided data collection (CADAC) will become more important in the near future. CADAC can reduce measurement error by utilizing automatic question skips and range and edit checks. But CADAC has far greater potentials. For instance, the internal computer clock can be used to record interview length or to measure latency time between questions and answers (cf. Bassili & Fletcher, 1991). Randomization of questions and answers can be used to avoid order effects. Complex questions can be asked and continuous response scales can be used in standard interviews (e.g., repertory grids, vignettes, magnitude estimation). Using a computer to interact with the respondent makes answering this kind of questions a natural process (cf. Saris, 1988). "Tailored" versions of a questionnaire may be offered to different respondents, in which the question sequences change

on the basis of the respondent's answer to previous questions. In the past researchers too often employed computer assisted versions of standard paper and pencil questionnaires. But CADAC can be used in a far more creative way. The available tools do affect the type of questions we can ask, and CADAC is offering a large and sophisticated toolkit!

Interviewer training should be adapted to the changes in data collection methods discussed above. Telephone interviewers should be explicitly trained in the use of explicit verbal and paralinguistic cues to overcome the absence of nonverbal communication in telephone interviews (cf. section 2.3). When CAPI or CATI is used interviewers should be trained in simple computer skills. More important however is that interviewers are trained in maintaining a high quality interaction with the respondents, even with a computer standing between them.

Finally, there are reasons to be optimistic about the future. Differences in data quality between data collection methods are mostly small, and new tools are available to collect the data. When these tools are used *intelligently*, measurement errors could be reduced even further. There is also some reason for concern: response rates in interview surveys have been falling for most countries (cf. De Heer & Israëls, 1990). At the same time response rates for mail surveys have reached acceptable heights (cf. Goyder, 1987). These rising response rates are the result of considerable research on response enhancing factors in mail surveys (cf. Dillman, 1978; Heberlein & Baumgartner, 1978). Therefore, in my view more research on response inducement in interview surveys would be a wise investment.

SAMENVATTING

Een Methodologische Vergelijking van de Datakwaliteit bij Face to Face, Telefonische en Schriftelijke Ondervraging

In dit proefschrift worden drie belangrijke dataverzamelingsmethoden voor sociaal-wetenschappelijk survey onderzoek, te weten de postenquête, het telefonische interview en het 'face-to-face' interview met elkaar vergeleken. Centraal in dit onderzoek staat de vraag of, en zo ja, in hoeverre de gegevens verkregen via deze drie dataverzamelingsmethoden van elkaar verschillen.

In het eerste hoofdstuk wordt een korte omschrijving gegeven van deze drie methoden voor dataverzameling en worden de voor- en nadelen van elke methode op een rijtje gezet.

Hoofdstuk 2 geeft een overzicht van verschillende theoretische overwegingen omtrent het ontstaan van mogelijke methodeverschillen.

In hoofdstuk 3 wordt de bestaande empirische onderzoeksliteratuur samengevat. De hierbij gebruikte methode is die van de meta-analyse. Op grond van deze meta-analyse kan geconcludeerd worden dat bij goed uitgevoerde surveys met gestructureerde vragenlijsten er slechts kleine verschillen in datakwaliteit zijn tussen de gebruikte survey methoden. Geen van de drie methoden was de beste op alle vergelijkingspunten (responsvaliditeit, sociale wenselijkheid, item nonrespons, aantal verschillende antwoorden op een open vraag, en overeenkomst tussen de methoden in antwoordverdelingen bij een meerkeuze vraag). De gevonden verschillen in data kwaliteit wijzen op een tweedeling in dataverzamelingsmethoden met en dataverzamelingsmethoden zonder interviewers.

Vervolgens is een grootschalig veldexperiment uitgevoerd, waarin een face-to-face interview, een telefonisch interview en een postenquête met elkaar werden vergeleken. Drie verschillende soorten methodeneffecten werden onderzocht: univariate effecten (hoofdstuk 5), psychometrische effecten (hoofdstuk 6), en multivariate effecten (hoofdstuk 7).

In hoofdstuk 4 wordt de opzet van het veldexperiment gegeven. Dit omvat een beschrijving van de instrumentatiefase waarin voor iedere dataverzamelingsmethode een equivalente versie van de vragenlijst geconstrueerd werd, een beschrijving van de gevolgde procedures bij het steekproeftrekken en bij de selectie en training van de interviewers, en een beschrijving van de wijze waarop de dataverzamelingsmethoden geïmplementeerd werden. Mogelijke bedreigingen van de interne en van de externe validiteit werden zorgvuldig tegen elkaar afgewogen. De experimentele procedures werden in een pilotonderzoek uitgetest en daarna toegepast in het hoofdonderzoek.

Hoofdstuk 4 besluit met een overzicht van de respons in het hoofdonderzoek. Deze verschilde significant per methode. Het face-to-face interview leverde de laagste respons (51%). De postenquête resulteerde in een respons van 66% en het telefonische interview eveneens in een respons van 66%. Dit komt overeen met recente bevindingen van het CBS. Nadere analyse van de nonrespons toonde aan dat in het algemeen de nonrespondenten minder welvarend waren dan de respondenten. Dit gold in gelijke mate voor elk van de drie onderscheiden dataverzamelingsmethoden.

De belangrijkste bevindingen uit de meta-analyse werden door de univariate analyses uit hoofdstuk 5 gerepliceerd. De postenquête resulteerde in meer partiële nonrespons, maar ook in meer 'zelf-onthulling' en minder sociaal-wenselijke antwoorden bij 'gevoelige' vragen (b.v. vragen naar eenzaamheid, inkomen). De data verkregen door middel van telefonische en face-to-face interviews verschilden niet op deze punten. Aanvullende analyses toonden kleine verschillen in antwoordtendenties aan. Zo kozen respondenten, die telefonisch ondervraagd werden, vaker voor een extreem positieve antwoordmogelijkheid.

In sociaal-wetenschappelijk onderzoek worden vaak schalen of subtests gebruikt die uit meerdere vragen bestaan. Uit de psychometrische analyses in hoofdstuk 6 blijkt een lichte invloed van de gebruikte dataverzamelingmethode op zowel de betrouwbaarheid als de schaalbaarheid. Wanneer de vragen gesteld werden in een postenquête dan was de klassieke betrouwbaarheid van de schaal hoger dan in beide interview-condities. Ook de resultaten van een Mokken schaalanalyse geven aan dat de gegevens verkregen via de post-enquête beter aan het schaalmodel voldoen. Tevens bleek dat bij de postenquête minder individuele respondenten met afwijkende antwoordpatronen gevonden werden. Opnieuw bleken er weinig verschillen tussen het telefonische en het face-to-face interview gevonden te worden.

In hoofdstuk 7 werden twee inhoudelijke modellen - een pad-model over gevoelens van eenzaamheid en een factor-analytisch meetmodel over de structuur van het begrip welbevinden - via een Lisrel multi-groep analyse met elkaar vergeleken. De resultaten geven redenen voor bezorgdheid. Weliswaar werden steeds dezelfde dimensie en structuur teruggevonden voor de drie verschillende dataverzamelingsmethoden, maar de restricties met betrekking tot gelijke parameterwaarden voor alle drie de dataverzamelingsmethoden konden niet gehandhaafd blijven. De geschatte parameterwaarden verschilden dermate tussen de dataverzamelingsmethoden dat bij verschillende dataverzamelingsmethoden ook verschillende inhoudelijke conclusies getrokken kunnen worden over de sterkte van de invloed van de ene variabele op de andere variabele.

Tot slot wordt in hoofdstuk 8 een korte samenvatting van de resultaten gegeven en worden de bevindingen geëxtrapoleerd naar computergestuurde dataverzamelingsmethoden.

APPENDIX A

BIBLIOGRAPHY AND CONCISE SUMMARY

A.1 Bibliography of Mode Comparison Studies

Aakster, C.W. (1968). Vergelijking van schriftelijke en mondelinge enquête [Comparison of mail and home survey]. *Sociologische Gids, 15*, 322-326.

Aneshensel, C.S., Frerichs, R.R., Clark, V.A., & Yokopenic, P.A. (1982). Measuring depression in the community. A comparison of telephone and personal interviews. *Public Opinion Quarterly, 46*, 110-121.

Assael, H., & Keon, J. (1982). Nonsampling versus sampling errors in survey research. *Journal of Marketing, 46*, 114-123.

Ayidiya, S.A., & McClendon, M.J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229-247.

Bishop, G.F., Hippler, H-J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 321-340). New York: Wiley.

Bushery, J.M., Cowan, C.D., & Murphy, L.R. (1978). Experiments in telephone-personal visit surveys. American Statistical Association, *1978 Proceedings of the Section on Survey Research Methods*, 564-569.

Cahalan, D. (1960). Measuring newspaper readership by telephone: two comparisons with face to face interviews. *Journal of Advertising Research, 1*, 1-6.

Cannell, Ch.F., & Fowler, F.J. jr. (1963). Comparison of a self-enumerative procedure and a personal interview: a validity study. *Public Opinion Quarterly, 27*, 251-263.

Cannell, Ch.F., & Fowler, F.J. jr. (1964). A note on interviewer effect in self-enumerative procedures. *American Sociological Review, 24*, 270.

Colombotos, J. (1965). The effects of personal versus telephone interviews on socially acceptable responses. *Public Opinion Quarterly, 29*, 457-458.

Colombotos, J. (1969). Personal versus telephone interviews; effect on responses. *Public Health Report, 84*, 773-782.

Dillman, D.A., & Mason, R.G. (1984). *The influence of survey method on question response.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Delavan, Wisconsin.

Ellis, A. (1947). Questionnaire versus interview methods in the study of human love relationships. *American Sociological Review, 12*, 541-553.

Groves, R.M. (1978). On the mode of administering a questionnaire and responses to open-ended items. *Social Science Research, 7*, 257-271.

Groves, R.M. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly, 43*, 190-205.

Groves, R.M., & Kahn, R.L. (1979). *Surveys by telephone; A national comparison with personal interviews.* New York: Academic Press.

Henson, R., Cannell, Ch.F., & Roth, A. (1978). Effects of interview mode on reporting of moods, symptoms, and need for social approval. *Journal of Social Psychology, 105*, 123-129.

