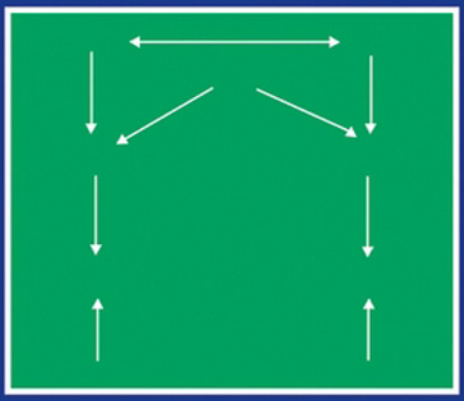
WILEY

Design, Evaluation, and Analysis of Questionnaires for Survey Research



Willem E. Saris
Irmtraud N. Gallhofer

WWW. LINK AVAILABLE

WILEY SERIES IN SURVEY METHODOLOGY

DESIGN, EVALUATION, AND ANALYSIS OF QUESTIONNAIRES FOR SURVEY RESEARCH

WILLEM E. SARIS
ESADE
Universitat Ramon Llull
Barcelona, Spain

IRMTRAUD N. GALLHOFER
ESADE
Universitat Ramon Llull
Barcelona, Spain

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

This Page Intentionally Left Blank

DESIGN, EVALUATION, AND ANALYSIS OF QUESTIONNAIRES FOR SURVEY RESEARCH

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

DESIGN, EVALUATION, AND ANALYSIS OF QUESTIONNAIRES FOR SURVEY RESEARCH

WILLEM E. SARIS
ESADE
Universitat Ramon Llull
Barcelona, Spain

IRMTRAUD N. GALLHOFER
ESADE
Universitat Ramon Llull
Barcelona, Spain

WILEY-INTERSCIENCE
A John Wiley & Sons, Inc., Publication

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:

Saris, Willem E.
Design, evaulation, and analysis of questionnaires for survey research /
Willem E. Saris, Irmtraud N. Gallhofer.
p. cm. — (Wiley series in survey methodology)
Includes bibliographical references.
ISBN 978-0-470-11495-7 (cloth)
1. Social surveys. 2. Social surveys—Methodology. 3. Questionnaires. 4. Interviewing. I. Gallhofer, Irmtraud N. II. Title.
HN29.S29 2007
300.72'3—dc22 2007001697

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Preface

Designing a survey involves many more decisions than most researchers realize. Survey specialists, therefore, speak of the art of designing survey questions (Payne 1951). However, this book introduces methods and procedures that can make questionnaire design a scientific activity. This requires knowledge of the consequences of the many decisions that researchers take in survey design and how these decisions affect the quality of the questions.

It is desirable to be able to evaluate the quality of the candidate questions of the questionnaire before collecting the data. However, it is very tedious to manually evaluate each question separately on all characteristics, mentioned in the scientific literature, that predict the quality of the questions. It may even be said that it is impossible to evaluate the effect of the combination of all these characteristics. This requires special tools that did not exist to date. A computer program that can evaluate all the questions of a questionnaire on a number of characteristics and provide an estimate of the quality of the questions based on the coded question characteristics would be very helpful. This program could be a tool for the designer of the survey who can determine, on the basis of the computer output, which questions in the survey require further study in order to improve the quality of the data collected.

Furthermore, after a survey is completed it is useful to have information about the data quality collected in order to correct for the errors in the data. Therefore, there is a need for a computer program that can evaluate all questions of a questionnaire on a number of characteristics and provide an estimate of the quality of the questions. Such information can be used to improve the quality of the data analysis.

In order to further such an approach, we have

1. Developed a system for coding characteristics of survey questions and the more general survey procedure
2. Assembled a large set of studies that used multitrait-multimethod (MTMM) experiments to estimate the reliability and validity of questions

3. Carried out a meta-analysis that relates these question characteristics to the reliability and validity estimates of the questions
4. Developed a semiautomatic program that predicts the validity and reliability of new questions based on the information available from the meta-analysis of MTMM experiments

We think that these four steps are necessary to change the development of questionnaires from an "art" into a scientific activity.

While this approach helps to optimize the formulation of a single question, it does not necessarily improve the quality of survey measures. Often researchers use complex concepts in research that cannot be measured by a single question. So, several indicators are used. Moving from complex concepts to a set of questions that together may provide a good measure for the concept is called operationalization. In order to develop a scientific approach for questionnaire design we have also provided suggestions for *operationalization* of complex concepts.

The purpose of this book is, first, to specify a three-step procedure which will generate questions to measure the complex concept defined by the researcher. The approach of operationalization is discussed in Part I of this book.

The second purpose of the book is to introduce to survey researchers the different choices researchers can make and are making while designing survey questionnaires. This topic is covered in Part II of this book.

Part III of this book discusses quality criteria for survey questions, the way these quality criteria have been evaluated in experimental research and the results of a meta-analysis over many of such experiments that allow researchers to determine the size of the effects of the different decisions on the quality of the questions.

Part IV indicates how all this information can be used efficiently in the design and analysis of surveys. Therefore, the first chapter introduces the program "survey quality predictor" (SQP), which can be used for the prediction of the quality of survey items on the basis of cumulative information concerning the effect of different characteristics of the different components of survey items on the data quality. The discussion of the program will be specific enough so that the reader can use it to improve his/her own questionnaires.

The information about data quality can and should also be used after a survey has been completed. Measurement error is unavoidable, and this information is useful for how to correct it. The exact mechanics of it are illustrated in several chapters of Part IV. We start out by demonstrating how this information can be applied to estimate the quality of measures of complex concepts, followed by a discussion on how to correct for measurement error in survey research and how to cope with measurement error in cross cultural research. In the last chapter we discuss how one can cope with measurement error in cross-cultural research.

In general, we hope to contribute to the scientific approach of questionnaire design and the overall improvement of survey research with this book.

Before closing this preface we have to mention the important contribution of Frank Andrews, who suggested this approach in the early 1990s and inspired us to continue his important work. The data sets of his studies are included in the database on which the SQP program is based. A very important contribution to this book has also been made by Annette Scherpenzeel and Richard Költringer, who designed and performed many of the experiments on the basis of which we can now make more general statements. We would also like to thank Albert Satorra for his helpful discussions about statistical problems. An important contribution in the development of this approach has been the cooperation over the years with the members of the International Research group of Methodology and Comparative Surveys (IRMCS) that came together on a yearly basis in Slovenia, among other places to discuss the latest developments. We especially have to mention Anuska Ferligoj, who organized most of the meetings. We are also very grateful to William van der Veld and Daniel Oberski, who made the first versions of the SQP program. Important advice concerning sentence grammar in Chapter 2 of the book has been obtained from Dr. P.J. van der Voort. We are very grateful for his help. In the last phase of producing the book we got a lot of helpful comment from Christine Punzo (Wiley) and without the help of our type setter Kjeld de Ruyter of Puntspatie we would not have been able to produce this book.

In addition we like also to thank the Netherlands Organization for Scientific Research (NWO) for its financial support not only for the data collection but also for the development of the programs. A very important role was also played by Sophia Kusyk, who was able to transform some of our awkward English phrases into proper ones.

Last but not least, we would like to thank the many students who have commented on the different versions of this book and the program. Without their stimulating support and criticism, this book would not have been made.

Willem E. Saris
Irmtraud Gallhofer

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Introduction

In order to emphasize the importance of survey research for the social, economic and behavioral fields, we have elaborated on a study done by Stanley Presser, originally published in 1984. In this study Presser performed an analysis of papers published in the most prestigious journals within the scientific disciplines of economics, sociology, political science, social psychology, and public opinion (or communication) research. His aim was to investigate to what extent these papers were based on data collected in surveys.

Presser did his study by coding the data collection procedures used in the papers that appeared in the following journals. For the economics field he used the *American Economic Review*, the *Journal of Political Economy*, and the *Review of Economics and Statistics*. To represent the sociology field he used the *American Sociological Review*, the *American Journal of Sociology* and *Social Forces*; and for the political sciences, the *American Journal of Political Science*, the *American Political Science Review*, and the *Journal of Politics*. For the field of social psychology he chose the *Journal of Personality and Social Psychology* (a journal that alone contains as many papers as each of the other sciences taken together). Finally, for public opinion research the *Public Opinion Quarterly* was elected. For each selected journal all papers published in the years 1949–1950, 1964–1965, and 1979–1980 were analyzed.

We have updated Presser's analysis of the same journals for the period of 1994–1995, a period that is consistent with the interval of 15 years to the preceding measurement. Presser (1984: 95) suggested using the following definition of a survey:

> ..any data collection operation that gathers information from human respondents by means of a standardized questionnaire in which the interest is in aggregates rather than particular individuals. (...) Operations conducted as an integral part of laboratory experiments are not included as surveys, since it seems useful to distinguish between the two methodologies. The definition is silent, however, about the method of respondent selection and the mode of data collection. Thus, convenience samples as well as census, self-administered questionnaires as well as face-to-face interviews, may count as surveys.

The results obtained by Presser, and completed by us for the years 1994–1995, are presented in Table I.1. For completing the data we stayed consistent with the procedure used by Presser except in one point: we did not automatically subsume studies performed by organizations for official statistics (statistical bureaus) under the category "surveys." Our reason was that at least part of the data collected by statistical bureaus is based on administrative records and not collected by survey research as defined by Presser. Therefore, it is difficult to decide on the basis of the description of the data in the papers whether surveys have been used. For this reason we have not automatically placed this set of papers, based on studies by statistical bureaus, in the class of survey research.

The difference in treating studies from statistical bureaus is reflected in the last column of Table I.1, relating to the years 1994–1995. We first present (within parentheses) the percentage of studies using survey methods based on samples (our own classification). Next, we present the percentages that would be obtained if all studies conducted by statistical bureaus were automatically subsumed under the category survey (Presser's approach).

Table I.1: Percentage of articles using survey data by discipline and year (number of articles excluding data from statistical offices in parentheses)

	Period				
Discipline	1949–50	1964–65	1979–80	1994–95	
Economics	5.7%	32.9%	28.7%	(20.0%)	42.3%
	(141)	(155)	(317)		(461)
Sociology	24.1%	54.8%	55.8%	(47.4%)	69.7%
	(282)	(259)	(285)		(287)
Political science	2.6%	19.4%	35.4%	(27.4%)	41.9%
	(114)	(160)	(203)		(303)
Social psychology	22.0%	14.6%	21.0%	(49.0%)	49.9%
	(59)	(233)	(377)		(347)
Public opinion	43.0%	55.7%	90.6%	(90.3%)	90.3%
	(86)	(61)	(53)		(46)

Depending on how the studies of the statistical offices are coded, the proportion of survey research has increased, or slightly decreased, over the years in economics, sociology and political science. Not surprisingly, the use of surveys in public opinion research is still very high, and stable.

Most remarkable is the increase of survey research in social psychology: the proportion of papers using survey data has more than doubled over the last 15-year interval. Surprisingly this outcome contradicts Presser's assumption that the limit of the survey research growth in the field of social psychology might already have been reached by the end of the 1970s, due to the "field's embracing

the laboratory/experimental methodology as the true path to knowledge."

Presser did not refer to any other method used in the papers he investigated, except for the experimental research of psychologists. For the papers published in 1994–1995 we, however, also categorized non-survey methods of the papers. Moreover, we checked whether any empirical data were employed in the same papers.

In economics, sociology, and political science many papers are published that are purely theoretical, that is, formulating verbal or mathematical theories or discussing methods. In economics this holds for 36% of the papers, in sociology this figure is 26%, and in political science it is 34%. In the journals representing the other disciplines, such papers have not been found for the period analyzed.