Herman, M.B. (1977). Mixed-mode data collection: Telephone and personal interviewing. *Journal of Applied Psychology, 62*, 399-404.

Herzog, A. R., Rodgers, W.L., & Kulka, R.A. (1983). Interviewing older adults: A comparison of telephone and face to face modalities. *Public Opinion Quarterly, 47*, 405-418.

Hinkle, A.L., & King, G.D. (1978). A comparison of three survey methods to obtain data for community mental health program planning. *American Journal of Community Psychology, 6*, 389-397.

Hochstim, J.R. (1962). Comparison of three information gathering strategies in a population study of sociometrical variables. American Statistical Association, *1962 Proceedings of the Social Statistics Section*, 154-159.

Hochstim, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62*, 976-989.

Janofsky, A.I. (1971). Affective self-disclosure in telephone vs face to face interviews. *Journal of Humanistic Psychology, 11*, 93-103.

Johnson, T.P., Hougland, J.G. jr., & Clayton, R.R. (1987). *Obtaining reports of sensitive behavior: A comparison from telephone and face to face interviews.* Paper presented at the International Conference on Telephone Survey Methodology. Charlotte, North Carolina. (see also Social Science Quarterly, 70, 174-183).

Jordan, L.A., Marcus, A.C., & Reeder, L.G. (1978). Response styles in telephone and household interviewing: A field experiment from the Los Angeles health survey. American Statistical Association, *1978 Proceedings of the Section on Survey Research Methods*, 362-366.

Jordan, L.A. Marcus, A.C., & Reeder, L.G. (1980). Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly, 44*, 210-222.

Kerssemakers, F.A.M. (1983). *An empirical comparison of two modes of data collection: The same survey by telephone and in person.* Den Haag: Centraal Bureau voor de Statistiek.

Kersten, H.M.P., & Moning, H.J. (1985). Differences in estimates due to changes in methods of data collection. *Kwantitatieve Methoden, 19*, 31-47.

Klecka, W.R., & Tuchfarber, A.J. (1978). Random digit dialing: A comparison to personal surveys. *Public Opinion Quarterly, 42*, 105-114.

Knudsen, D.D., Pope, H., & Irish, D.D. (1967). Response differences to questions on sexual standards: An interview-questionnaire comparison. *Public Opinion Quarterly, 31*, 290-297.

Körmendi, E. (1988). The quality of income information in telephone and face to face surveys. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 341-356). New York: Wiley.

Körmendi, E., & Noordhoek, J. (1989). *Data quality and telephone interviews*. Copenhagen: Danmarks Statistik.

Krohn, M., Waldo, G.P., & Chiricos, Th.G. (1975). Selfreported delinquency: A comparison of structured interviews and self administered checklists. *Journal of Criminal Law and Criminology, 65*, 545-553.

Kulka, R.A., Weeks, M.F., Lessler, J.T., & Whitmore, R.W. (1982). A comparison of the telephone and personal interview modes for conducting local household health surveys. *NCHSR Proceedings on Health Survey Research Methods*, 116-127.

Larsen, O.N. (1952). The comparative validity of telephone and face to face interviews in the measurement of message diffusion from leaflets. *American Sociological Review, 17*, 471-476.

Locander, W., Sudman, S., & Bradburn, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association, 71*, 269-275.

Mangione, Th.W., Hingson, R., & Barrett, J. (1982). Collecting sensitive data. *Sociological Methods and Research, 10*, 337-346.

McDonagh, E.C., & Rosenblum, A. L. (1965). A comparison of mailed questionnaires and subsequent structured interviews. *Public Opinion Quarterly, 29*, 131-136.

McGuire, B., & Leroy, D.J. (1977). Comparison of mail and telephone methods of studying media contactors. *Journal of Broadcasting, 21*, 391-400.

Miller, P.V. (1982). A comparison of telephone and personal interviews in the health interview survey. *NCHSR Proceedings on Health Survey Research Methods*. 135-145.

Nederhof, A.J. (1984). Visibility of response as a mediating factor in equity research. *Journal of Social Psychology, 122*, 211-215.

Nuckols, R.C. (1964). Personal interview versus mail panel survey. *Journal of Marketing Research, 1*, 11-16.

Oakes, R.H. (1954). Differences in responsiveness in telephone versus personal interviews. *The Forum, 19*, 169.

O'Dell, W.F. (1962). Personal interviews or mail panels. *Journal of Marketing, 26*, 34-39.

O'Toole, B.I., Battistutta, D., Long, A., & Crouch, K. (1986). A comparison of costs and data quality of three health survey methods: Mail, telephone and personal home interview. *American Journal of Epidemiology, 124*, 317-328.

Prawl, W.L., & Jorns, W.J. (1976). Reviewing county extension programs. *Journal of Extension, 14*, 11-17.

Rogers, T.F. (1976). Interviews by telephone and in person: Quality of response and field performance. *Public Opinion Quarterly, 40*, 51-65.

San Augustine, A.J., & Friedman, H.H. (1978). The use of the telephone interview in obtaining information of a sensitive nature: A comparative study. American Statistical Association, *1978 Proceedings of the Section on Survey Research Methods*, 559-561.

Schmiedeskamp, J.W. (1962). Reinterviews by telephone. *Journal of Marketing, 26*, 28-34.

Siemiatycki, J. (1979). A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health, 69*, 238-245.

Siemiatycki, J., & Campbell, S. (1984). Non-response bias and early versus all responders in mail and telephone surveys. *American Journal of Epidemiology, 120*, 291-301.

Siemiatycki, J., Campbell, S., Richardson, L., & Aubert, D. (1984). Quality of response in different population groups in mail and telephone surveys. *American Journal of Epidemiology, 120*, 302-314.

Sudman, S., & Ferber, R. (1974). A comparison of alternative procedures for collecting consumer expenditure data for frequently purchased products. *Journal of Marketing Research, 11*, 128-135.

Sudman, S., Greely, A., & Pinto, L. (1965). The effectiveness of self-administered questionnaires. *Journal of Marketing Research, 2*, 293-297.

Sykes, W., & Collins, M. (1988). Effects of mode of interview: experiments in the U.K. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 301-320). New York: Wiley.

Van Amstel, R. (1981). Postenquete of bezoekenquete? [Mail survey or personal delivered questionnaire?]. *Tijdschrift voor Sociale Geneeskunde, 59*, 164-169.

Van Sonsbeek, J.L.A., & Stronkhorst, L.H. (1983). *Vergelijking van drie waarnemingsvarianten bij de meting van medische consumptie* [A comparison of data collection methods in the measurement of medical consumption]. Den Haag: Centraal Bureau voor de Statistiek.

Walsh, W. B. (1967). Validity of self-report. *Journal of Counseling Psychology, 14*, 18-23.

Walsh, W. B. (1968). Validity of self-report: Another look. *Journal of Counseling Psychology, 15*, 180-188.

Walsh, W. B. (1969). Self-report under socially undesirable and distortion conditions. *Journal of Counseling Psychology, 16*; 569-574.

Wheatly, J.J. (1973). Self-administered written questionnaires or telephone interviews. *Journal of Marketing Research, 10*, 94-96.

Wierdsma, A.I., & Garretsen, H.F.L. (1985). Gezondheidsenquete per post of op bezoek? Resultaten van een vooronderzoek in Rotterdam [Health surveys by mail or home interview?]. *Tijdschrift voor Sociale Gezondheidszorg, 63*, 592-595.

Williams, W. jr., & LeRoy, D. (1976). Alternative methods of measuring public radio audiences: A pilot project. *Journalism Quarterly, 53*, 516-521.

Wiseman, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly, 36*, 105-108.

Woltman, H.F., Turner, A.G., & Bushery, J.M. (1980). A comparison of three mixed-mode interviewing procedures in the national crime survey. *Journal of the American Statistical Association, 75*, 534-543.

Yaffe, R., Shapiro, S., Fuchsberg, R.R., Rhode, Ch.A., & Corpeno, H.C. (1978). Medical economics survey-methods study, cost effectiveness of alternative survey strategies. *Medical Care, 16*, 641-659.

Zeiner-Henrikson, T. (1972). Comparison of personal interview and postal inquiry methods for assessing prevalence of angina and possible infarction. *Journal of Chronical Disease, 25*, 433-440.

A.2 Concise Summary of the Conclusions Quoted in the Studies Reviewed

When studies are partly reported in more than one article, the first author and year of publication of the additional articles are given in parentheses.
First author, year of publication, subject, type of comparison (e.g., face to face versus telephone, face to face versus mail, mail versus telephone) and summary conclusion as given in the original articles.

First author	Year	Subject	Comparison and Conclusion
Aakster	1968	health	Mail vs self-administered questionnaire in presence of interviewer. Mail survey more item non-response on compl x questions, but S.A.Q. with interviewer presents more item non-response on sensitive questions.
Aneshensel	1982	health/ depression	Face to face vs telephone. No significant mode effects.
Assael	1982	consumer/ business	Face to face, telephone and mail compared. Telephone less accurate; mail most effective in reducing response error.
Ayidiya	1990	various topics	Mail vs interview (Face to face and telephone). In general, order effects less likely in mail, but form effects and a recency effect equally likely.
Bishop	1987	various topics	Mail vs telephone. Order effects less likely in mail, form effects as likely.
Bushery	1978	victimization	Face to face vs telephone. Personal visit interviews tend to produce slightly better data.
Cahalan	1960	consumer/ newspaper	Face to face vs telephone. No differences.
Cannell (also Cannell 1964)	1963	health	Self-administered vs face to face. When respondent has records, self-administered is more accurate, no difference in social desirability bias.
Colombotos (also Colombotos 1965)	1969	health	Face to face vs telephone. Essentially no differences.
Dillman	1984	housing	Face to face/telephone/mail. Some evidence of telephone extremeness, mail less extremeness.
Ellis	1947	relationships	Face to face vs mail. Answers on questionnaire more incriminating than in previous interview.
Groves (also Groves 1979a & Groves 1979b)	1978	several topics	Face to face vs telephone. Telephone tends to yield fewer and faster answers.
Henson	1978	health/moods	Face to face vs telephone. Telephone fewer symptoms and more social desirability.
Herman	1977	voting	Face to face vs telephone. In general, no mode effects, but telephone respondents less willing to reveal sensitive information.

First author	Year	Subject	Comparison and Conclusion
Herzog	1983	reanalysis older subjects	Face to face vs telephone. Elderly in general under-represented; little evidence for mode by age interaction.
Hinkle	1978	health/ mental	Face to face/telephone/mail. Both interview methods yield comparable data; mail resulted in more neutral and negative answers.
Hochstim (also Hochstim 1962)	1967	health	Face to face/telephone/mail. Data collection strategies proved to be practically interchangeable.
Janofsky	1971	feelings	Face to face vs telephone. In both modes respondents equally willing to express feelings.
Johnson	1987	drug use	Face to face vs telephone. In person interviews resulted in more reported drug use.
Jordan (also Jordan 1978)	1980	health	Face to face vs telephone. Telephone has more missings on income data, more extremeness, acquiescence & evasiveness.
Kerssemakers	1983	consumer	Face to face vs telephone. Telephone higher percentage don't know. In general, results of the two modes in good agreement.
Kersten	1985	travel	Face to face vs telephone. Small differences. (both strategies used additional diary)
Klecka	1978	victimization	Face to face vs telephone. Telephone survey with RDD can replicate face to face survey with complex sampling.
Knudsen	1967	relations/ sex	Face to face vs self-administered questionnaire. Questionnaire lower proportion women with restrictive norms.
Körmendi (also Körmendi 1989)	1988	various topics	Face to face vs telephone. No differences in general; no differences on income.
Krohn	1975	selfreported delinquency	Face to face interview vs self-administered questionnaire. No reason to assume one technique is any more valid than other.
Kulka	1982	health	Face to face vs telephone. No important mode effects.
Larson	1952	leaflet messages	Face to face vs telephone. Serious doubt on validity of telephone responses.
Locander	1976	facts (sensitive)	Face to face/telephone/ self-administered questionnaire/randomized response. None of the methods differed significantly.
Mangione	1982	drinking	Face to face/telephone/self-administered questionnaire. In person more drinking.
McDonagh	1965	general	Face to face vs mail. No statistically significant difference.
McGuire	1977	Media habits	Telephone vs mail. Combination of mail and telephone is best.

First author	Year	Subject	Comparison and Conclusion
Miller	1982	health	Face to face vs telephone. Telephone surveys do not necessarily produce lower quality data.
Nederhof	1984	equity	Face to face vs mail. More altruistic answers in face to face interviews.
Nuckols	1964	finance	Face to face vs mail. Mail panel showed up well: answers more accurate.
Oakes	1954	consumer	Face to face vs telephone. Average number of answers less in telephone survey.
O'Dell	1962	consumer (panels)	Face to face vs mail. Selection of method is decision based on the optimum allocation of the research dollar.
O'Toole	1986	health	Face to face/telephone/mail. Overall no mode differences; mail less complete.
Prawl	1976	education	Telephone vs mail. Telephone data seem highly credible.
Rogers	1976	housing/ services	Face to face vs telephone. Quality of data collected is comparable.
San Augustine	1978	attitudes	Telephone/mail/self-administered questionnaire. Mail low response and more liberal answers; telephone survey preferable.
Schmiedeskamp	1962	finances	Face to face vs telephone reinterview. Telephone some avoiding of definite positions.
Siemiatycki (also Siemiatycki, 1984a & Siemiatycki, 1984b)	1979	health	Face to face/telephone/mail. Mail surveys more valid answers and more willingness to answer sensitive questions.
Sudman	1965	religion/ education	Face to face vs self-administered questionnaire. No large differences, S.A.Q. seems to give better measure of true feelings.
Sudman	1974	consumer	Telephone vs diary. Daily telephone interview not as complete as diary.
Sykes	1988	various topics	Face to face vs telephone. Similarity of answers obtained under different modes.
Van Amstel	1981	health	Mail vs self-administered questionnaire with interviewer. In mail survey more personal problems are reported than in the presence of a interviewer.
Van Sonsbeek	1983	health	Face to face/mail/mixture. Results on medical consumption are very similar.
Walsh Walsh Walsh	1967 1968 1969	education (three replications)	Face to face interview vs (group) questionnaire. No method elicits more accurate selfreports than another.
Wheatly	1973	consumer	Telephone vs questionnaire. No difference in nature of response.
Wierdsma	1985	health	Face to face vs mail. Mail questionnaires are not second to the interview.
Williams	1976	media	Telephone vs mail. Mail surveys more likely premeditated responses.

First author	Year	Subject	Comparison and Conclusion
Wiseman	1972	various topics	Face to face/telephone/mail. Responses not always independent of method.
Woltman	1980	victimization	Mixtures of face to face and telephone interviews. Reported victimization less with telephone interviews as major mode.
Yaffe	1978	health	Face to face vs telephone. In person strategies result in higher accuracy.
Zeiner-Henrikson	1972	cardiac pain	Face to face vs mail (reinterview). Two methods yield much variety, and are not interchangeable.

Note. Country of origin of the studies was the U.S.A., with the exception of Aakster, Kersten, Kerssemakers, Nederhof, Van Amstel Van Sonsbeek, and Wierdsma (The Netherlands), Bishop (America/Germany), Körmendi (Denmark), O'Toole (Australia), Siemiatycki (Canada), Sykes (Great Britain), and Zeiner-Henrikson (Norway).

APPENDIX B

CONTENT OF THE QUESTIONNAIRES

A short description of each section of the questionnaire is given. For each section at least one example is given of the type of questions asked. Appendix B1 includes an English translation of the question text as found in the self-administered questionnaire. Appendix B2 contains the same example questions now worded us used in the telephone survey, appendix B3 contains the wording used in the face to face survey. The complete Dutch text of the final equivalent versions for the mail, telephone, and face to face survey, including the text of interviewer instructions and the response cards, is available as technical report No. 6 (De Leeuw, 1991).

B.1 Mail Survey Questionnaire

Section 1: General happiness question, graphical representation (cf. Cantril, 1965; Hox, 1986).

Here is a picture of a ladder. At the top of the ladder, on the seventh rung, is the best life you might reasonably expect to have. At the bottom, on the first rung, is the worst life you might reasonably expect to have.

DRAWING OF LADDER WITH SEVEN STEPS

Where on the ladder would you say was how happy you felt in the past year, on which rung would you be?

On rung number: .....

Section 2: Five general satisfaction questions; closed questions, five response categories (cf. Andrews & Whithey, 1976; Hox, 1986).

Taking all things together, how satisfied or dissatisfied are you with the home in which you live?

1 VERY DISSATISFIED
2 DISSATISFIED
3 NEITHER SATISFIED NOR DISSATISFIED
4 SATISFIED
5 VERY SATISFIED

Section 3: Eighteen well-being questions; closed questions, two response categories. Both positively and negatively formulated questions were used (Extended Affect Balance Scale; see Bradburn, 1969; Hox, 1986).

During the past few weeks, did you ever feel that things were going your way?

1 NO
2 YES

During the past few weeks, did you ever feel depressed or very unhappy?

1 NO
2 YES

Section 4: Eleven loneliness questions; closed questions, three response categories. Both positively and negatively formulated questions were used (cf. De Jong-Gierveld & Kamphuis, 1985), followed by eight self-evaluation questions; closed questions, three response categories (cf. Dykstra, forthcoming).

Loneliness:

There is always someone that I can talk to about my day to day problems

1 YES
2 MORE OR LESS
3 NO

I miss having a really close friend

1 YES
2 MORE OR LESS
3 NO

Self-evaluation:

I am rather sure of myself

1 YES
2 MORE OR LESS
3 NO

Section 5: Four questions on the social network (one open question on the extension of the network and three checklists asking for core network members; eleven response categories).

Are there people around (in your proximity) who are very important to you?

1 NO
2 YES ——————————→ How many? ..... people

Who is -for you- the most important person to discuss personal topics with.
(Circle your answer).

- spouse, partner/significant other
- (male) friend
- (female) friend
- father/mother
- brother/sister
- son/daughter
- other relative
- neighbor
- acquaintance
- colleague, former colleague
- someone else, that is .....

Section 6: Ten questions on the financial situation (open questions, closed questions with response categories ranging from three to five categories, and checklists with nine to eleven response categories).

[In every household people have to spend money on food, clothes, housing, etc. How do you finance this, or in other words]
What is the main source of income in your household?

- Earned income
- Unearned income
- Pension, Life annuity, Early retirement pension
- General Retirement Pension Act, General Widow & Orphans Act
- Income support, social security
- Disability benefit
- Reduced pay, Unemployment Act, Unemployment Assistance Act
- Other social security benefits: .....
- Scholarship, grant
- Alimony
- Financial support by parents/guardians
- Other: .....

Compared to other people you know, would you say you are much better off, somewhat better off, just as well off, worse off, or much worse off?

1 MUCH BETTER
2 SOMEWHAT BETTER
3 JUST AS WELL
4 WORSE
5 MUCH WORSE

Are there things that are important to you, but that you cannot afford financially?

1 NO
2 YES → Could you give a short description?

What is the *net* monthly income of your household?

Section 7: Five questions on survey preference and participation (open questions and closed questions with two to four response categories); followed by five questionnaire threat questions (closed, two response categories).

Survey preference and participation:

Have you ever refused to participate in a survey?

1 NO
2 YES → Why?

Questionnaire threat:

[On the whole, how do *you* think people feel about completing *this* questionnaire]

Most people will find the questions threatening

1 YES
2 NO

Section 8: Ten standard demographic questions (open questions and closed questions with two to eight response categories).

Do you have children?

1 NO
2 YES: ..... children

Section 9: Ending the questionnaire (one closed, one open question).

How did you feel about completing this questionnaire; was it

1 VERY ENJOYABLE
2 ENJOYABLE
3 NEITHER ENJOYABLE NOR UNPLEASANT
4 UNPLEASANT
5 VERY UNPLEASANT

Is there anything else you would like to tell us? If so, please use this space for that purpose. Also, any comments you wish to make about this questionnaire or about this survey will be highly appreciated.