Given the large number of theoretical papers it makes sense to correct the percentages of Table I.1, by ignoring the purely theoretical papers, and considering only empirical studies. The results of this correction for 1994–95 are presented in Table I.2.

Table I.2: Use of different data collection methods in different disciplines as found in the major journals in 1994–1995 expressed in percentages with respect to the total number of empirical studies published in these years

	Disciplines				
Method	Economics	Sociology	Political science	Psychology	Public opinion
Survey	39.4	59.6	28.9	48.7	95.0
Experimental	6.0	1.7	5.4	45.6	5.0
Observational	3.2	0.6	31.9	4.1	0.0
Text analysis	6.0	4.6	7.2	0.6	0.0
Statistical data	45.4	33.5	26.6	9.0	0.0

Table I.2 shows the overwhelming importance of the survey research methodology for public opinion research, but also for sociology and even for social psychology. For social psychology the survey method is at least as important as the experimental design, while hardly any other method is employed. In economics and sociology, existing statistical data also are frequently used, but it has to be considered that these data sets themselves are often collected through survey methods.

The situation in political science in the period of 1994–1995 is somewhat different, although political scientists also use quite a number of surveys and statistical data sets based on surveys, they also make observations in many papers of the voting behavior of representatives.

We can conclude that survey research has become even more important than it was 15 years ago, as shown by Presser. All other data collection methods

are only used infrequently with the exception of what we have called "statistical data." These data are collected by statistical bureaus and are at least partially based on survey research and on administrative records. Observations, in turn, are used especially in the political sciences for researching voting behavior of different representative bodies, but hardly in any other science. The psychologists naturally use experiments, but with less frequency than was expected from previous data. In communication science, experiments are also utilized on a small scale. All in all, this study clearly demonstrates the importance of survey research for the fields of the social and behavioral sciences.

DESIGNING A SURVEY

As a survey is a rather complex procedure to obtain data for research, in this section we will briefly discuss a number of decisions a researcher has to take in order to design a survey.

1. Choice of a topic

The first choice to be made concerns the substantive research in question. There are many possibilities, depending on the state of the research in a given field what kind of research problem will be identified. Basic choices are whether one would like to do a *descriptive* or *explanatory* study and in the latter case whether one would like to do *experimental* research or *nonexperimental* research.

Survey research is often used for descriptive research. For example, in newspapers and also in scientific journals like *Public Opinion Quarterly* many studies can be found which merely give the distribution of responses of people on some specific questions such as: satisfaction with the economy, government, and functioning of the democracy. Many polls are done to determine the popularity of politicians, to name just a few examples.

On the other hand, studies can also be done to determine the reasons for the satisfaction with the government or the popularity of a politician. Such research is called *explanatory research*. The class of explanatory studies includes nonexperimental as well as experimental studies in a laboratory. Normally we classify research as *survey research* if large groups of a population are asked questions about a topic. Therefore, even though laboratory experiments employ questionnaires they are not treated as surveys in this book. However, nowadays experimental research can also be done with survey research. In particular, computer assisted data collection facilitates this kind of research by random assignment procedures (De Pijper and Saris 1986; Piazza and Sniderman 1991), and such research is included here as survey research. The difference between the two experimental designs is where the emphasis is placed, either on the data of individuals or small groups or on the data of some specified population.

2. Choice of the most important variables

The second choice is that of the variables to be measured. In the case of a descriptive study the choice is rather simple. It is directly determined by the

purpose of the study. For example, if a study is measuring the satisfaction of the population with the government, it is clear that questions should be asked about the "satisfaction with the government."

On the other hand, to study what the effects of different variables are on participation in elections, the choice is not so clear. In this case it makes sense to develop an inventory of possible causes and to develop from that list a preliminary model that indicates the relationships between the variables of interest. An example is given in Figure I.1. We suppose that two variables have a *direct effect* on "participation in elections" (voter participation): "political interest" and "the adherence to the norm that one should vote" (norm).

Furthermore we hypothesize that "age" and "education" have a direct influence on these two variables but only an *indirect effect* on "participation in elections." One may wonder why the variables age and education are necessary in such a study if they have no direct effect on "voter participation." The reason is that these variables cause a relationship between the "norm" and "voter participation" and, in turn, between "political interest" and "voter participation." Therefore, if we use the correlation between, for example, "political interest" and "voter participation" as the estimate of the effect of "political interest," we would overestimate the size of the effect because part of this relationship is a "spurious correlation" due to "age" and "education."

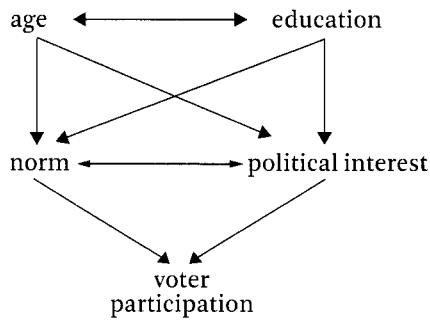


FIGURE I.1: *A model for the explanation of participation in elections by voting.*

For more details on this issue we recommend the following books on causal modeling by Blalock (1964), Duncan (1975), and Saris and Stronkhorst (1984). Therefore, in this research one not only has to introduce the variables "voter participation," "political interest," and "adherence to the norm," but also "age" and "education" as well as all other variables that generate spurious correlation between the variables of interest.

3. Choice of a data collection method

The third choice to be made concerns the data collection method. This is an important choice related to costs, question formulation and quality of data.

Several years ago the choices available were between personal interviews (face to face interviews), telephone interviews and mail surveys, all using paper questionnaires. A major difference in these methods was the presence of the interviewer in the data collection process. In personal interviews the interviewer is physically present, in telephone interviewing the interviewer is at a distance and the contact is by phone while in mail surveys the interviewer is not present at all. Nowadays each of these modes of data collection can also be computerized by computer-assisted personal interviewing (CAPI), computer- assisted telephone interviewing (CATI) and computer-assisted self interviewing (CASI) or Web surveys.

As was mentioned, these modes of data collection differ in their cost of data collection, where personal interviewing is the most expensive, telephone interviewing is less expensive, and mail interviewing is the cheapest. This holds true even with the aid of the computer. The same ordering can be specified for the response that one can expect from the respondents although different procedures have been developed to reduce the nonresponse (Dillman 2000).

Besides the abovementioned differences, there is a significant amount of literature on the variances in data quality obtained from these distinct modes of data collection. We will come back to this issue later in the book, but what should be clear is that the different modes require a corresponding formulation of the questions and due to these differences in formulation differences in responses can also be expected. Therefore, the choice of the mode of data collection is of critical importance not only for the resulting data quality but also for the formulation of the questions, which is the fourth decision to be made while designing a survey.

4. Choice of operationalization

Operationalization is the translation of the concepts to the questions. Most people who are not familiar with designing questionnaires think that making questionnaires is very simple. This is a common and serious error. To demonstrate our point, let us look at some very simple examples of questions:

I.1 Do you like football?

Most women probably answered the question: *Do you like to watch football on TV?*
Most young men will answer the question: *Do you like to play football?*
Some older men will answer the former question, some others the latter one, depending on whether they are still playing football.
This example shows that the interpretation of the question changes for the age and gender of the respondents.

Let us look at another example of a question that was frequently asked in 2003.

I.2.a Was the invasion of Iraq in 2003 a success?

In general the answer to this question is probably "yes." President Bush declared the war over in a relatively short time. But the reaction would have been quite different in 2004 if it had been asked:

I.2.b Is the invasion of Iraq in 2003 a success?

Probably the answer would be "no" for most people because after the end of the war the initial problem was not solved.

While there is only a one word difference in these questions the responses of the people would have been fundamentally different because in the first question (2a) people answer a question about the invasion, but in the second question (2b) they shift the object to evaluating the consequences of the invasion at that later point in time.

Given that such simple questions can already create a problem, survey specialists speak of "the art of asking questions" (Payne 1951; Dillman 2000: 78). We think that there is a third position on this issue: that it is possible to develop scientific methods for questionnaire design. In designing a question many decisions are made. If we know the consequences of these decisions on the quality of the responses then we can design *optimal questions* using a scientific method.

Now, let us consider some decisions that have to be made while designing a question.

Decision 1: *subject and dimension*
A researcher has to choose a subject and a dimension on which to evaluate the subject of the question. Let's expand on examples I.2a and I.2b:

I.2c Was the invasion a success?
I.2d Was the invasion justified?
I.2e Was the invasion important?

For examples I.2c–I.2e, there are many more choices possible but what is done here is that the *subject* (the invasion) has been kept the same and the *dimension* on which people have to express their answer (*concept* asked) changes. The researcher has to make the choice of the dimension or concept depending on the purpose of the study.

Decision 2: *Formulation of the question*
Many different formulations of the same question are also possible. For example:

I.2f Was the invasion a success?
I.2g Please tell me if the invasion was a success.
I.2h Now I would like to ask you whether the invasion was a success?
I.2i Do you agree with the statement that the invasion was a success?

Again, there are many more formulation choices possible, as we will show later.

Decision 3: *The response categories*
The next decision is choosing an appropriate response scale. Here again are some examples:

I.2j	*Was the invasion a success?*	*Yes/no*
I.2k	*How successful was the invasion?*	*Very much / quite / a bit / not at all*
I.2l	*How successful was the invasion?*	*Express your opinion with a number between 0 and 100 where 0 = no success at all and 100 = complete success*

Again there are many more formulation options, as we will discuss later in the book.

Decision 4: *Additional text*
Besides the question and answer categories, it is also possible to add
- An introduction
- Extra information
- Definitions
- Instructions
- A motivation to answer

It is clear that the formulation of a single question has many possibilities. Study of these decisions and their consequences on the quality of the responses will be the main topic of this book. But before we discuss this issue, we will continue with the decisions that have to be made while designing a survey study.

5. Test of the quality of the questionnaire

The next step in designing a survey study is to conduct a check of the quality of the questionnaire. Some relevant checks are
- Check on face validity
- Control of the routing in the questionnaire
- Prediction of quality of the questions with some instrument
- Use of a pilot study to test the questionnaire

It is always necessary to ask yourself and other people whether the concepts you want to measure are really measured by the way the questions are formulated. It is also necessary to control for the correctness of all routings in the questionnaire. This is especially important in computer-assisted data collection because otherwise the respondent or interviewer can be guided completely in the wrong direction, which normally leads to incomplete responses.

There are also several approaches developed to control the quality of questions. This can be done by an expert panel (Presser and Blair 1994) or on the

basis of a coding scheme (Forsyth et al. 1992 or Van der Zouwen 2000) or by using a computer program (Graesser et al. 2000a,b). Another approach that is now rather popular is to present respondents with different formulations of a survey item in a laboratory setting in order to understand the effect of wording changes (Esposito et al. 1991; Esposito and Rothgeb 1997). For an overview of the different possible cognitive approaches to the evaluation of questions, we recommend Sudman et al. (1996).

In this book we will provide our own tool namely SQP (survey quality predictor) which can be used to predict the quality of questions before they are used in practice.

6. Formulation of the final questionnaire

After corrections in the questionnaire have been made, the ideal scenario would be to test the new version again. With respect to the routing of computer assisted data collection, that is certainly the case because of the serious consequences if something is off route. Also to ensure that people actually understand a question better after correction. However, it will be clear that there is a limit to the iteration of tests and improvements.

Another issue is that the final layout of the questionnaire has to be decided on. This holds equally for both the paper-and-pencil approach as for questionnaires designed for computer-assisted data collection. However, research has only started on the effects of the layout on quality of the responses. For further analysis of the issue, see Dillman (2000).

After all these activities, the questionnaires can be printed if necessary to follow through with the data collection.