B.2 Telephone Survey Questionnaire

Interviewer instructions are written in the text between parentheses, using italic script.
A general rule was that only texts printed in lowercase are spoken by the interviewer. Everything in UPPERCASE is not read out aloud.

Section 1: General happiness question (cf. Cantril, 1965; Hox, 1986).

First of all: Suppose you have a ladder with seven rungs. At the top of the ladder, on the seventh rung, is the best life you might reasonably expect to have. At the bottom, on the first rung, is the worst life you might reasonably expect to have. Where on the ladder would you say was how happy you felt in the past year, on which rung would you be?

(INT: ONE ANSWER; WHEN NECESSARY REPEAT: the first rung is the worst life, the seventh rung the best life you might reasonable expect to have. [On which rung of the ladder would you be, on the first, the second, the third, the fourth, the fifth, the sixth, or the seventh rung].)

("worst") 1 2 3 4 5 6 7 ("best")
88 (Do not know) 99 (no answer)

Section 2: Five general satisfaction questions; closed questions, five response categories (cf. Andrews & Whithey, 1976; Hox, 1986).

Taking all things together, how satisfied or dissatisfied are you with the home in which you live. Are you very dissatisfied, dissatisfied, neither satisfied nor dissatisfied, satisfied, or very satisfied?

1 VERY DISSATISFIED
2 DISSATISFIED
3 NEITHER SATISFIED NOR DISSATISFIED
4 SATISFIED
5 VERY SATISFIED
8 DO NOT KNOW
9 NO ANSWER

(INT: WHEN NECESSARY : Shall I repeat the possibilities? REPEAT: Taking all things together are you very dissatisfied, dissatisfied, neither satisfied nor dissatisfied, satisfied, or very satisfied).

Section 3: Eighteen well-being questions; closed questions, two response categories. Both positively and negatively formulated questions were used (Extended Affect Balance Scale; see Bradburn, 1969; Hox, 1986). At the end of the first two questions, the interviewer explicitly said: 'no or yes' (see first example), in the next twelve questions this was not done (see second example).

(INT: WHEN NECESSARY REPEAT AFTER EACH QUESTION: no or yes?)

During the past few weeks, did you ever feel that things were going your way: no or yes?

1 NO
2 YES
8 DO NOT KNOW
9 NO ANSWER

During the past few weeks, did you ever feel depressed or very unhappy?

1 NO
2 YES
8 DO NOT KNOW
9 NO ANSWER

Section 4: Eleven loneliness questions; closed questions, three response categories. Both positively and negatively formulated questions were used (cf. De Jong-Gierveld & Kamphuis, 1985), followed by eight self-evaluation questions; closed questions, three response categories (cf. Dykstra, forthcoming). At the end of the first three questions the interviewer explicitly said: 'yes, more-or-less, or no?' (see first example loneliness). In the next fifteen questions this was not done (second example loneliness).

(INT: WHEN NECESSARY REPEAT RESPONSE CATEGORIES: 'yes', more-or-less, no')

Loneliness:

There is always someone that I can talk to about my day to day problems

1 YES
2 MORE OR LESS
3 NO

I miss having a really close friend

1 YES
2 MORE OR LESS
3 NO

Self-evaluation:

I am rather sure of myself

1 YES
2 MORE OR LESS
3 NO

Section 5: Four questions on the social network (one open question on the extension of the network and three checklists asking for core network members; eleven response categories).

Are there people around (in your proximity) who are very important to you?

(INT: IF YES THEN QUESTION 45, OTHERWISE NEXT PAGE)

1 NO
2 YES ——→
8 DO NOT KNOW
9 NO ANSWER

Q-45 How many?

..... people
77 NOT APPLICABLE
88 DO NOT KNOW
99 NO ANSWER

(the following question was on the next page)

The following list contains people, who you may meet in your day to day life.

(READ LIST)

- spouse, partner/significant other
- (male) friend
- (female) friend
- father/mother
- brother/sister
- son/daughter
- other relative
- neighbor
- acquaintance
- colleague, former colleague
- someone else, that is .....

Please indicate who are -for you- the three most important people. That is, people who are so important to you that you will discuss personal topics with them. You may choose from the list I just read to you.

Who is -for you- the **most important** person to discuss personal topics with. Shall I repeat the list? *(INT: REPEAT LIST IF NECESSARY)*

The most important person is .....

88 DO NOT KNOW
99 NO ANSWER

Section 6: Ten questions on the financial situation (open questions, closed questions with response categories ranging from three to five categories, and checklists with nine to eleven response categories).

[In every household people have to spend money on food, clothes, housing, etc. How do you finance this, or in other words]
What is the main source of income in your household, is that?

1 Earned income
2 Unearned income
3 Pension, Life annuity, Early retirement pension
4 General Retirement Pension Act, General Widow & Orphans Act
5 Income support, social security
6 Disability benefit
7 Reduced pay, Unemployment Act, Unemployment Assistance Act
8 Other social security benefits *(INT: PROBE: which?)*
.....
9 Scholarship, grant
10 Alimony
11 Financial support by parents/guardians
12 Other *(INT: PROBE: what is the main source of income?)*
.....

Shall I repeat the possibilities? *(INT: REPEAT IF NECESSARY)*
(88 DO NOT KNOW)
(99 NO ANSWER)

(INT: IF MORE THAN ONE ANSWER, FIRST REPEAT what is the main source of income?. IF RESPONDENT STILL GIVES MORE THAN ONE SOURCE, ACCEPT IT AND CIRCLE THOSE RESPONSES)

Compared to other people you know, would you say you are much better off, somewhat better off, just as well off, worse off, or much worse off?

1 MUCH BETTER
2 SOMEWHAT BETTER
3 JUST AS WELL
4 WORSE
5 MUCH WORSE

Are there things that are important to you, but that you cannot afford financially?

1 NO *(CONTINUE Q. 56)*
2 YES *(CONTINUE Q. 55*

Q. 55 Yes?, could you give a short description?

What is the *net* monthly income of your household?

(INT: ROUND OFF TO GUILDERS)

..... guilders net each month

INT: RESPONSE WAS:
1 ROUNDED OFF IN GUILDERS BY RESPONDENT
2 REPORTED IN GUILDERS AND CENTS
3 APPROXIMATE
7 NOT APPLICABLE

(INT: ACCEPT A REFUSAL WITHOUT COMMENT AND CONTINUE WITH NEXT QUESTION. This was followed by several scripts for angry or anxious respondents).

Section 7: Five questions on survey preference and participation (open questions and closed questions with two to four response categories); followed by five questionnaire threat questions (closed, two response categories).

Survey preference and participation:

Have you ever refused to participate in a survey?

1 NO (continue Q63)
2 YES → Q62 Why?
.....
8 DO NOT KNOW
9 NO ANSWER

? NOT APPLICABLE
8 DO NOT KNOW
9 NO ANSWER

Questionnaire threat:

[On the whole, how do *you* think people feel about completing *this* questionnaire]

Most people will find the questions threatening: yes or no?

1 YES
2 NO
8 DO NOT KNOW
9 NO ANSWER

Section 8: Ten standard demographic questions (open questions and closed questions with two to eight response categories).

Do you have children?
(INT: IF YES THAN PROBE: how many?)

1 NO (continue Q77
2 YES: ..... children
88 DO NOT KNOW (continue Q77)
99 NO ANSWER (continue Q77)

Section 9: Ending the interview (one closed, one open question).

How did you feel about completing this questionnaire; was it very enjoyable, enjoyable, neither enjoyable nor unpleasant, unpleasant or very unpleasant?

1 VERY ENJOYABLE
2 ENJOYABLE
3 NEITHER ENJOYABLE NOR UNPLEASANT
4 UNPLEASANT
5 VERY UNPLEASANT
8 DO NOT KNOW
9 NO ANSWER

Is there anything else you would like to tell us?
(INT: WRITE DOWN THE ANSWERS IN THE SPACE BELOW. YOU CAN ALSO USE THE SPACE ON THE LEFT PAGE).

B.3 Face to Face Survey Questionnaire

Interviewer instructions are written in the text between parentheses, using italic script.
A general rule was that only texts printed in lowercase are spoken by the interviewer. Everything in UPPERCASE is not read out aloud.

Section 1: General happiness question (cf. Cantril, 1965; Hox, 1986).

(INT: HAND OVER BOOKLET OPEN AT RESPONSE CARD A)

Here on this card is a picture of a ladder with seven rungs. At the top of the ladder, on the seventh rung, is the best life you might reasonably expect to have. At the bottom, on the first rung, is the worst life you might reasonably expect to have. Where on the ladder would you say was how happy you felt in the past year, on which rung would you be?

("worst") 1 2 3 4 5 6 7 ("best")
88 (Do not know) 99 (no answer)

Section 2: Five general satisfaction questions; closed questions, five response categories (cf. Andrews & Whithey, 1976; Hox, 1986).

Please look at card B

Taking all things together, how satisfied or dissatisfied are you with the home in which you live? You may choose from the responses on the card

1 VERY DISSATISFIED
2 DISSATISFIED
3 NEITHER SATISFIED NOR DISSATISFIED
4 SATISFIED
5 VERY SATISFIED
8 DO NOT KNOW
9 NO ANSWER

(INT: IF NECESSARY REPEAT: Please choose that answer that is closest to your own feeling [you may choose from the responses on the card]

Section 3: Eighteen well-being questions; closed questions, two response categories. Both positively and negatively formulated questions were used (Extended Affect Balance Scale; see Bradburn, 1969; Hox, 1986). At the end of the first two questions, the interviewer explicitly said: 'no or yes' (see first example), in the next twelve questions this was not done (see second example).

(INT: WHEN NECESSARY REPEAT AFTER EACH QUESTION: no or yes?)

During the past few weeks, did you ever feel that things were going your way: no or yes?

1 NO
2 YES
8 DO NOT KNOW
9 NO ANSWER

During the past few weeks, did you ever feel depressed or very unhappy?

1 NO
2 YES
8 DO NOT KNOW
9 NO ANSWER

Section 4: Eleven loneliness questions; closed questions, three response categories. Both positively and negatively formulated questions were used (cf. De Jong-Gierveld & Kamphuis, 1985), followed by eight self-evaluation questions; closed questions, three response categories (cf. Dykstra, forthcoming). At the end of the first three questions the interviewer explicitly said: 'yes, more-or-less, or no?' (see first example loneliness). In the next fifteen questions this was not done (second example loneliness).

(INT: WHEN NECESSARY REPEAT RESPONSE CATEGORIES: 'yes', more-or-less, no')

Loneliness:

There is always someone that I can talk to about my day to day problems

1 YES
2 MORE OR LESS
3 NO

I miss having a really close friend

1 YES
2 MORE OR LESS
3 NO

Self-evaluation:

I am rather sure of myself

1 YES
2 MORE OR LESS
3 NO

Section 5: Four questions on the social network (one open question on the extension of the network and three checklists asking for core network members; eleven response categories).

Are there people around (in your proximity) who are very important to you?

(INT: IF YES THEN QUESTION 45, OTHERWISE NEXT PAGE)

1 NO
2 YES → Q-45 How many?
..... people
77 NOT APPLICABLE
88 DO NOT KNOW
99 NO ANSWER
8 DO NOT KNOW
9 NO ANSWER

(the following question was on the next page)

Please take card C
On this card is a list containing people, who you may meet in your day to day life.
Please indicate who are -for you- the three most important people. That is, people who are so important to you that you will discuss personal topics with them. You may choose from the list you have in front of you.

(INT: DO NOT READ THE LIST OUT LOUD. IF NECESSARY: 'Please choose from the list on the card' OR IF RESPONDENT HAS TROUBLE READING: 'the choices are: READ LIST)

Who is -for you- the **most important** person to discuss personal topics with. Shall I repeat the list?

The most important person is .....

88 DO NOT KNOW
99 NO ANSWER

INT: LIST THAT IS ON CARD C

- *spouse, partner/significant other*
- *(male) friend*
- *(female) friend*
- *father/mother*
- *brother/sister*
- *son/daughter*
- *other relative*
- *neighbor*
- *acquaintance*
- *colleague, former colleague*
- *someone else, that is .....*

Section 6: Ten questions on the financial situation (open questions, closed questions with response categories ranging from three to five categories, and checklists with nine to eleven response categories).
Please look at the next card (CARD E).

[In every household people have to spend money on food, clothes, housing, etc. How do you finance this, or in other words]
What is the main source of income in your household, is that?

1 EARNED INCOME
2 UNEARNED INCOME
3 PENSION, LIFE ANNUITY, EARLY RETIREMENT PENSION
4 GENERAL RETIREMENT PENSION ACT, GENERAL WIDOW & ORPHANS ACT
5 INCOME SUPPORT, SOCIAL SECURITY
6 DISABILITY BENEFIT
7 REDUCED PAY, UNEMPLOYMENT ACT, UNEMPLOYMENT ASSISTANCE ACT
8 OTHER SOCIAL SECURITY BENEFITS *(INT: PROBE: which?)*
.....
9 SCHOLARSHIP, GRANT
10 ALIMONY
11 FINANCIAL SUPPORT BY PARENTS/GUARDIANS
12 OTHER *(INT: PROBE: what is the main source of income?)*
.....

(88 DO NOT KNOW)
(99 NO ANSWER)

(INT: IF MORE THAN ONE ANSWER, FIRST REPEAT what is the main source of income? IF RESPONDENT STILL GIVES MORE THAN ONE SOURCE, ACCEPT IT AND CIRCLE THOSE RESPONSES)

Compared to other people you know, would you say you are much better off, somewhat better off, just as well off, worse off, or much worse off?

1 MUCH BETTER
2 SOMEWHAT BETTER
3 JUST AS WELL
4 WORSE
5 MUCH WORSE

Are there things that are important to you, but that you cannot afford financially?

1 NO *(CONTINUE Q. 56)*
2 YES *(CONTINUE Q. 55*

Q. 55 Yes?, could you give a short description?

What is the *net* monthly income of your household?

(INT: ROUND OFF TO GUILDERS)

..... guilders net each month

INT: RESPONSE WAS: *1 ROUNDED OFF IN GUILDERS BY RESPONDENT* *2 REPORTED IN GUILDERS AND CENTS* *3 APPROXIMATE* *7 NOT APPLICABLE*

(INT: ACCEPT A REFUSAL WITHOUT COMMENT AND CONTINUE WITH NEXT QUESTION. This was followed by several scripts for angry or anxious respondents).

Section 7: Five questions on survey preference and participation (open questions and closed questions with two to four response categories); followed by five questionnaire threat questions (closed, two response categories).

Survey preference and participation:

Have you ever refused to participate in a survey?

1 NO (continue Q63)
2 YES → Q62 Why?
.....

7 NOT APPLICABLE
8 DO NOT KNOW
9 NO ANSWER

Questionnaire threat:

[On the whole, how do *you* think people feel about completing *this* questionnaire]

Most people will find the questions threatening: yes or no?

1 YES
2 NO
8 DO NOT KNOW
9 NO ANSWER

Section 8: Ten standard demographic questions (open questions and closed questions with two to eight response categories).

Do you have children?
(INT: IF YES THAN PROBE: how many?)

1 NO (continue Q77)
2 YES: ..... children
88 DO NOT KNOW (continue Q77)
99 NO ANSWER (continue Q77)

Section 9: Ending the questionnaire/interview (one closed, one open question).

Please take the last card in front of you.

How did you feel about completing this questionnaire.

1 VERY ENJOYABLE
2 ENJOYABLE
3 NEITHER ENJOYABLE NOR UNPLEASANT
4 UNPLEASANT
5 VERY UNPLEASANT
8 DO NOT KNOW
9 NO ANSWER

Is there anything else you would like to tell us?

(INT: WRITE DOWN THE ANSWERS IN THE SPACE BELOW. YOU CAN ALSO USE THE SPACE ON THE LEFT PAGE).

APPENDIX C

MARGINAL DISTRIBUTIONS OF BACKGROUND VARIABLES

C.1 Gender by Method.

	Mail	Method F to F	Tel.	CATI
Male	55.5%	41.6%	47.7%	45.5%
Female	45.5%	58.4%	52.3%	54.5%
	100%	100%	100%	100%
N	254	243	266	77

C.2 Marital Status by Method

	Mail	Method F to F	Tel.	CATI
Never married	26.0%	35.0%	35.8%	31.2%
Married	63.8%	44.4%	47.2%	58.4%
Divorced	5.1%	10.3%	6.0%	5.2%
Widowed	5.1%	10.3%	10.9%	5.2%
	100%	100%	100%	100%
N	254	243	265	77

C.3 Age Distribution by Method

	Mail	Method F to F	Tel.	CATI
Mean	44.7	44.8	45.3	42.6
Stand. Dev.	15.5	17.5	18.3	16.4
N	254	243	265	77

C.4 Education by Method

	Mail	Method F to F	Tel.	CATI
Elementary (1)	11.2%	10.7%	11.3%	15.6%
(2)	15.6%	14.8%	18.1%	9.1%
(3)	15.2%	10.7%	14.7%	16.9%
(4)	14.0%	10.3%	15.1%	14.3%
(5)	15.2%	16.5%	12.8%	20.8%
(6)	19.6%	23.9%	16.2%	13.0%
University (7)	9.2%	13.2%	11.7%	10.4%
	100%	100%	100%	100%
N	250	243	265	77

C.5 Having Children by Method

	Mail	Method F to F	Tel.	CATI
No	36.2%	41.2%	44.4%	35.1%
Yes	63.8%	58.8%	55.6%	64.9%
	100%	100%	100%	100%
N	254	243	266	77

C.6 Previous Interview Experience by Method

	Mail	Method F to F	Tel.	CATI
No	26.9%	19.8%	27.5%	18.4%
Yes	73.1%	80.2%	72.5%	81.6%
	100%	100%	100%	100%
N	253	243	265	76

REFERENCES

Akuto, H. (1992). Current status of research in telecommunication in Japan. In: L. Lebart (Ed.) *Quality of information in sample surveys* (pp. 169-183). Paris: Dunod.

Andrews, F.M., & Withey, S.B. (1978). *Social indicators of well-being.* New York: Plenum.

Aneshensel, C.S., Frerichs, R.R., Clark, V.A., & Yokopenic, P.A. (1982). Measuring depression in the community. A comparison of telephone and personal interviews. *Public Opinion Quarterly, 46*, 110-121.

Argyle, M. (1973). *Social Interaction.* London: Tavistock.

Argyle, M. & Dean, J. (1965). Eye-contact, distance and affiliation. *Sociometry, 28*, 289-304.

Asimov, I. (1971). *Foundation.* London: Panther Books.

Ayidiya, S.A., & McClendon, M.J. (1990). Response effects in mail surveys. *Public Opinion Quarterly, 54*, 229-247.

Bailar, B. (1984). The quality of survey data. *American Statistical Association, 1984 Proceedings of the section on survey research methods*, 43-52.

Bangert-Drowns, R.L. (1986). Review of developments in meta- analytic method. *Psychological Bulletin, 99*, 388-399.

Bassili, J.N., & Fletcher, J.F. (1991). Response time measurement in survey research: A method for CATI and a new look at nonattitudes. *Public Opinion Quarterly, 55*, 331-346.

Belson, W.A. (1981). *The design and understanding of survey questions.* Aldershot: Gower.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-600.

Bentler, P.M., Jackson, D., & Messick, S. (1971). Identification of content and style: A two dimensional interpretation of acquiescence. *Psychological Bulletin, 76*, 186-204.

Betlehem, J.G., & Kersten, H.M.P. *Werken met non-respons* [Working with non-response]. Doctoral dissertation, University of Amsterdam, Amsterdam.

Betlehem, J.G., & Kersten, H.M.P. (1981). The nonresponse problem. *Survey Methodology, 7*, 130-156.

Bishop, G.F., Hippler, H-J., Schwarz, N., & Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 321-340). New York: Wiley.

Biemer, P.P. (1988). Measuring data quality. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 273-282). New York: Wiley.