So far we have concentrated on the design of the questionnaire. There is, however, another line of work that also has to be done. This concerns the selection of a population and sampling design and organization of the fieldwork, which will be discussed in the subsequent sections.

7. Choice of population and sample design

With all survey research a decision about what *population* to report on, has to be made. One possible issue to consider is whether to report about the population of the country as a whole or about a specific subgroup. This decision is important because without it a sampling design cannot be specified. *Sampling* is a procedure to select a limited number of units from a population in order to describe this population. From this definition it is clear that a population has to be selected first.

The sampling should be done in such a way that the researcher has no influence on the selection of the respondents; otherwise the researcher can influence the results. The recommended procedure to satisfy this requirement is to select the respondents at random. Such samples based on a selection at random are called *random samples*.

If a random sampling procedure is used with a known selection probability for all respondents (not zero and not necessarily equal for all people), then it is

in principle possible to *generalize* from the sample results to the population. The precision of the statements one can make about the population depends on the design of the sample and the size of the sample.

In order to draw a sample from a population a *sampling frame*, such as a list of names and addresses of potential respondents is needed. This can be a problem for specific populations, but if such a list is missing, there are also procedures to create a sampling frame. For further details we refer to the standard literature in the area (Kalton 1983; Kish 1965; Cochran 1977). It should, however, be clear that this is a very important part of the design of the survey instrument that has to be worked out very carefully and on the basis of sufficient knowledge of the topic.

8. Decide about the fieldwork

At least as important as the design of the sample is the design of the fieldwork. This stage determines the amount of cooperation and refusals from respondents and the quality of the work of the interviewers. In order to generate an idea of the complexity of this task we provide an overview of the decisions that have to be made:

- Number of interviews for each interviewer
- Number of interviewers
- Recruitment of interviewers: where, when, how
- How much to pay: per hour/ per interview
- Instruction: kind of contacts, number of contacts, when to stop, administration
- Control procedures: interviews done/not done
- Registration of incoming forms
- Coding of forms
- Necessary staff

All these decisions are rather complex and require special attention in survey research, which are beyond the scope of this book.

9. What we know about these decisions

In his paper mentioned at the beginning of this introduction, Presser (1984) complained that, in contrast with the importance of the survey method, methodological research was directed mainly at statistical analysis, and not at the methods of data collection itself. That his observation still holds, can be seen if one looks at the high proportion of statistical papers published in *Sociological Methodology* and in *Political Analysis*, the two most prestigious methodological outlets in the social sciences. However, we think that the situation has improved over the last 15 years in that research has been done, directed at the quality of the survey method. The following section will be a brief review of this research.

In psychology large sets of questions are used to measure a concept. The quality of these so called tests are normally evaluated using factor analysis,

classical test theory models and reliability measures like Cronbach's α or item response theory (IRT) models. In survey research such large sets of questions are not commonly used. Heise (1969) presented his position for a different approach. He argued that the questions used by sociologists and political scientists cannot be seen as alternative measures for the same concept as in psychology. Each question measures a different concept and therefore a different approach for the evaluation of data quality is needed. He suggested the use of the quasi-simplex models, evaluating the quality of a single question in a design using panel studies. Saris (1981) showed that different questions commonly used for the measurement of "job satisfaction" cannot be seen as indicators of the same concept. Independently of these theoretical arguments, survey researchers are frequently using single questions as indicators for the concepts they want to measure.

In line with this research tradition many studies have been done to evaluate the quality of single survey questions. Alwin and Krosnick (1991) followed the suggestion by Heise and used the quasi-simplex model to evaluate the quality of survey questions. They suggested that on average approximately 50% of the variance in survey research variables is due to random measurement error. Split ballot experiments are directed at determining bias due to question format [Schuman and Presser 1981; Tourangeau et. al. 2000; Krosnick and Fabrigar (forthcoming)]. Nonexperimental research has been done to study the effect of question characteristics on nonresponse and bias (Molenaar 1986). Multitrait-multimethod (MTMM) studies have been done to evaluate the effects of design characteristics on reliability and validity (Andrews 1984; Költringer 1995; Scherpenzeel 1995; Scherpenzeel and Saris 1997). Saris et al. (2004b) have suggested the use of split-ballot MTMM experiments to evaluate single questions with respect to reliability and validity. Cognitive studies concentrate on the aspects of questions that lead to problems in the understanding, retrieval, evaluation, and response of the respondent (Belson 1981, Schwarz and Sudman 1996; Sudman et al. 1996). Several studies have been done to determine the positions of category labels in metric scales (Lodge et al. 1976; Lodge 1981). Interaction analysis has been done to study the problems that certain question formats and question wordings may cause with the interaction between the interviewer and the respondent (Van der Zouwen et al. 1991; Van der Zouwen and Dijkstra 1996). If no interviewer is used, the respondent can also have problems with the questions. This has been studied using keystroke analyses and response latency analyses (Couper et al. 1997).

A lot of attention has also been given to sampling. Kish (1965) and Cochran (1977) have published standard works in this context. More detailed information about new developments can be found in journals like the *Journal of Official Statistics* (JOS) and *Survey Methodology*. More recently nonresponse has become a serious problem and has been given a lot of attention in publication as in the *Journal of Public Opinion Quarterly* and in books by Groves (1989), de Heer (1992), Groves and Couper (1998), Voogt (2003) and Stoop (2005).

As this brief review has demonstrated, the literature on survey research is expanding rapidly. In fact, the literature is so expansive that the whole process cannot be discussed without being superficial. Therefore we have decided to concentrate this book on the process of designing questionnaires. For the more statistical aspects like sampling and the nonresponse problems, we refer the reader to other books.

10. Summary

In this chapter we have described the different choices of survey design. It is a complex process that requires different kinds of expertise. A lot of information about sampling and nonresponse problems can be found in the statistical literature. Organization of fieldwork requires a different kind of expertise. The fieldwork organizations know more about this aspect of survey research. Designing survey questions is again a distinct kind of work, and we do not recommend relying on the statistical literature or the expertise of fieldwork organizations. The design of survey questions is the typical task and responsibility of the researcher. Therefore we will concentrate in this book on questionnaire design. For other aspects of survey research, we refer the reader to the standard literature on sampling and fieldwork. This does not mean that we will not use statistics. In the third part of this book we will discuss the evaluation of the quality of questions through statistical models and analysis. The fourth part of this book will make use of statistical models to show how information about data quality can be employed to improve the analysis of survey data.

Exercises

1. Choose a research topic that you would like to study. What are the most important concepts for this topic? Why are they important?
2. Try to make a first questionnaire to study this topic.
3. Go through the different steps of the survey design mentioned in this chapter and make your choices for the research on which you have chosen to work.

Part I

The three-step procedure to design requests for an answer

In this part we explain a three-step procedure for the design of questions or as we call it requests for answers. First we distinguish between concepts-by-intuition, for which obvious questions can be formulated, and concepts-by-postulation, which are formulated on the basis of concepts-by-intuition. A common mistake made by researchers is that they do not indicate explicitly how the concepts-by-postulation they use are operationalized in concepts-by-intuition. They immediately formulate questions they think are proper ones. For this reason many survey instruments are not clear in their operationalization or even do not measure what they are supposed to measure.

In this part we suggest a three-step approach that, if properly applied, will always lead to a measurement instrument that measures what is supposed to be measured.

The three steps are:

1. Specification of the concept-by-postulation in concepts-by-intuition (Chapter 1)
2. Transformation of concepts-by-intuition in statements indicating the requested concept (Chapter 2)
3. Transformation of the statement into a question (Chapter 3).

This Page Intentionally Left Blank

CHAPTER 1

Concepts-by-postulation and concepts-by-intuition

The effects of the wording of survey questions on their responses have been studied in depth by Sudman and Bradburn (1983), Schuman and Presser (1981), Andrews (1984), Alwin and Krosnick (1991), Molenaar (1986), Költringer (1993), Scherpenzeel and Saris (1997). In contrast, very little attention has been given to the problem of translating concepts into questions (De Groot and Medendorp 1986, Hox 1997). Blalock (1990) and Northrop (1947), distinguish between concepts-by-intuition and concepts-by-postulation.

1.1 Concepts-by-intuition and concepts-by-postulation

Blalock (1990: 34) asserts the following about differentiating between the concepts of intuition and postulation:

> Concepts-by-postulation receive their meaning from the deductive theory in which they are embedded. Ideally, such concepts would be taken either as primitive or undefined or as defined by postulation strictly in terms of other concepts that were already understood. Thus, having defined mass and distance, a physicist defines density as mass divided by volume (distance cube). The second kind of concepts distinguished by Northrop are concepts-by-intuition, or concepts that are more or less immediately perceived by our sensory organs (or their extensions) without recourse to a deductively formulated theory. The color "blue," as perceived by our eyes, would be an example of a concept-by-intuition, whereas "blue" as a wavelength of light would be the corresponding concept-by-postulation.

The distinction he makes between the two is that concepts-by-intuition are simple concepts whose meaning is immediately obvious while concepts-by-postulation are less obvious concepts that require explicit definitions. Concepts-by-postulation are also called constructs. Examples of concepts-by-intuition include judgments, feelings, evaluations, norms, and behaviors. Most of the time it is very obvious that a text presents a feeling (x likes y) or a norm (people should behave in a certain way) or behavior (x does y). We will return

to their classification later. Examples of concepts-by-postulation might include "ethnocentrism," different forms of "racism," and "attitudes toward different objects." One item in a survey cannot present an attitude or racism. For such concepts more items are necessary and, therefore, these concepts need to be defined. This is usually done using a set of items that represent concepts-by-intuition. For example attitudes were originally defined (Krech et al. 1962) by a combination of a cognitive, affective, and action tendency component. In Figure 1.1 an operationalization of the concept-by-postulation "an attitude toward Clinton" is presented in terms of concepts-by-intuition, questions, and assertions representing the possible responses.

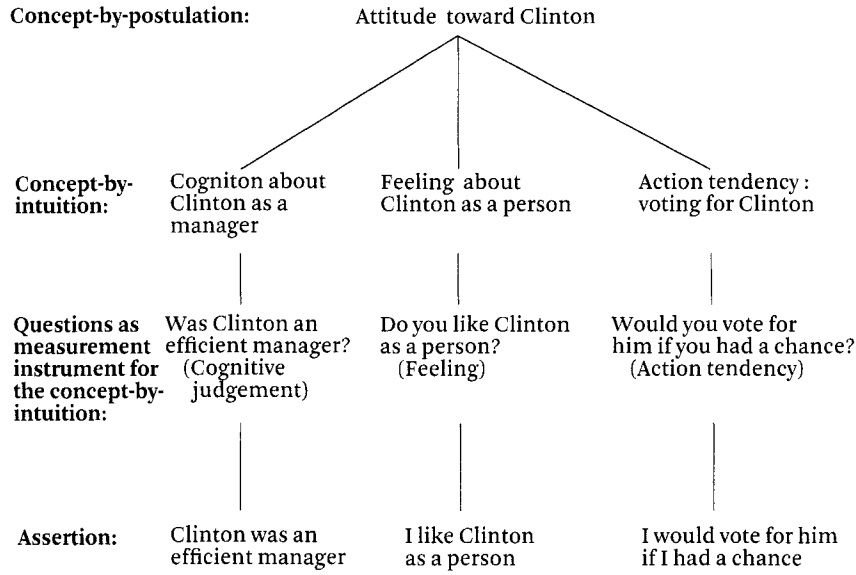


FIGURE 1.1: *Operationalization of an attitude toward Clinton.*

At the bottom of Figure 1.1 three assertions are mentioned. There is no doubt that the assertion "Clinton was an efficient manager" represents a cognitive judgment, that the assertion "I like Clinton as a person" represents a feeling and that the assertion "I would vote for him if I had a chance" represents an action tendency. From this it follows that the questions asking for such assertions represent measurement instruments for "cognitions", "feelings", and "action tendencies," respectively. Given that there is hardly any doubt about the link between these assertions, questions, and the concepts mentioned, these concepts are called concepts-by-intuition. However, the reverse relationship is not necessarily true. There are many different cognitive judgments to formulate about Clinton, including, as leader of his party or as world leader. From this example we can conclude that there are many different possible "cognitions,"

"feelings," and "action tendencies" with respect to Clinton. But normally, after selecting a specific aspect of the topic a question, linked to that concept can be formulated, that reflects the "concept-by-intuition."