Block, J. (1971). On further conjecture regarding acquiescence. *Psychological Bulletin, 76*, 205-210.

Bollen, K.E. (1989). *Structural equations with latent variables.* New York: Wiley.

Bradburn, N.M. (1969). *The structure of well-being.* Chicago: Aldine.

Bradburn, N.M. (1983). Response effects. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), *Handbook of survey research* (pp. 289-328). New York: Academic Press.

Breed, P.C.M., & Swaans-Joha, B.C. (1986). *Doven in Nederland* [Deaf people in the Netherlands]. Doctoral dissertation, University of Amsterdam, Amsterdam.

Brinkman, W. (1987). *Een assertiviteitsschaal II* [Measuring assertivity II]. Amsterdam: University of Amsterdam, Department of Psychology.

Bronner, A.E. (1980). Telefonisch onderzoek [Telephone surveys]. *Methoden en Data Nieuwsbrief van de Sociaal Wetenschappelijke Sectie van de Vereniging voor Statistiek, 5*, 145-155.

Bronner, A.E. (1991). Recente ontwikkelingen in markt- en opinieonderzoek [Recent developments in opinion and marketing research]. In *Aspecten van onderzoek; Theorie, variabelen en praktijk* (pp. 63-80). Utrecht: University of Utrecht (Available from ISOR, Rijksuniversiteit Utrecht, POB 80140, 3508 TC Utrecht).

Burt, R.S., Fischer, M.G., & Christman, K.P. (1979). Structures of well-being; sufficient conditions for identification as restricted covariance models. *Sociological Methods and Research, 8*, 111-120.

Burt, R.S., Wiley, J.A., Minor, M.J., & Murray, J.R. (1978). Structure of well-being; Form, content, and stability over time. *Sociological Methods and Research, 6*, 365-407.

Cannell, C.F., & Fowler, F.J. (1963). Comparison of a self- enumerated procedure and a personal interview: A validity study. *Public Opinion Quarterly, 27*, 250-264.

Cannell, C.F., Miller, P.V., & Oksenberg, L. (1981). Research on interviewing techniques. In: S. Leinhardt (Ed.), *Sociological Methodology* (pp. 389-437). San Francisco: Jossey-Bass.

Cantril, H. (1965). *The pattern of human concerns.* New Brunswick: Rutgers University Press.

Carroll, L. (1976). *The annotated Alice; Alice's adventures in wonderland and through the looking glass* (illustrated by John Teniel, with an introduction and notes by Martin Gardner). Harmondsworth: Penguin books.

CBS (1988). *Bevolking der gemeenten van Nederland op 1 januari 1988* [Population of Dutch municipalities 1988; A publication of the Netherlands Central Bureau of Statistics]. The Hague: Staatsuitgeverij.

CBS (1990). *Statistisch jaarboek 1990* [Statistical yearbook 1990; A publication of the Netherlands Central Bureau of Statistics]. The Hague: Staatsuitgeverij.

Cochran, W.G. (1977). *Sampling techniques.* New York: Wiley.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Collins, M., Sykes, W., Wilson, P., & Blackshaw, N. (1988). Nonresponse: The UK experience. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 213-231). New York: Wiley.

Conan Doyle, A. (1981). The copper beeches. In: *The adventures of Sherlock Holmes* (pp. 260-285). London: Penguin Books.

Cook, T.D., & Campbell, D.T. (1979). *Quasi experimentation: Design and analysis issues for field studies.* Chicago: Rand McNally.

Couch, A. & Keniston, K. (1960). Yeahsayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60*, 151-174.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

De Bie, S.E., & Dijkstra, W. (1989), *Interviewen cursusboek* [Interview Manual, Society of Research Centers]. Leiden: Vereniging van Onderzoek Instituten.

De Bock, H. (1987). Technologische innovatie in sociaal- wetenschappelijk onderzoek: maatschappelijke randvoorwaarden [Technological innovations in social sciences research]. In J. de Jong-Gierveld, & J. van der Zouwen (Eds.), *De vragenlijst in het sociaal onderzoek* (pp. 49-64). Deventer: Van Loghum Slaterus.

De Greef, P., Breuker, J., & Wielinga, B. (1988). Kennisverwerving voor het bouwen van expertsystemen [Knowledge acquisition for the construction of expertsystems]. In J.J. Hox, & G. de Zeeuw (Eds.), *De microcomputer in sociaal- wetenschappelijk onderzoek* (pp. 115-137). Amsterdam/Lisse: Swets & Zeitlinger.

De Groot, A.D. and Van Naerssen, R.F. (1969). *Studietoetsen: construeren, afnemen, analyseren [The construction and analysis of tests].* The Hague: Mouton.

De Heer, W.F., Akkerboom, J.C., & Israëls, A.Z. (1990). *Ideas for nonresponse investigations; contribution to the nonresponse workshop.* Voorburg: CBS Netherlands Central Bureau of Statistics.

De Heer, W.F., & Israëls, A.Z. (1990). *Verslag van de "Workshop on household survey nonresponse" in Stockholm [Report on the Stocholm workshop on household survey nonresponse]* (unpublished memo). Voorburg: CBS Netherlands Central Bureau of Statistics.

De Jong-Gierveld (1987). Developing and testing a model of loneliness. *Journal of Personality and Social Psychology, 53*, 119-128.

De Jong-Gierveld, J., & Kamphuis, F. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement, 9*, 289-299.

De Leeuw, E.D. (1991). *The influence of data collection procedure on psychometric reliability and scaling properties.* (Response effects in Surveys, Technical report No 5). Amsterdam: Vrije Universiteit, Department of Social Research Methodology.

De Leeuw, E.D. (1991). *Een vergelijking van de datakwaliteit bij gegevens verkregen met een postenquête, een telefonisch interview, en een face to face interview; De gebruikte vragenlijsten* [Data quality in mail, telephone, and face to face surveys; The questionnaires] (Response effects in Surveys, Technical report No 6). Amsterdam: Vrije Universiteit, Department of Social Research Methodology.

De Leeuw, E.D., & Hox, J.J. (1988). Response stimulating factors in mail surveys. *Journal of Official Statistics, 4*, 241-249.

De Leeuw, E.D., & Hox, J.J. (1989a). *Telefonisch interviewen; Veldgids voor interviewers* [Telephone interviewing; A field guide] (Methods & Statistics Series No 43). Amsterdam: University of Amsterdam, Department of Education

De Leeuw, E.D., & Hox, J.J. (1989b). *Interviewen in een face to face situatie; Veldgids voor interviewrs* [Face to face interviewing; A field guide] (Methods & Statistics Series No 44). Amsterdam: University of Amsterdam, Department of Education

De Leeuw, E.D. & Hox, J.J. (forthcoming). Mode effects in structural modeling; A Lisrel multi-group comparison of mail, telephone, and face to face survey data.

De Leeuw, E.D., & Van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283-299). New York: Wiley.

De Maio, T.J. (1984). Social desirability and survey measurement: A review. In Ch.F. Turner & M.E. Martin (Eds.), *Surveying subjective phenomena, vol 2* (pp. 257-282). New York: Russell Sage Foundation

Deming, W.E. (1944). On errors in surveys. *American Sociological Review, 9*, 359-369.

Dijkstra, W. (1983). How interviewer variance can bias the results of research on interviewer effects. *Quality and Quantity, 17*, 179-187.

Dijkstra, W., & Van der Zouwen, J. (1977). Testing auxiliary hypothesis behind the interview. *Annals of System Research, 6*, 49-63.

Dillman, D.A. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley.

Dillman, D.A. (1991). The design and administration of mail surveys. *Annual Review of Sociology, 17*, 225-249.

Dillman, D.A. (1992). Recent advances in survey data collection methods and their implications for meeting rural data needs. In: R. Buse & J. Driscoll (Eds.), *New directions in data and information systems.* Ames: Iowa state university press.

Dillman, D.A., & Mason, R.G. (1984). *The influence of survey method on question response.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Delavan, Wisconsin.

Dillman, D.A., & Tarnai, J. (1988). Administrative issues in mixed mode surveys. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 509-528). New York: Wiley.

Dykstra, P.A. (1990). *Next of (non)kin: The importance of primary relationships for older adults' well-being.* Amsterdam/Lisse: Swets & Zeitlinger.

Dykstra, P.A. (forthcoming). Alternative for the absence of a partner: the presence of supportive relationships and the desire for independence as factors that serve to mitigate loneliness. *Ageing and Society.*

Ellis, A. (1947): Questionnaire versus interview methods in the study of human love relationships. *American Sociological Review, 12*, 541-553.

Feldt, L.S. (1969). A Test of the Hypothesis that Cronbach's Alpha or Kuder-Richardson Coefficient Twenty is the Same for Two Tests. *Psychometrika, 34*, 363-373.

Fienberg, S.E. (1978). *The analysis of cross-classified categorical data.* Cambridge: MIT Press.

Fodor, J.A. (1981). *Representations. Philosophical essays on the foundation of cognitive science.* Brighton: Harvester Press.

Forsyth, B.H., & Lessler, J.T. (1991). Cognitive laboratory methods; A taxonomy. In: P.P. Biemer, R.M. Groves. L.E. Lyberg, N.A. Mathiowetz, & S.Sudman (eds). *Measurement errors in surveys* (pp. 393-418). New York: Wiley.

Fowler, F.J., Jr. (1991). Reducing interviewer-related error through interviewer training, supervision and other means. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 259-278). New York: Wiley.

Frey, J.M. (1983). *Survey research by telephone.* Beverly Hills: Sage.

Galtung, J. (1967). *Theory and methods of social research.* London: Allen.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chicester: Wiley.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research.* Beverly Hills: Sage.

Gouaux, Ch. (1971). Induced affective states and interpersonal attraction. *Journal of Personal and Social Psychology, 20,* 37-43.

Goyder, J. (1982). Further evidence on factors affecting response rates to mailed questionnaires. *American Sociological Review, 47,* 550-553.

Goyder, J. (1987). *The silent minority; Nonrespondents on sample surveys.* Cambridge: Policy Press.

Groves, R.M. (1978). On the mode of administering a questionnaire and responses to open-ended items. *Social Science Research, 7,* 257-271.

Groves. R.M. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly, 43,* 190- 205.

Groves, R.M. (1989). *Survey errors and survey costs.* New York: Wiley.

Groves, R.M., Bierner, P.P., Lyberg, L.E., Massey, J.T., Nicholls, W.L. II, & Waksberg, J. (Eds.). (1988). *Telephone survey methodology.* New York: Wiley.

Groves, R.M., & Kahn, R.L. (1979). *Surveys by telephone.* New York: Academic Press.

Groves, R.M., & Lyberg, L.E. (1988). An overview of nonresponse issues in telephone surveys. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 191-211). New York: Wiley.

Groves, R.M., & Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly, 50,* 251-266.

Groves, R.M., & Nicholls, W.L. II. (1986). The status of computer assisted telephone interviewing: Part II - Data quality issues. *Journal of Official Statistics, 2,* 117-134.

Hakstian, A.R. and Whalen, T.E. (1976). A K-sample Significance Test for Independent Alpha Coefficients. *Psychometrika, 41,* 219- 231.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

Heberlein, T.A., & Baumgartner, R.M. (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review, 43,* 447-462.

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

Herman, M.B. (1977). Mixed-mode data collection: Telephone and personal interviewing. *Journal of Applied Psychology, 62,* 399- 404.

Hippler, H-J., & Schwarz, N. (1992). *The impact of administration modes on response effects in surveys* (ZUMA-Arbeitsbericht Nr. 92/14). Mannheim: ZUMA.

Hochstim, J. R. (1967): A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62,* 976-989.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65-70.

Hox, J.J. (1986). *Het gebruik van hulptheoriën bij operationaliseren* [Using auxiliary theories for operationalization; A study of the construct of subjective well- being] (doctoral dissertation). Amsterdam: University of Amsterdam, Department of Education.

Hox, J.J. (1992). Modeling interviewer effects with multilevel models. *Kwantitatieve Methoden* (in press).

Hox, J.J., De Bie, S.E., & De Leeuw, E.D. (1990). Computer assisted (telephone) interviewing: A review. In: J. Gladitz & K.G. Troitzsch (Eds.), *Computer aided sociological research* (pp. 305-317). Berlin: Akademie-Verlag.

Hox, J.J., & De Jong-Gierveld, J. (1990). *Operationalization and research strategy.* Amsterdam/Lisse: Swets & Zeitlinger.

Hox, J.J., De Leeuw, E.D., & Kreft, I.G.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In: P.P. Biemer, R.M. Groves. L.E. Lyberg, N.A. Mathiowetz, & S.Sudman (eds). *Measurement errors in surveys* (pp. 439-461). New York: Wiley.

Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis.* Beverly Hills: Sage.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). *Meta-analysis: Cumulating research findings across studies.* New York: Sage.

Jackson, G.B. (1980). Methods for integrative reviews. *Review of Educational Research, 50,* 428-460.

Jordan, L. A., Marcus, A. C. and Reeder, L. G. (1980): Response styles in telephone and household interviewing: A field experiment. *Public Opinion Quarterly, 44,* 210-222.

Jöreskog, K.G., & Sörbom, D. (1989). *Lisrel 7; A guide to the program and applications* (second edition). Chicago: SPSS Inc.

Kalfs, N., & Saris, W.E. (1991). Mode effects in time diary research. *Kwantitatieve Methoden, 37,* 65-86.

Kalton, G., Kasprzyk, D., & McMillen, D.B. (1989). Nonsampling errors in panel surveys. In D. Kasprzyk, G.J. Duncan, G. Kalton, & M.P. Singh (Eds). *Panel surveys* (pp. 249-270). New York: Wiley.

Kahn, R.L., & Cannell, C.F. (1957). *The dynamics of interviewing.* New York: Wiley.

Kendall, M.G. (1959). Hiawatha designs an experiment. *American Statistician, 1959, 13,* 23-24.

Kerssemakers, F.A.M. (1985). Telefonisch enquêteren [Telephone interviewing]. In *CBS-select 3* (pp. 211-230). Voorburg/Heerlen: Centraal Bureau voor de Statistiek.

Kerssemakers, F.A.M., De Mast, F.A.C., & Remmerswaal, P.W.M. (1987). Computer assisted telephone interviewing, some response findings. In *CBS-select 4* (pp. 119-131). Voorburg/Heerlen: CBS Netherlands Central Bureau of Statistics.

Kidd, A. (1986). *Knowledge elicitation for expertsystems: A practical handbook.* New York: Plenum Press.

Kirk, R.E. (1968). *Experimental Design: Procedures for the Behavioral Sciences.* Belmont: Wadsworth Company.

Kish, L. (1949). A procedure for objective respondent selection in the household. *Journal of the American Statistical Association, 44,* 380-387.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association, 57,* 92-115.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

Kish, L. (1987). *Statistical design for research.* New York: Wiley.

Kogut, J. (1986). *A review of IRT-based indices for detecting and diagnosing aberrant response patterns* (Report No 86-4). Enschede: Toegepaste Onderwijskunde, Universiteit van Twente.

Körmendi, E. (1988): The quality of income information in telephone and face to face surveys. In R. M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 341-356). New York: Wiley.

Körmendi, E., & Noordhoek, J. (1989). *Data quality and telephone interviews.* Copenhagen: Danish Statistical Office (Danmarks Statistik).

Krosnick, J.A., & Alwin, D.F. (1987). An evaluation of cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51,* 201-219.

Kruskall, W. (1991). Introduction. In: P.P. Biemer, R.M. Groves. L.E. Lyberg, N.A. Mathiowetz, & S.Sudman (eds). *Measurement errors in surveys* (pp. xxiii-xxxiii). New York: Wiley.

Kviz, F.J. (1977). Towards a standard definition of response rate. *Public Opinion Quarterly, 41,* 265-267.

Lavrakas, P.J.(1987). *Telephone survey methods; Sampling, selection and supervision.* Beverly Hills: Sage.

Lepkowski, J.M. (1988). Telephone sampling methods in the United States. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 73-98). New York: Wiley.

Lévy-Leblond, J-M. (1990, January). Une recherche qui se fait comme elle se parle..[Research evolves while talking to each other..]. *Le Monde Diplomatique; Supll. Langues et Science,*, pp. 25-26.

Light, R.J. & Pillemer, D.B. (1984). *Summing up; The science of reviewing research.* Cambridge Ma: Harvard University Press.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Marascuilo, L.A. (1966). Large-sample multiple comparisons. *Psychological Bulletin, 65,* 280-290.

Lyberg, L., & Kasprzyk, D. (1991). Data collection methods and measurement errors: An overview. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 237-258). New York: Wiley.

Martin, J., & O'Muircheartaigh, C. (1991). *The use of CAPI for attitude surveys: An experimental comparison with traditional methods* (Working paper series No. 8). London: Joint Centre for Survey Methods.

McClendon, M.J. (1991). Acquiescence and recency response order effects in interview surveys. *Sociological Methods and Research, 20,* 60-103.

Meijer, R.R. (1990). *Detecting and diagnosing aberrant response patterns within the context of nonparametric IRT and by means of group based indices.* Unpublished manuscript, Vrije Universiteit, Department of Industrial and Organizational Psychology, Amsterdam.

Meijer, R.R., & De Leeuw, E.D. (1992). *Person fit indices in survey research; A mode comparison on the "De Jong-Gierveld loneliness scale"* (Response effects in surveys, Report No 7).

Amsterdam: Vrije Universiteit.

Meijer, R.R., Sijtsma, K., & Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283-298.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton.

Mokken, R.J., & Lewis, C. (1982). A nonparametric approach to the analysisof dichotomous item responses. *Applied Psychological Measurement, 6,*417-430.

Molenaar, I.W. (1982). Mokken scaling revisited. *Kantitatieve Methoden, 3,* 145-164.

Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75-106.

Nicholls, W.L. II, & Groves, R.M. (1986). The status of computer assisted telephone interviewing: Part I - Introduction and impact on cost and timeliness of survey data. *Journal of Official Statistics, 2,* 93-115.

Nuckols, R.C. (1964). Personal interview versus mail panel survey. *Journal of Marketing Research, 1,* 11-16.

Nunnally, J.C. (1967). *Psychometric theory.* New York: McGraw-Hill.

O'Muircheartaigh, C.A. (1977). Response error. In C.A. O'Muircheartaigh, & C. Payne (Eds.), *The analysis of survey data* (pp. 193-239) London: Wiley.

O'Toole, B.I., Battistutta, D., Long, A., & Crouch, K. (1986). A comparison of costs and data quality of three health survey methods: Mail, telephone and personal home interview. *American Journal of Epidemology, 124,* 317-328.

Oldendick, R.W., Bishop, G.F., Sorenson, S.B., & Tuchfarber, A.J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics, 4,* 307-318.

PTT [PTT Telecom Netherlands] (1989). *De maatschappij verandert, PTT verandert mee* [Changes in society, changes in telecommunication in the Netherlands]. Den Haag: PTT

Rogers, T.F. (1976). Interviews by telephone and in person: Quality of response and field performance. *Public Opinion Quarterly, 40,* 51-65.