In contrast to concepts-by-intuition, concepts-by-postulation are less obvious. In our example in Figure 1.1, the concept-by-postulation "attitude toward Clinton" has been defined according to the attitude concept with the three selected components. However, this choice is debatable. In fact, currently attitudes are often defined on the basis of "evaluations" (Fishbein and Ajzen 1980) and not the components mentioned previously. Although these two operationalizations of attitudes differ, both define attitudes on the basis of concepts-by-intuition.

Blalock as early as in 1968 (Blalock 1968) complained about the gap between the language of theory and research. More than two decades later, when he raised the same issues again, the gap had not been reduced (Blalock 1990). Although he argues that there is always a gap between theory and observations, he also asserts that not enough attention is given to the proper development of the concepts-by-postulation. As an illustration of this we present measurement instruments for different forms of racism.

Several researchers have tried to develop instruments for new constructs related to racism. Typical examples are the following constructs: "symbolic racism" (McConahay and Hough 1976; Kinder and Sears 1981); "aversive racism" (Kovel 1971, Gaertner and Dovidio 1986), "laissez-faire racism" (Bobo et al. 1997), "new racism" (Barker 1981); "everyday racism" (Essed 1984), and "subtle racism" (Pettigrew and Meertens 1995). In all these instruments, similar statements have been employed in different combinations and using different interpretations and terms. Table 1.1 illustrates this point for the operationalizations of symbolic and subtle racism. Table 1.1 shows that five items of the two constructs are the same but each construct is also connected with some specific items. The reason for including these different statements is unclear; nor is there a theoretical reason given for their operationalizations.

The table identifies that "subtle racism" is defined by two norms (items 1 and 2), two feelings (items 5 and 6), four cognitive judgments (items 7a–7d and some other items). It is not at all clear why the presented combination of concepts-by-intuition should lead to the concept-by-postulation "subtle racism." Nor is the overlap in the items and the difference in items between the two concepts-by-postulation, subtle and symbolic racism at all clear. Even the distinction between the items for "blatant racism" and the items of the other two constructs has been criticized (Sniderman and Tetlock, 1986; Sniderman et al. 1991).

One of the major problems in the operationalization process is that the researchers do not, as Blalock suggested, think in terms of concepts-by-intuition but only in terms of questions. They form new constructs without a clear awareness of the basic concepts-by-intuition being represented by the questions. This observation suggests that it would be useful to study the link

Table 1.1: Operationalization of subtle and symbolic racism.

Items		Subtle	Symbolic
1	Os living here should not push themselves where they are not wanted.	+	+
2	Many other groups have come here and overcame prejudice and worked their way up. Os should do the same without demanding special favors.	+	+
3	It is just a matter of some people not trying hard enough. If Os would only try harder, they could be as well off as our people.	+	+
4	Os living here teach their children values and skills different from those required to be successful here.	+	
5	How often have you felt sympathy for Os?	+	+
6	How often have you felt admiration for Os?	+	+
7	How different or similar do you think Os living here are to other people like you		
7a	In the values that they teach their children?	+	
7b	In the religious beliefs and practices?	+	
7c	In their sexual values or practices?	+	
7d	In the language that they speak?	+	
8	Has there been much real change in the position of Os in the past few years?	+	
9	Generations of slavery and discrimination have created conditions that make it difficult for Os to work their way out of the lower class.		+
10	Over the past few years Os have gotten less than they deserve.		+
11	Do Os get much more attention from the government than they deserve?		+
12	Government officials usually pay less attention to a request or complaint from an O person than from "our" people.		+

"O" stands for member(s) of the outgroup, which include "visible minorities" or "immigrants."

between a set of concepts-by-intuition and questions for questionnaires. If such a link could be established, these basic concepts could then be used in a more systematic way to formulate higher-order concepts-by-postulation such as attitudes and others. Therefore, let us shift our attention to the relationship between concepts-by-intuition and the concepts-by-postulation in the following section.

1.2 THE DISTANCE BETWEEN CONCEPTS-BY-INTUITION AND CONCEPTS-BY-POSTULATION

We think that the best way to discuss the issue of the gap between concepts-by-intuition and concepts-by-postulation is to give an example. The example

we use is the measurement of "political interest." In democratic countries it is assumed that people should be sufficiently interested in politics to participate at least in the selection of candidates who represent them in the political institutions. Therefore, "political interest" is a concept that is often at the top of the list of variables to be included in survey research. The measurement of "political interest" at first glance appears deceptively straight forward.

1.2.1 Use of the concept of a direct question

The measurement appears to be simple because a direct question about "political interest" can be formulated.

1.1 How interested are you in politics?

1. *Very interested*
2. *Somewhat interested*
3. *Not very interested*
4. *Not at all interested*

Indeed, in election studies this question is frequently used. This operationalization assumes that political interest can be measured directly by the answers to the direct question and initially it appears that question 1.1 is a measure for the concept-by-intuition "political interest." However a deeper analysis reveals otherwise.

In Figure 1.2 we present this assumption in a path model allowing for random measurement errors (e) due to mistakes in the answer or the recording of the interviewer. This model suggests that people express their political interest directly in their response except for possible random errors. The variable of interest is "political interest." This variable cannot be observed directly because the scores on this variable are in the mind of the respondent. This is called a latent or unobserved variable and is presented in the circle. The responses to question 1.1 can be observed directly. Such variables are usually presented in squares while the random errors, inherent in the registration of any response, are normally denoted by an "e." This model suggests that the verbal report of the question is determined by the unobserved variable "political interest" and random errors. We will use this notation throughout the book.

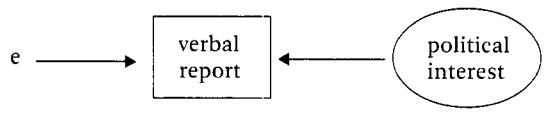


FIGURE 1.2: *A measurement model for a direct measure of political interest.*

Thomassen (2002) suggested question 1.1 for the European Social Survey (ESS), but he comments on this question: "As common as this measurement is, it might have a clear disadvantage. It is not unlikely that people in general will

associate politics with traditional politics and will claim not to be interested in politics, although they are interested in the activities of, for instance, "new social movements." Kriesi (1993) suggested that the "participation in politics" changes from the "classical participation in parties" to "participation in new social movements."

This discussion suggests that the measurement of "political interest" is not so simple. "Political interest" consists of two components: "Interest in activities of party and government organizations" and "interest in activities of nongovernment organizations" (NGOs). Assuming that both components can be measured by a direct question, we get the measurement model presented in Figure 1.3.

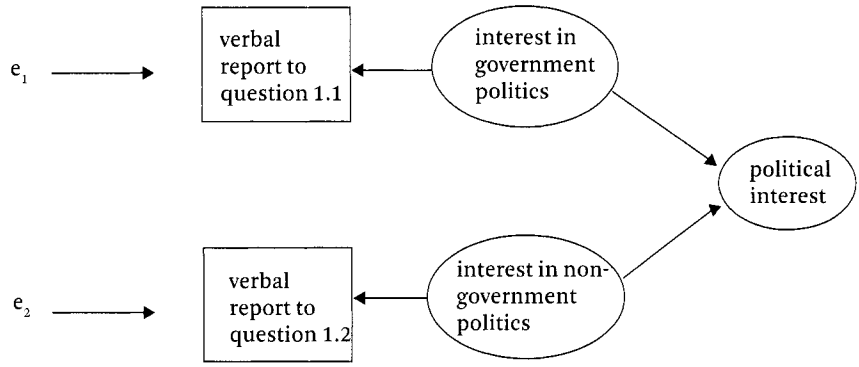


FIGURE 1.3: *A measurement model for a measure of political interest based on two components.*

It is assumed that the second component can be measured by a direct question. However, if this second question could be formulated there still is the issue of how the responses to these two questions can be combined in one index for "political interest." Should the responses get equal weight or is one component more important than the other one? Should they be added or multiplied or should we make an elaborate typology on the basis of the two characteristics? These questions are not simple to solve and will get extra attention in Chapter 14 of this book. In this way the concept "political interest" has become a concept-by-postulation and its operationalization is not immediately apparent anymore. However the two components "interest in party and government politics" and "interest in nongovernment politics" can be seen as concepts-by-intuition, which can be measured by direct questions like question 1.1 and 1.2.

1.2 *There are many organizations that try to influence political decisions in your country and the world, for example, the trade unions, employers organizations, environmental protection organizations.*

How interested, would you say, you are in the activities of such organizations; are you very interested, somewhat interested, not much interested, or not at all interested?

This question was also suggested for use in the ESS. The comment of Thomassen on this ESS proposal was: "A problem with the suggested question is that people probably consider only the specified organizations and do not give a general judgement." This discussion suggests that omitting the direct question 1.2 probably does not cover all the interests people have in politics. On the other hand, adding the direct question 1.2 about politics not connected to party and government politics has the problem that people probably account for the examples mentioned in the question and no other organizations.

However, there are still more issues to consider. When we ask people about their "political interest" in this format, there will be a number of people who want to make a good impression. Therefore, people who are not so interested in politics may have a tendency to exaggerate their interest, called a "social desirable answer" [see, e.g., Schuman and Presser (1981)].

A last problem is that people may not know how to answer such a question. They may ask themselves (and possibly the interviewer) when should I say "somewhat interested" and when "very interested." To simplify, one could also ask a relative judgment by formulating the following:

1.3 *Are you more or less interested in politics than the average citizen?*
1. *Much more*
2. *A bit more*
3. *A bit less*
4. *Much less*

This question is simpler to answer because our judgments are in general relative and not absolute (Poulton 1968). Still, this question requires that one has an impression of the "political interest" of the average citizen. It might be that people have very different impressions of the average citizen and then the responses are incomparable.

This exercise demonstrated that there is sufficient reason not to immediately choose the first direct question that comes to mind, even if in combination with another second direct question. One can also consider if there is a more objective approach to the problem of measuring "political interest."

1.2.2 The use of indirect measures

It is also possible to derive measures for concepts on the assumption that there is a strong relationship between the variable of interest and another variable that can more easily be measured. Again we will illustrate this approach using "political interest" as an example in which we discuss two alternatives: one based on passive behavior and one based on active behavior.

1.2.2.1 The use of passive behavior

A different approach is to use passive behavior as an indicator for political interest. Using this approach people is asked how much they inform themselves about politics through TV, radio, and newspapers. This operationalization assumes that people who are more interested in politics will spend more time on the media to follow what is happening in politics. The latter variable is the variable that is transformed into a direct question. This leads to the measurement model for "political interest" presented in Figure 1.4.

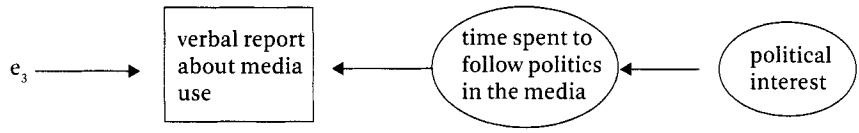


FIGURE 1.4: *Measurement of political interest through media use.*

If the relationship between "political interest" and "the time spent to follow politics in the media" were perfect, then there would be no problem with this approach. However, it is clear that the relationship between "political interest" and "the time spent to follow politics in the media" is not perfect. There are good reasons to argue that other variables will also influence the latter variable. For example, "the amount of leisure time" available to a person may influence the time a person spends on the media. As a consequence, the time spent to follow politics in the media will not be a perfectly valid measure for political interest. Therefore, the validity of this measure is an interesting question for empirical research; it depends on the strength of the relationship between "political interest" and "the time spent to follow politics in the media."

Let us now look at possible measures for this so-called indicator of political interest by examining another ESS question that was proposed for the use of the media. We start with question 1.4. It will be clear that question 1.4 measures the use of the media with certainty, and it is a clear concept-by-intuition. This matrix operationalization is attractive because adding up the answers in a row, the use of the media of a person becomes apparent. The cells in each row also give the relative importance of the different purposes of use of each medium. Adding up the answers in the columns the use of the media for different purposes on individual level becomes apparent. For example, the total in the second column will provide an estimate of the total amount of time spent by respondents on politics and current affairs and a precise measure of the amount of time spent on politics can be derived.

However, this measure also has its problems, even though it is relatively objective and exact. It really is a measure of the use of media and not directly of "political interest." Whether this measure can be used as a valid indicator depends on the strength of the relationship between the concept-by-intuition ("use of the media") and the concept-by-postulation ("political interest").

1.4 Media can be used for different purposes, see the Media Card. Can you estimate for how many minutes you use the TV, radio, and newspaper on a normal day for these different purposes?

Media Card

Different purposes for use of the media:

- Entertainment = quizzes, lotteries, games, shows etc.
- Politics = news, current affairs, political discussions
- Business = financial information, business information
- Sport = reports about sport events or previews
- Hobbies = gardening, home improvement, painting, holidays, etc.
- Education = educational programs, science and technology
- Arts = movies, music, discussions about art

The responses can be registered in the following matrix:

	Entertainment	Politics	Business	Sport	Hobbies	Education	Arts
TV							
Radio							
Newspaper							

But there are two more issues: (1) this measure cannot be used to study the relationship between "political interest" and "media use" and (2) it is questionable whether people can give such precise information about their activities. Perhaps this information can be determined by using hours instead of minutes, but then the answers may not be precise enough. A third issue is that the question asks for a "normal day," but the definitional difficulties for a normal day have been discussed in the literature. The alternative of asking for "a normal weekday " is also imprecise because it has been found that people ignore, for example, daytime viewing on TV (Belson 1981). Another alternative could be to ask the media use for yesterday. In that case people probably remember what they have done and can provide the information reliably, but "yesterday" may have been a very unusual day for some people.

A solution for this problem is to ask people to fill in diaries for a number of days to achieve a stable measure where unusual events are canceled out. A problem with this measure is that the task asked from the respondent is labor-intensive. As a consequence, people will reduce the amount of information they are giving day by day, which will create a downward bias in the measure. It has been documented as a common problem of the use of diaries (Kalfs 1993; Kaper 1999).

In order to solve this problem, automatic registration procedures have been developed at least for TV. Now respondents do not have to answer questions but only push a button on a remote control to register when they begin to watch TV and when they stop watching it. This is an efficient method for TV and probably also for radio (although we are not aware of a similar application for radio). However, it is also not a comprehensive solution because not all media are covered and it involves expensive equipment that does not pay off for just one study.

Our small discussion brings us back to the original idea that for the ESS there were too many questions about media in the set. The original proposal would cost 21 questions because the respondent would have to check all cells in the matrix, otherwise, the researcher would not know whether to code the empty cells zero or as skipped questions. However, given that the measurement of "political interest" by direct questions requires only one or two questions, the discussed alternative procedure, asking about media use with 21 questions, must be much better; otherwise, the direct question would have been preferred.

The alternative chosen by the ESS simplifies the operationalization by asking that for each medium only the total amount of time per activity be indicated, as well as, the time spent on "politics and current affaires." This method preserves the idea of the proposal since it is possible to estimate the amount of time spent on politics relative to the total amount of time spent on the media. Hence, the new version requires only 6 questions instead of the original 21questions of the matrix. The questions are as follows:

1.5a *In total, how much time on an average weekday do you generally spend watching television?*
(RECORD in HOURS/MINUTES)
(filter if respondent does not watch television)

1.5b *And how much of this time on an average weekday (again in minutes) do you spend watching politics and/or current affairs on television?*
(RECORD in HOURS/MINUTES)

1.5c *In total, how much time on an average weekday do you generally spend listening to the radio?*
(RECORD in HOURS/MINUTES)
(filter if respondent does not listen to radio)

1.5d *And how much of this time on an average weekday (again in minutes) do you spend listening to politics and/or current affairs on the radio?*
(RECORD in HOURS/MINUTES)

1.5e *In total, how much time on an average weekday do you generally spend reading the newspaper?*
(RECORD in HOURS/MINUTES)
(filter if respondent does not read newspapers)

1.5f *And how much of this time on an average weekday (again in minutes) do you spend reading about politics and/or current affairs?*
(RECORD in HOURS/MINUTES

This operationalization demonstrated data collection based on passive behavior: "use of the media for political information;" however, it would be equally justified to consider the possibility of active participation in political activities, which will be discussed in the next section.

1.2.2.2 The use of active participation

If one is sure that "political interest" is strongly related with "participation in political activities" direct questions about this participation can also be used for measuring political interest. So a third possibility is that people are asked to indicate to what extent they participate in all kinds of political activities. The idea behind this operationalization is presented in Figure 1.5.

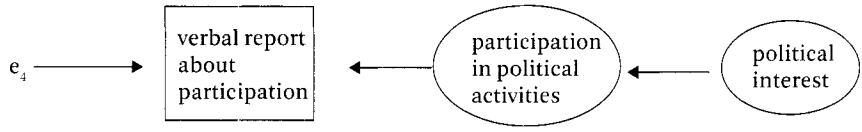


FIGURE 1.5: *Measurement of political interest by participation in political activities.*

The active participation model is similar to the passive participation in the use of the indicator; this is a valid measure only if there is a perfect relationship between "political interest" and "participation in political activities." However, the participation may be completely determined not only by "political interest" but also by "leisure time," "age," and "education." Therefore the indicator may not be perfectly valid for "political interest" and empirical research has to show the validity of the measure for "political interest." Therefore, "participation in political activities" is, since the research of Kaase and Barnes (1979) normally asked using a question battery like 1.6.

In order to get a score for "political participation" the number of times a respondent says "yes" is determined. Although there is no doubt that the items in question 1.6 measure participation in political activities (a concept-by-intuition), it is not so clear that these questions can also be used for measuring the concept-by-postulation "political interest."

This indicator for "political interest" cannot be used to study the relationship between "political interest" and "political participation;" however, there are also other issues with this measure even while employing it as a measure of "political participation." One issue is that the question does not have a comprehensive list of activities. Activities not mentioned are voting, talking about politics with others, and becoming a member of a pressure group. So the question arises as to which actions should be included in the list and which are not and why.

1.6 *There are different ways of attempting to bring about improvements or counteract deterioration in society. During the last 12 months, have you done any of the following?*

	yes	*no*
a. Contacted a politician		
b. Contacted an association or organization		
c. Contacted a national, regional or local civil servant		
d. Worked/volunteered for a political party		
e. Worked/volunteered for a (political) action group		
f. Worked/volunteered for another organization or association		
g. Worn or displayed a campaign badge/sticker		
h. Signed a petition		
i. Taken part in a public demonstration		
j. Taken part in a strike		
k. Boycotted certain products		

Another unconscious choice that the researcher has made is to ask whether the respondent has performed one or more of the activities in the past 12 months but not how frequently. It is probably just as reasonable to assume that the more frequent activities are done (the more time is spent on these activities), the more interested a person is in politics. Therefore, a valid design question is whether the respondents who are very interested in politics do many of the specified activities only once or one activity frequently. It is also an unresolved research design problem with the abovementioned type of questions as to whether the different actions should be weighed equally or that some are more important than others.

We see here again that the concept-by-intuition ("participation in political activities") is clear but the derived scores for the concept-by-postulation "political participation in general" is not that clear; and that holds even more weight for the concept-by-postulation "political interest" as indicated by "political participation."

1.2.3 A last alternative

In each of these operationalizations the same respondents will get a different score even though they are measuring just one concept ("political interest") and therefore should get one score. It is also not certain that these different measures would correlate strongly with each other even though they are supposedly measuring the same variable.

A solution is to hypothesize that each of these scores measures "political interest" but besides those also other factors will influence these scores. This means that there is one common source of correlation between these different

measures, and so one could suggest that the variable "political interest" is the common cause of the three indicators. This idea is presented in the measurement model of Figure 1.6. This figure indicates that each of the three indicators is influenced by the score people have on the unmeasured variable of "political interest." Besides that, each indicator is also influenced by its own unique components (u_j). Furthermore, it is supposed that for each indicator direct questions can be formulated, where the responses to these questions provide the observed score on each indicator. Finally it is supposed that these scores are not perfect measures of the indicators but that each observed score also contains errors (e_j).

This is a rather complex measurement model, but it provides an answer for the problems that people can have different scores on the different indicators while measuring the same variable. In this approach it is assumed that all people have only one score for "political interest." Chapter 14 in this book will address how such a theory can be tested, how these scores can be obtained, and how to evaluate these scores as measures for "political interest."

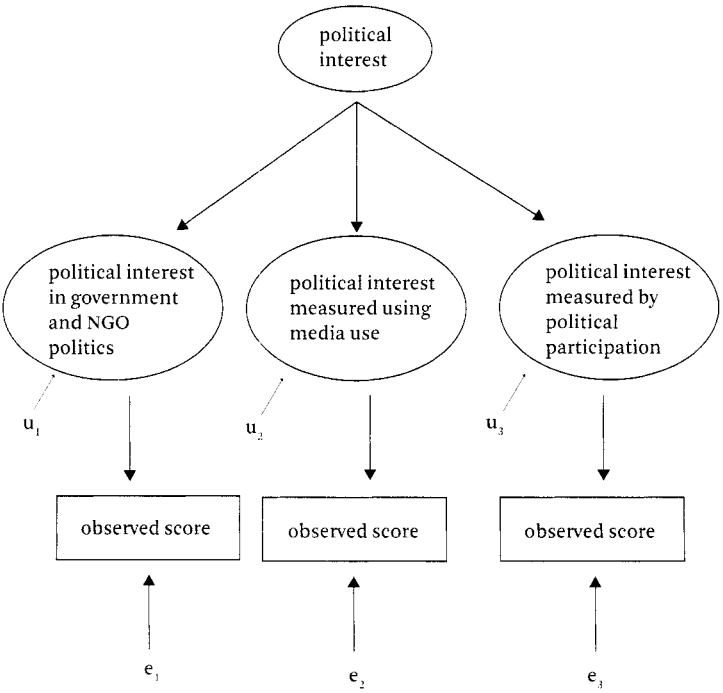


FIGURE 1.6: *Relationships between the latent variable "political interest" and three possible indicators and the observed scores for these indicators.*

1.3 CONCLUSIONS

The first issue discussed in this chapter was the distinction between concepts-by-intuition and concepts-by-postulation. We have seen that concepts-by-intuition are easily transformed into questions. Concepts-by-postulation cannot be operationalized directly in survey questions. They are normally defined by some combination of concepts-by-intuition.