Rorer, L.G. (1965). The great response style myth. *Psychological Bulletin, 63*, 129-156.
Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills: Sage.
Rosenthal, R., & Rubin, D.B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.
Rossi, P.H., Wright, J.D., & Anderson, A.B. (1983). Sample surveys: History, current practice, and future prospects. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), *Handbook of survey research* (pp. 1-20). San Diego: Academic Press.
Salmon, C.T., & Nichols, J.S. (1983). The next birthday method for respondent selection. *Public Opinion Quarterly, 47*, 270-276.
Saris, W.E. (1988). *Variation in response functions: A source of measurement error*. Amsterdam: Sociometric Research Foundation.
Saris, W.E. (1989). A technological revolution in data collection. *Quality and Quantity, 23*, 33-349.
Saris, W.E. (1991). *Computer assisted interviewing* (Quantitative applications in the social sciences, No 80). Newbury Park: Sage.
Sayers, D.L. (1975). *Murder must advertise*. London: New English Library.
Schaeffer, N.C. (1991). Conversation with a purpose-or conversations? Interaction in the standardized interview. In: P.P. Biemer, R.M. Groves. L.E. Lyberg, N.A. Mathiowetz, & S.Sudman (eds). *Measurement errors in surveys* (pp. 367-391). New York: Wiley.
Schuman, H., & Presser, S., (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
Schwarz, N., Strack, F., Hippler, H.J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Aplied Cognitive Psychology, 5*, 193-212.
Scott, W.A. (1968). Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology, second edition, Vol. 2* (pp. 204-273). Reading: Addison-Wesley.
SCP [Social and Cultural Planning Office] (1988). *Sociaal en cultureel rapport 1988* [Social and cultural report 1988]. Alphen aan de Rijn: Samson.
Siemiatycki, J. (1979). A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health, 69*, 238-245.
Sigelman, L. (1982): The uncooperative interviewie. *Quality and Quantity, 16*, 345-353.
Sijtsma, K. (1988). *Contributions to Mokken's Nonparametric Item Response Theory*. Amsterdam: Free University Press.
Sijtsma, K., & Molenaar, I.W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*, 79-97.
Sikkel, D. (1988). *Quality aspects of statistical data collection*. Amsterdam: The Sociometric Research Foundation.
Smith, T.W. (1987). The art of asking questions, 1936-1985. *Public Opinion Quarterly, 51*, S95-S108.
Snijkers, G.J.M.E. (1992). Computer gestuurd enquêteren: Telefonisch of persoonlijk? [Computer assisted interviewing: By telephone or in person?]. *Kwantitatieve Methoden, 39*, 53-69.
S.R.C. (1976). *Interviewer's manual; Revised edition*. Ann Arbor: University of Michigan, Survey Research Center and Institute for Social Research.
Steeh, C.G. (1981). Trends in nonresponse rates, 1952-1979. *Public Opinion Quarterly, 45*. As reprinted in E. Singer, & S. Presser (1989), Survey Research Methods, A reader. Chicago: university of Chicago Press.
Strack, F., & Martin, L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In: H-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methodology* (pp. 123-148). New York: Springer Verlag.
Sudman, S., & Bradburn, N.M. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.
Sudman, S., & Bradburn, N.M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey- Bass.
Sugiyama, M. (1992). Responses and non-responses. In: L. Lebart (Ed.) *Quality of information in sample surveys* (pp. 227-239). Paris: Dunod.

Sykes, W., & Collins, M. (1988). Effects of mode of interview: Experiments in the U.K. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 301-320). New York: Wiley.

Sykes, W., & Hoinville, G. (1985). *Telephone interviewing on a survey of social attitudes: A comparison with face-to-face procedures* (SCPR Survey Research Publication). London: Social And Community Planning Research.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982). Detection of aberrant response patterns. *Journal of Educational Statistics, 7*, 215-231.

Tarnai, J.; & Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail vs. telephone surveys. In N. Schwarz & S. Sudman, *Order effects in social and psychological research* (pp. 115-129). New York: Springer Verlag.

Thornberry, O. Jr., Nicholls, W.L. II, & Kulpinsky, S. (1982). Data collection methods in federal statistical surveys. *American Statistical Association, 1982 Proceedings of the section on survey research methods*, 185-190.

Trewin, D., & Lee, G. (1988). International comparisons of telephone coverage. In R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 9-24). New York: Wiley.

Tufte, E.R. (1991). *Envisioning information*. Cheshire: Graphic Press.

Tull, D.S., & Hawkins, D.I. (1984). *Marketing research: Measurement and method*. New York: McMillan.

Van Bastelaar, A.M.L., Kerssemakers, F.A.M., & Sikkel, D. (1987). A test of The Netherlands Continuous Labour Force survey with hand-held computers; interviewer behaviour and data quality. In *CBS-select 4* (pp. 37-54). Voorburg/Heerlen: CBS Netherlands Central Bureau of Statistics.

Van de Geer, J.P. (1985). *Homals* (Report UG-85-02). Leyden: University of Leyden, Department of Data Theory.

Van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]*. Lisse: Swets & Zeitlinger.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

Van der Zouwen, J., Dijkstra, W., & Smit, J.H. (1991). Studying respondent-interviewer interaction: The relationship between interviewing style, interviewer behavior, and response behavior. In: P.P. Biemer, R.M. Groves. L.E. Lyberg, N.A. Mathiowetz, & S.Sudman (eds). *Measurement errors in surveys* (pp. 419-437). New York: Wiley.

Van Rooy, C. (1987). Responsvoorspellingen: Toverformules of realisme? [Response prediction: A magic formula or realism?]. In A.E. Bronner (Ed). *Jaarboek van de vereniging van marktonderzoekers 86-87* (pp. 36-41). Haarlem: De Vrieseborch.

Van Sonsbeek, J.L.A., & Stronkhorst, L.H. (1983). *Vergelijking van drie waarnemingsvarianten bij de meting van medische consumptie* [A comparison of data collection methods in the measurement of medical consumption]. Den Haag: Centraal Bureau voor de Statistiek.

Van Tilburg, T.G. and De Leeuw, E.D. (1991). Stability of scale quality under various data collection procedures: A mode comparison of the 'De Jong-Gierveld loneliness scale.' *International Journal of Public Opinion Research*, 3, 69-85.

Walberg, H.J., & Haertel, E.H. (1980). Research integration: An introduction and overview. *Evaluation in Education, 4*: 5-10.

Waterton, J.J. (1984). Reporting alcohol consumption: The problem of response validity. *American Statistical Association, 1984 Proceedings of the section on survey research methods*, 664-669.

Willis, G.B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology, 5*, 251-268.

Wiseman, F. (1972): Methodological bias in public opinion surveys. *Public Opinion Quarterly, 36*, 105-108.

Wolf, F.M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills: Sage.

Wortman, P.M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology, 34*, 223-260.
Wortman, P.M., & Bryant, F.B. (1985). School desegregation and black achievement; An integrative review. *Sociological Methods and Research, 13*, 289-324.
Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: Mesa Press.
Zoon, C. (1992, April 18). Dial C for chaos. *Volkskrant*, p. 2.

AUTHOR INDEX

TOPIC INDEX

Asking questions off respondents is one of the main data collection methods in social science and its associated applied fields. The oldest survey methods are the face to face interview and the mail questionnaire. After 1970, telephone interviews have become increasingly popular. A new trend is mixed mode surveys; surveys that combine more than one data collection mode within one study.

One of the most important questions for both survey researchers and for consumers of survey research is whether the data collected by one method differ from the data collected by another. This book compares three major modes of survey research: face to face interviews, telephone interviews, and mail questionnaires. After a theoretical discussion why mode effects may occur, the book presents a comprehensive overview based on a meta-analysis of the research literature. This is followed by the results of a controlled field experiment. The analysis goes beyond the usual reports of univariate differences between the methods, by testing the psychometric properties of scales and the results of multivariate models for mode effects.

The combination of an incisive analysis of issues in survey methodology with sophisticated data analysis techniques gives this book a broad scope. It will be of interest to both social science methodologists and people who work in theoretical or applied social research.

'T-Publikaties

ISBN 90-801073-1-X NUGI 659

Stellingen behorende bij het proefschrift van Edith D. de Leeuw, "Data quality in mail, telephone and face to face surveys".

1. Het verdedigen van een proefschrift komt neer op het ingooien van de eigen glazen: Het grootste aantal academische vacatures is voor AIO's en OIO's.

2. De universele dataverzamelingsmethode bestaat niet: Geen van de gebruikelijke dataverzamelingsmethoden is onder alle omstandigheden superieur.

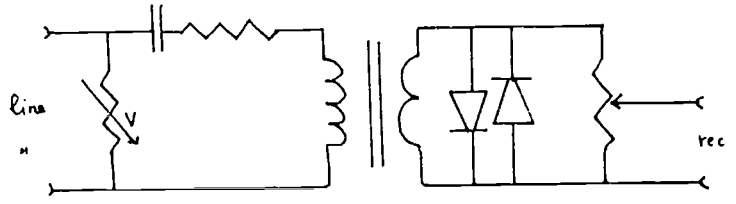
3. Het belangrijkste onderscheid tussen verschillende vormen van dataverzameling is het onderscheid in vormen met en vormen zonder interviewer.

4. Bij mixed-mode surveys verdient het aanbeveling om 'methode van dataverzameling' expliciet als variabele in het statistische model op te nemen.

5. Meta-analyse is een inductieve procedure en alle bezwaren die Popper tegen het inductivisme heeft ingebracht zijn dan ook van toepassing op de meta-analyse. Daarom is één enkele studie met totaal onverwachte resultaten interessanter dan een meta-analytische samenvatting van alle eerdere studies.

6. Bij meerdere significantietoetsingen is voor het corrigeren van het globale alphaniveau de multiplicatieve procedure van Holm verre te verkiezen boven de meer gebruikelijke additieve Bonferroni-correctie.

7. De kwaliteit van de data bij telefonisch interviewen kan door zeer eenvoudige technische middelen worden verbeterd. Ter illustratie:



(© K. ten Hoeve)

8. Auteurs en redacteuren dienen bij beslissingen over de gewenste mate van detaillering in publikaties rekening te houden met de mogelijkheid dat de betreffende publikatie in een meta-analyse opgenomen kan worden.

9. De Amsterdamse VVV maakt aan buitenlandse toeristen onvoldoende duidelijk dat er een verschil is tussen voet- en fietspad.

10. De bevinding van Maarten 't Hart dat ratten met smaak zeep eten, is niet onafhankelijk repliceerbaar. Gezien het *proefleider-verwachtingseffect* roept dit vragen op over de eetgewoonten van 't Hart. (cf. Joh. Hoogstraten, *De machteloze onderzoeker*, Meppel, Boom, 1979; Maarten 't Hart, *Ratten*, Amsterdam, Wetenschappelijke uitgeverij, 1980).

11. Het huidige academische rangenstelsel lost in ieder geval een probleem op voor lezers van de Bommel-saga: Herr Pieps is een AIO.

12. Met het verdwijnen van Jool Hul is één van de redenen verdwenen waarom deze tijd een VU nodig heeft.