We have also seen in the example of "political interest" that the concept changes through the operationalization. First "political interest" was measured by a direct question and it seemed to be a concept-by-intuition. Then it became a concept-by-postulation based on the combination of two concepts-by-intuition: "interest in government politics" and "interest in nongovernment politics." After that, two operationalizations have been developed on the basis of indirect measures. One operationalization was based on the use of a passive indicator "media use"; the other one, on the use of an active indicator "political participation."

We have also suggested that in each of these operationalizations respondents will get a different score even though we thought to be measuring just one concept "political interest" on which people should have just one score. It is not at all certain that these different measures would correlate very strongly with each other even though these measures are supposed to measure the same variable "political interest." The difference can be due to different systematic components in these measures that reduce their validity. The differences can also be due to incidental errors that occur more in one measure than in another, which would lead to differences in reliability. In any case, it is important to have a scientific method for deriving optimal questions.

This book is directed to answer such questions. However, instead of immediately proceeding to analyze the relationships between concepts-by-postulation and responses to questions, we will concentrate first on the link between concepts-by-intuition and their questions. Only once we know this relationship and can say something about the quality of a single question will we discuss the quality of concepts-by-postulation. The idea is that in order to speak on the quality of concepts-by-postulation, the elements on which the concepts-by-postulation are built need to be identified. For example, if we realize that questions about participation in political activities measure "political behavior," we will be more reluctant to use them as an indicator for "political interest." This prudence is necessary to prevent the construction of concepts-by-postulation that are unclear and will produce confusing results in data analysis. Therefore, we will return to the construction and the evaluation of concepts-by-postulation in Chapter 14 of this book.

EXERCISES

1. Try to formulate questions that represent concepts-by-intuition and concepts by postulation for the following concepts:
   a. Job satisfaction
   b. Happiness
   c. The importance of the value "honesty"
2. In practice it is seldom clear whether the questions suggested measure what they are supposed to measure. Some examples follow below. The following proposal has been made to measure "left-right orientations" in politics. The authors said:
   "The left-right orientation contains two components:
   - Egalitarianism: a policy of equality of incomes
   - Interventionism: a policy of government intervention in the economy by, e.g., Nationalization."

   Items 1–3 in the following list are supposed to measure the egalitarian element; the next two, the interventionism element.
   How strongly do you agree or disagree with the following items?
   Agree completely, agree very much, agree, neither agree nor disagree, disagree, disagree very much, disagree completely
      1. *It is not the government's role to redistribute income from the better off to the worse off.*
      2. *It is the government's responsibility to provide a job for everyone who wants one.*
      3. *Management will always try to get the better of employees if it gets a chance.*
      4. *Private enterprise is the best way to solve Britain's economic problems.*
      5. *Major public services and industries ought to be under state ownership.*

   a. Check whether these assertions represent the concepts they are supposed to represent.
   b. Try to improve the assertions that seem incorrect.
3. Let us now look at the questionnaire you have developed yourself:
   a. Do the questions measure what they are supposed to measure?
   b. Did you use concepts-by-intuition or concepts-by-postulation?
   c. Is it possible that other variables affect the responses than just the variables you would like to measure?
   d. If you think that some of your questions are wrong, try to improve them.

This Page Intentionally Left Blank

CHAPTER 2

From social science concepts-by-intuition to assertions

The first chapter discussed that there are different ways to operationalize a variable of interest. The distinction between concepts-by-postulation and concepts-by-intuition was that some concepts are intuitively clear while others require theoretical and empirical support. In the remainder of Part I of the book we concentrate on the operationalization of concepts-by-intuition. In this chapter we will show how assertions can be formulated for the most common concepts-by-intuition of the social sciences. In doing so, we discuss the linguistic links between these concepts-by-intuition and the different possibilities of their verbal expression in assertions. By use of these rules one can be sure that the assertions generated represent the concept-by-intuition of interest. In the next chapter we will discuss the transition from an assertion to a request for an answer. Any verbal expression of an assertion should at minimum refer to a concept-by-intuition (e.g. behavior, norm, or evaluation) for an object of interest (e.g. government, family, or work). The selection of the concept and object of a request are rather arbitrary but depends mainly on the issue of investigation. Therefore, before we discuss these choices, we will talk about survey items and the link between requests for answers and assertions.

2.1 Description of the components of a survey item

Andrews (1984) defined a survey item as consisting of three different parts of text or components, namely, an introduction, one or more requests for an answer and a response scale. Molenaar (1986) also uses quite similar survey item components. In this chapter we propose to build on their work but to distinguish even more components of a survey item.

In our opinion a survey items can contain the following: an introduction, a motivation, information regarding the content, information regarding a definition, an instruction of the respondent, an instruction of the interviewer, the request for an answer, and response categories or scales. Figure 2.1 summarizes the basic components of a survey item.

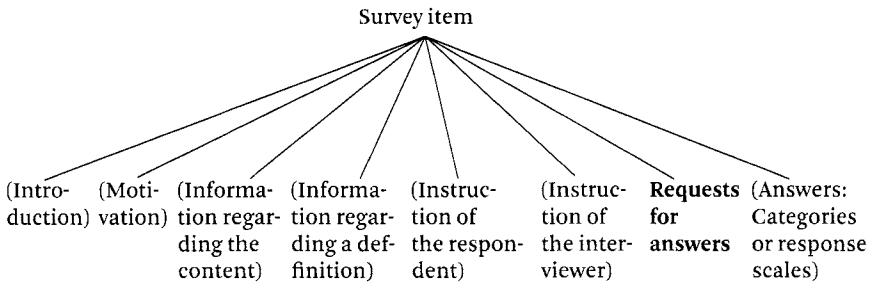


FIGURE 2.1: *Decomposition of a survey item into its components.*

The components indicated within parentheses in Figure 2.1 are optional for the designer of the survey. This implies that the request for an answer is the core unit of a survey item. It also means that the simplest form of a survey item is just an open request for an answer and nothing more. However, Figure 2.1 shows that a survey item can consist of many more components. How many and which ones are frequently used in survey research will be discussed further in Chapter 6. In this chapter we concentrate on the request for an answer.

2.2 ASSERTIONS AND REQUESTS FOR AN ANSWER

In order to clarify the link between basic concepts-by-intuition and verbal expressions of requests, the linguistic components of the sentences that represent the different concepts must be discussed first. The starting point of the discussion is the sentence structure. A *sentence* is defined as a group of words that when written down begins with a capital letter and ends with a full stop, a question mark or an exclamation mark. But, a sentence also can be classified according to its *linguistic meaning* where a distinction is made between *declarative* sentences or *assertions*, *interrogative* sentences or *requests*, *imperative* sentences or *orders*, and *exclamations*. As we will see later in this section, the first three linguistic forms of sentences are used to elicit answers from a respondent, and not only the interrogative form. Therefore, we speak of "requests for answers" and not of questions. The fourth form is not used in survey research.

Most of the items in Table 1.1 (Chapter 1) were declarative sentences or assertions representing specific concepts-by-intuition. The respondents are asked whether they agree or disagree with these assertions. It is not necessary to use such statements. It is also possible to use normal requests. But we will show how an assertion (example 2.1) can be transformed into a request (2.2). The assertion is

2.1 *Immigrants living here should not push themselves where they are not wanted.*

To transform this assertion into a request, we only have to add "Do you think that" then we get

2.2 *Do you think that immigrants living here should not push themselves where they are not wanted?*

In this or similar ways, any statement can be transformed into a request.

It is also possible to transform any request into an assertion (Harris 1978; Givon 1990). The assertion corresponding to the abovementioned request has already been given. Another example of a request is item 8 in Table 1.1. The request was as follows:

2.3 *Has there been much real change in the position of black people in the past few years?*

By inverting the term "there" and the auxiliary verb "has," we obtain from this request the following assertion:

2.4 *There has been much real change in the position of black people in the past few years.*

Similar changes can be performed on any request in order to get an assertion.

Instead of requests or assertions, surveys sometimes use instructions or directives that are called "imperatives" in linguistic terminology. These imperatives can also be transformed into assertions. The following example illustrates this:

2.5a *Tell me if you are in favor of the right of abortion.*

This imperative can be transformed into an assertion as follows:

2.5b *I am in favor of the right of abortion.*

We have shown above that imperatives and interrogatives can be used to elicit answers from the respondents and can also be linguistically transformed into assertions or statements. Although this is true, it should be clear that there are fundamental differences between "requests requiring an answer" and the related assertions. In fact, a request for an answer, whatever the form of the request may be, presents the respondent with a set of possible answers, called the *uncertainty space* by Groenendijk and Stokhof (1997). On the other hand, an assertion is a specific choice from the set. Take example 2.5a, where the request was:

2.5c *Tell me if you are in favor of the right of abortion.*

This request for an answer allows not only for the assertion 2.5d:

2.5d *I am in favor of the right of abortion.*

but equally for the assertion 2.5e

2.5e *I am not in favor of the right of abortion.*

Although this inequality exists between the requests for an answer and the assertions, we prefer to discuss the link between concepts and requests for an answer on the basis of the related assertions. (We need to keep in mind that there is an almost unlimited number of forms for the requests of an answer[1]). The use of assertions therefore simplifies the discussion. In Chapters 3 and 4 we will discuss how these assertions can be transformed into requests for an answer. In order to discuss the link between the basic concepts and their related assertions, the next section introduces the structure of assertions.

2.3 THE BASIC ELEMENTS OF ASSERTIONS

Sentences can be divided into sentence constituents or phrases and their syntactic functional components. In this section we will discuss the decomposition of assertions into these elements in order to determine how concepts-by-intuition can be formulated in assertions and what parts of assertions can indicate the concept-by-intuition that is represented.

In linguistics a simple assertion is decomposed in two main components: a noun phrase (NP) and a verb phrase (VP). A *noun phrase (NP)* consists of one or more words with a noun or pronoun as the main part. A *verb phrase (VP)* is a phrase that has one or more verbs. But next to the verb, verb phrases contain all the remaining words in the sentence outside the noun phrase, which can be complements, objects, or adverbials. The reader should be aware that we use here the definition of verb phrase as employed in transformational generative grammar (Richards et al. 1993: 399). Example 2.6a might illustrate this:

2.6a	*Clinton*	*was a good president.*
	NP +	VP

Example 2.6a shows a *simple sentence or clause* where the NP is "Clinton" and "was a good president" is the VP. Although this decomposition in NP and VP is very common, for our purposes a more detailed decomposition is more useful. This decomposition is indicated in 2.6b and all the following examples. One can always use the distinction between NP and VP but we will concentrate on the parts of these components:[2]

2.6b	*Clinton*	*was*	*a good president.*
	Subject +	Predicator +	Subject Complement.

As example 2.6b illustrates "Clinton" functions as the *subject* that indicates what is being discussed in the sentence. The *predicator* or the verb is "was" and connects the subject with the remaining part of the sentence, which is again

---

1 In the survey literature the term "stem" of a question is used (Bartelds et al. 1994; Dillman 2000) in a similar manner to the term assertion, but the term "stem" is used for different meanings. Consequently we prefer the term "assertion."

2 The linguistic aspect of this section is based on the work of Koning and Van der Voort (1997). We would like to thank dr. Van der Voort for his useful comments on this chapter.

a noun ("president") with an adjective ("good") as modifier of the noun. This specific remaining part expresses what the *subject is* and is therefore called a *subject complement*. Predicators that indicate what a subject is/was or becomes/became are called *link verbs (LV predicator)*. Other examples of verbs that can function as link verbs (connecting a subject with a subject complement) are "get," "grow," "seem," "look like," "appear," "prove," "turn out," "remain," "continue," "keep," "make," and so on. (Koning and Van der Voort 1997: 48–49). We suggest that the negations of these verbs are also classified as link verbs, for example: "not look like," "being unlike," and "being different from." According to the linguistic functions of the words, the sentence structure of example 2.6b can be formalized as structure 1:

Structure 1: Subject + LV predicator + subject complement.

It can easily be shown that one can make different assertions that refer to different concepts using this structure. As an illustration, we will create different assertions using as subject "my work" and as link verb "is" while the subject complement varies across the examples:

2.7a My work is useful.
2.7b My work is pleasant.
2.7c My work is important.
2.7d My work is visible.

We see by these examples that changing the subject complement (which is each time a different adjective) the sentence refers to a different concept-by-intuition. These examples refer to an evaluation, a feeling, an importance judgment, and a neutral cognitive judgment. We will see later that structure 1 is the basic structure for assertions expressing evaluations, feelings, importance, demographic variables, values, and cognitive judgments.

A second relevant linguistic structure is illustrated in example 2.8a:

2.8a My mother had washed the clothes.
Subject + predicator + direct object.

This example has a subject ("my mother"), the predicator "had washed," and a direct object is "clothes." Koning and Van der Voort (1997: 52) define a *direct object* as the person, thing, or animal that is "affected" by the action or state expressed by the predicator. The linguistic structure of example 2.8a thus can be summarized as structure 2:

Structure 2: Subject + predicator + direct object.

It can easily be shown through examples that changing the predicator in this structure, changes the concept-by-intuition that the assertion refers to. In the

examples we always use "I" as subject and "my clothes" as direct object. By varying the predicator we formulate different sentences which refer to different concepts-by-intuition:

2.8b *I washed my clothes.*
2.8c *I should wash my clothes.*
2.8d *I shall wash my clothes.*
2.8e *I prefer to wash my clothes myself.*
2.8f *I hate to wash my clothes.*

Although the subject and the direct object remain the same, variation in the verb changes the meaning of the assertion. In sequence of appearance above, the sentences refer to a behavior, a norm, a behavioral intention, a preference, and a feeling. Note that example 2.8e even displays a second direct object ("myself").

As we will show structure 2 can be used to formulate relations, preferences, duties, rights, actions, expectations, feelings, and behaviour, to name only a few examples. These will be discussed further in the following sections. Structure 2 has predicators called *lexical verbs* in linguistic terminology. This means that these verbs have full meanings on diverse topics in contrast with link verbs (structure 1) Thus the use of various lexical verbs in predicators explains for a great deal why the concepts change in these assertions. Sometimes the lexical verb is preceded by an auxiliary verb such as "should" (2.8c) and "shall" (2.8d). Its function in 2.8c is to modify the lexical verb in the predicator into an obligation, and in this way it contributes to the change of the concept by intuition. In example 2.8d the auxiliary "shall" modifies the lexical verb into the future tense and this contributes again to the change of the concept-by-intuition.

There is a third linguistic structure relevant to the context of expressing assertions. Example 2.9a illustrates its structure:

2.9a *The position of the blacks has changed.*
Subject + predicator.

Example 2.9a has a subject "the position of the black" and a predicator "has changed." In linguistics these verbs which are not followed by a direct object are called *intransitive*. The basic structure of these assertions can be summarized in structure 3:

Structure 3: Subject + predicator

It can be shown that the meaning of the sentences is easily changed by changing the predicator as previously in structure 2. However, the number of possibilities is much more limited because of the reduced number of intransi tive verbs. Some examples are provided below:

2.9b *I will go to sleep.*
2.9c *I slept.*

Here the subject is "I" and the first sentence (2.9b) indicates a behavioral intention while the second (2.9c) is a behavior. Here are two more examples:

2.9d The position of blacks will change.
2.9e The position of blacks has changed.

In 2.9d the subject is "the position of blacks" and the first sentence indicates a future event and the second, 2.9e, a past event. This structure is frequently used to present behavior, behavioral intentions and past and future events.

So far, we have discussed the basic components of three possible linguistic structures of assertions that can be extended with other components, as will be explained in the next sections.

2.3.1 Indirect objects as extensions of simple assertions

The first extra component that can be added to the basic structures discussed above are indirect objects. An *indirect object* is defined as the person and sometimes also the thing that benefits from the action expressed by the predicator and the direct object (Koning and Van der Voort 1997: 56). Examples 2.10a and 2.10b are illustrations:

2.10a	*Honesty*	*is*	*very important*	*to me.*
	Subject +	LV predicator +	subject + complement	indirect object

Example 2.10a has structure 1 but an indirect object "to me" is added to it. Example 2.10b illustrates the same extension for structure 2:

2.10b	*He*	*bought*	*an apartment*	*for his mother.*
	Subject +	predicator +	direct object +	indirect object

In this example the subject "he" is connected by the predicator "bought" and followed by a direct object "apartment" and then an *indirect object* "for his mother." The general structure of this assertion is the same as structure 2 with the addition of an indirect object.

2.3.2 Adverbials as extensions of simple assertions

Another component that can be added to the basic structure is an adverbial. An *adverbial* gives information about when, where, why, how and under what circumstances, or to what degree something takes place, took place or will take place. Adverbials can occur in all three structures and can have quite different forms (Koning and Van der Voort 1997: 59). Examples 2.11, 2.12, and 2.13 illustrate this:

2.11	*Clinton*	*was*	*president*	*from 1992 to 2000.*
	Subject +	predicator +	subject complement +	adverbial.

This is an extension of structure 1 with an adverbial indicating *when* it happened.

2.12 *My mother had washed the clothes in the washing machine.*
Subject + predicator + direct object + adverbial.

This is an extension of structure 2 with an adverbial indicating *the way* it was done.

2.13 *He worked a lot.*
Subject + predicator + adverbial.

This is an extension of structure 3 with an adverbial indicating *a degree* of working.

2.3.3 Modifiers as extensions of simple assertions

Another very common component attached to nouns is a modifier. A *modifier* specifies a noun. The modifiers can be placed before and after the noun and can be related to the subject but also to the object. Examples 2.14, 2.15, and 2.16 illustrate the use of modifiers for the three basic structures.

2.14 *The popular Clinton was president.*
Subject (modifier + noun) + predicator + subject complement

This is an extension of structure 1 with a modifier for the subject Clinton.

2.15 *My mother had washed the dirty clothes.*
Subject + predicator + direct object (modifier + noun).

This is an extension of structure 2 with a modifier of the noun in the direct object.

2.16 *The son of my brother died.*
Subject (noun + modifier) + predicator

This is an extension of structure 3 with a modifier (of my brother) attached to the subject. The noun phrase as a whole including the modifier is seen as the subject not just the main word in the phrase. For that reason we have put the modifier and the noun in brackets because together they form the phrase mentioned before. In this way the basic structure is immediately evident.

2.3.4 Object complements as extensions of simple assertions

Koning and Van der Voort (1997: 54) define the object complement as a noun, adjective or prepositional phrase that follows the direct object and expresses what the direct object is or becomes. Please see examples 2.17 and 2.18 below:

2.17 *They are driving me crazy.*
Subject + predicator + direct object + object complement

2.18 *I consider him as a friend.*
Subject + predicator + direct object + object complement

These structures of 2.17 and 2.18 are the same as structure 2 with an additional object complement "crazy" or "as a friend." Although this kind of expression occurs seldom in survey research, for the sake of completeness it has been presented here.

2.3.5 Some notation rules

So far we have described three distinct forms of assertions that are relevant for concepts-by-intuition in the social sciences.

Structure 1 of an assertion connects the grammatical subject (x) by means of a link verb (I) in the predicator to a subject complement (sc). The form of this assertion is denoted simply by (xIsc). In principle the "sc" could be anything, but the most frequently occurring sc´s are denoted as follows:

c denotes a neutral judgment like "large/small," "active/passive," "obvious" etc.
ca denotes a relation such as "(to be) the cause/ reason /source of" etc.
d denotes a demographic variable like "age," "profession," "date of birth/ marriage" etc.
e denotes an evaluation like "good/bad," "valuable," "advantageous/disadvantageous," etc.
f denotes a feeling or affective evaluation such as "nice/awful," "pleasant/ unpleasant," "happy/unhappy," etc.
i denotes "important," "interesting"
pr denotes a preference such as "for/against," "in favor/in disfavor" etc.
ri denotes a right like "permitted/allowed/justified,/accepted" etc.
s denotes "similarity" or "dissimilarity" such as "alike/unlike," similar/ dissimilar" etc.

The subject (x) can also be represented by anything, but we use specific symbols for frequently occurring subjects for coding purposes:

g stands for government or politicians
o denotes anyone or everybody
r denotes the respondent himself
v denotes a value

Structure 2 is denoted by (xPy), where the grammatical subject (x) is connected by the lexical verb (P) to the predicator "y," which contains a direct object in the simplest form. Also the same subjects as mentioned previously can be applicable. In this structure the predicators play a major role. Since there are some very frequently employed lexical verbs for predicators that relate to the intuitional concepts of social science, we will denote them with specific symbols:

C indicates relationships where the subject causes or influences the object
D indicates deeds such as "does," "is doing," "did," or "has done"
E indicates predicators specifying expectations such as "expects," or "anticipates"

F specifies feelings as links such as "like/dislike", "feel"[3] "worry about," etc.
FD indicates a predicator referring to future deeds such as "will," "intends," "wishes"
(H+I) specifies a predicator which contains words like "has to" or "should," "is necessary," etc. followed by an infinitive
HR specifies predicators like "has the right to" or "is allowed to"
J specifies a judgment connector such as "consider," "believe," "think"
PR indicates predicators referring to preferences such as "preferred to"
S indicates relationships where a similarity (closeness) or difference (distance) between the subject and the object is indicated

Structure 3 for assertions will be denoted by (xP). Here the predicator (P) and a subject (x) are present without a direct object. An adverbial can follow the predicator. The same choices can be made for the subject and the predicator as enumerated previously.

Having discussed the basic structures of simple assertions in general the next section will discuss the characteristics of the typical assertions for the most commonly used concepts-by-intuition in survey research.

2.4 CONCEPTS-BY-INTUITION IN SURVEY RESEARCH

In this section we will describe how assertions that are characteristic of the concepts-by-intuition employed in survey research can be generated. Most researchers dealing with survey research (Oppenheim 1966; Sudman and Bradburn 1983; Bradburn and Sudman 1988; Smith 1987) make a distinction between factual or demographic requests, requests of "opinion" or "attitudes" and where they arise, requests of knowledge and behavior. The terms *opinion* and *attitude* are often used in these studies for any type of subjective variables. "Attitude" is not discussed here because we consider attitudes as concepts-by-postulation. Since we want to make a distinction between different kinds of opinions, the term "opinion" itself is also not used in this book.

In the sections that follow the structure of the connected assertions are introduced for different concepts. We start with so called subjective variables.

2.4.1 Subjective variables

By subjective variables, as stated, we understand variables for which the information can only be obtained from a respondent because the information exists only in his/her mind. The following concepts-by-intuition are discussed: evaluations, importance judgments, feelings, cognitive judgments, perceived rela-

[3] Note that verbs such as "like," "feel," and "resemble" are linguistically mostly considered as linking verbs followed by a subject complement. However, we prefer to classify them according to their semantic meaning as feelings and similarity like lexical verbs. But the part that follows should grammatically always be considered as a subject complement.

tionships between the x and y variables, evaluative beliefs, preferences, norms, policies, rights, action tendencies and expectations of future events. We begin with evaluations.

Evaluations are seen by most researchers as concepts-by-intuition of attitudes (Fishbein and Ajzen 1975; Bradburn and Sudman 1988; Van der Pligt and de Vries 1995; Tesser and Martin 1996). Their structure (xIe) generates assertions that certainly are expressions of "evaluations" (a_e). Typical for such assertions is that the subject complement is evaluative. Examples of evaluative words are good/bad, positive/negative, perfect/imperfect, excellent/poor, superior/inferior, favorable/unfavorable, satisfactory/unsatisfactory, sufficient/insufficient, advantageous/disadvantageous, useful/useless, profitable/unprofitable, lucrative/unlucrative, and so on. Examples 2.19 and 2.20 are typical examples of assertions indicating an evaluation:

2.19 *Clinton was a good president.*

It is very clear that this assertion indicates an evaluation: the (x) is "Clinton," the evaluative subject complement (e) is "a good president" and the link verb predictor (I) is "was."

2.20 *Their work was perfect.*

Also this is clearly an evaluative assertion where the subject is "their work," the linking verb is "was," and the subject complement is "perfect." Using structure 1 combined with an evaluative subject complement ensures that the assertion created is an evaluation of the chosen subject.

Importance is the next concept to discuss. The structure of an "importance" assertion (a_i) is (xIi) which means "x is important." This assertion has the same form as the assertions indicating evaluations. The only difference is that the subject complement is in this case an expression of "importance." Example 2.21 illustrates this:

2.21 *My work is important.*

"My work" is the subject (x) and "important" represents the subject complement (i), while "is" is the link verb (I). Values are often used as subjects. A value (v) can be defined as a basic goal or state for which individuals strive such as "honesty," "security," "justice," and "happiness" (Rokeach 1973; Schwartz and Bardi 2001). A typical example is:

2.22 *Honesty is very important to me.*

In example 2.22 (x) is the value "honesty," the predicator (I) is "is," and "very important" is the subject complement of "honesty," while "to me" is an indirect object. There is no doubt that assertions generated with structure 1 and an importance subject complement represent importance judgments.

Feelings or affective evaluations have in the past been considered as belonging to evaluations (Bradburn and Sudman 1988; Van der Pligt and de

Vries 1995). However, more recently a distinction has been made between cognitive evaluations and affective evaluations or feelings (Abelson et al. 1982; Zanna and Rempel 1988; Bagozzi 1989; Ajzen 1991). Three basic assertions can be formulated to express feelings. First, (a_f) can be in the form of (xIf) as example 2.23 illustrates:

2.23 My work is nice.

Example 2.23 reads as follows: (x) is "my work," (I) is the link verb predicator "is," and (f) is the *affective* subject complement "nice." It will become clear that other feeling words can be used as a subject complement, which will be discussed. However, structure 1 combined with a feeling subject complement generates an assertion that expresses a feeling with certainty.

The second structure that can be used to express feelings is (xFy), which is an example of structure 2 discussed previously. An example is assertion 2.24:

2.24 I like my work.

In the case of 2.24, "I" is (x), the verb in the predicator "like" is a feeling (F), and "my work" is grammatically a subject complement (see note 3). There is no doubt that this assertion expresses a feeling toward "my work." It is also quite clear that other feelings can be expressed by using a different feeling verb like "hate" or any other feeling word, as in 2.25. Therefore structure 2 with a predicator as a verb that expresses a feeling generates an assertion that represents a feeling.

The third possible structure is (xPy_f) as shown by example 2.25:

2.25 Politicians make me angry.

Example 2.25 reads as follows: (x) is "politicians," (P) stands for the verb form "make," while "me" is the direct object and "angry," expressing a feeling (f), is the object complement. This is one of the few examples of this structure in survey research. Nevertheless, combining structure 2 with a feeling object complement will generate an assertion that will also express a feeling.

Thus (f) or (F) stands for feelings (fear, disgust, anger, sadness, contempt, shame, humility, hope, desire, happiness, surprise, etc.) (Cornelius 1996) that could be grammatically either *lexical verbs* (frighten, fear, scare, terrify, disgust, offend, repulse, enrage, infuriate, despise, disdain, reject, surprise, amaze, astonish etc.) or *subject or object complements* (afraid, distressed, ashamed angry, disappointed, happy, lucky, crazy, etc.).

With "f" the subject or object complement form is denoted and with "F" the lexical verb in the predicator is indicated. The use of "f" or "F," makes a difference in the structure of the assertion but not in the concept presented.

Cognitions have been discussed in the psychology literature as one of the basic components of an attitude (Krech and Crutchfield 1948; Bradburn and Sudman 1988; Ajzen 1989; Eagly and Chaiken 1993; Van der Pligt and de Vries 1995). Two kinds of cognition have been mentioned in the literature. The first

is a *cognitive judgment*. The structure of an assertion representing a cognitive judgment (a_j) is (xIc), which denotes that x has characteristic c. We use c to indicate that a specific type of subject complement must be used. Subject complements of cognitive judgments pertain to neutral connotations such as active/passive, requestable/unrequestable, limited/unlimited, aware/unaware, reasonable/unreasonable, usual/unusual, regular/irregular, ordinary/extraordinary, conservative/progressive, direct/indirect, big/small, slow/quick, left/right, planned/unplanned, practical/impractical, flexible/inflexible, heavy/light, predictable/unpredictable, and so on. It is important to note that the main requirement is that the subject complements do not represent "evaluations," "feelings," and "importance." Example 2.26 displays a typical assertion of a cognitive judgment:

2.26 *Our family was large.*

In 2.26 the subject complement (sc) is the neutral term "large." This example shows that structure 1 combined with a neutral subject complement will generate assertions that express cognitive judgments.

The second concept in the class of cognitions is *a relationship* between a subject x and an object y. However, we need to make a distinction between two relationships: *causal relationships* and *similarity or dissimilarity and connectedness relationships*.

Causal relationships are, for example, studied in attribution theory (Kelley and Michela 1980). There are two structures for causal relationships (a_c): Structure 1 and structure 2. Structure 1 can be used if the subject complement indicates a cause (xCsc). Example 2.27 illustrates this possibility.

2.27 *New laws were the cause of the change of the position of black people.*

There is no doubt that example 2.27 represents a causal relationship where "new laws" (x) is the subject, "were" (I) is the link verb, and "the cause of the change of the position of black people" is the subject complement (sc) with several modifiers.

Structure 2 combined with *a causal or influence predicator* is also typical for assertions indicating a causal relationship. The formal structure can be represented by (xCy), which means (x has a causal relationship with y). Examples of cause or influence indicating lexical verbs are produce, bring about, provoke, create, replace, remove, alter, affect, accomplish, achieve, attain, or lead to. All are used in the sense of being the outcome or consequence of something. Note that relations are expressed by lexical verbs and not adjectives. Example 2.28 is an assertion which indicates a causal relationship:

2.28 *New laws have changed the position of black people.*

Example 2.28 indicates a causal relationship where the (x) "new laws" have changed (C) "the position of black people" (y). This example demonstrates that

structure 2 assertions with a causal predicator will always indicate a causal relationship.

Other types of relationships frequently studied in social science refer to the *similarity/dissimilarity* or *distance/closeness* between objects (e.g. Rabinowitz et al.1991; Stokes 1963) or *connectedness between subjects* (Harary 1971; Helmers et al. 1975, Knoke and Kuklinski 1982; Ferligoj and Hlebec 1999). Examples include being attached to, resembling, being similar, identical/different, being like/unlike, being close. To express such similarity relations in assertions (a_s) structure 1 can be used with a similarity or dissimilarity expressed in the subject complement (xIs) or structure 2 with a similarity or dissimilarity expressing predicator (xSy). We start by illustrating the use of structure 1. An example of the relationship in the sense of membership is found in 2.29:

2.29 *He is strongly attached to the Labor Party.*

In example 2.29 the (x) is "He," the link verb predicator (I) is "is," and the subject complement is " strongly attached" followed by an indirect object " to the Labor Party." To indicate dissimilarity one can use a negation of the assertion in 2.29 "do not resemble" or " are different from." Example 2.30 is an example of a dissimilarity assertion:

2.30 *The Republicans are different from Democrats.*

In example 2.30 the (x) is "Republicans," the link verb predicator (I) is "are" and then follows the subject complement "different" with the indirect object "from the Democrats," expressing the negation of similarity.

So far we have shown that structure 1 can be used to express similarity relations. However, structure 2 can also be used for the same purpose as the following three examples illustrate. A first example is given in 2.31:

2.31 *European Liberals resemble American Conservatives.*

Here the subject (x) "European Liberals" is said to "resemble," the predicator (S) and the (y) is "American Conservatives." The reader should be aware as stated previously (note 3) that we consider "resemble" as a lexical verb but the "y" (American Conservatives) that follows is grammatically a subject complement. A second example is presented in 2.32.

2.32 *Republicans differ from Democrats.*

In this example "Republicans" are again the subject (x), the predicator indicating dissimilarity (S) is "differ from" and the direct object (y) is "Democrats."

Example 2.33 expresses the same concept-by-intuition by employing structure 3

2.33 *Their opinions varied*

In this example "Their opinions" is the subject and the dissimilarity predicator is "varied."

Assertions about relationships indicate the views that respondents hold about the relationship between a subject and an object and not just about one subject. In this respect, relational assertions provide a different type of information than cognitive judgments, although both have been called *cognitions* in the academic literature as long as the assertions indicate neutral judgments.

Preferences are frequently asked in consumer research, election studies, and in studies of policies where items from the most preferred to the least preferred are compared (Torgerson 1958; Von Winterfeld and Edwards 1986). The structure of a preference assertion (a_{pr}) is embedded in structure 2 with a lexical verb in the predicator, indicating preference and denoted as xPRyz..., which means (x prefers y above z...) as in the example 2.34:

2.34 *I prefer the Socialist Party above the Conservative and Liberal Party.*

Here "I" indicates (x), "prefer" is the preference verb (PR), the direct object (y) is "the Socialist Party," and the text "above the Conservative and Liberal Party" indicates an object complement (z). As 2.34 demonstrates, several items are compared, and one is preferred to the others. Often no explicit comparison is made but the assertion is based on an implicit comparison. Example 2.35 displays this form:

2.35 *I favor a direct election of the president.*

In example 2.35 "I" indicates again (x), "favor" is the preference verb (P), and (y) contains only a direct object with a modifier "a direct election of the president." This assertion thus indicates explicitly the preference of a direct election of the president. Implicit in this assertion is the comparison with the opposite of direct elections which are indirect elections.

Another frequently occurring type of assertion indicating a preference in survey research pertains to structure 1and is indicated by (xIpr). Examples 2.36 and 2.37 illustrate this:

2.36 *I am for abortion.*

2.37 *I am against abortion.*

In these examples "I" indicates the subject, in this case, the respondent (r), the link verb predicator is "am", while "for abortion " (2.36) and "against abortion" (2.37) are preference subject complements (p). In these cases the explicit preference is expressed in the subject complement.

Norms are also central to social research (Sorokin 1928; Parsons 1951; Homans 1965). Coleman (1990: 242) defines them as specifications of "what actions are regarded by a set of persons (o) as proper or correct." Structure 2 with an obligation indicating word (H) in the predictor followed by an infinitive (I) can be used to express a norm (a_n)= (o(H+I)y), which means that someone should do something to the direct object (y). Example 2.38 illustrates this concept:

2.38 Immigrants should adjust to the culture in their new country.

In example 2.38 the "immigrants" are the persons (o) for whom this norm holds, "should" stands for the obligation indicating part (H) of the predicator, which also contains the infinitive "adjust to" (H + I), while the direct object (y) with a modifier is "the culture in their new country." For norms also structure 3 can be used as the following example illustrates:

2.39 The children should go to sleep.

This assertion also indicates a norm, but does not contain a direct object. In that case the structure indicates (o) "the children" and the predicator consists of the obligation indicating auxiliary (H) "should" and the infinitive (I) "go to sleep."

Policies are an important topic in political science research. They are used to determine what the public thinks about different measures of the government (Sniderman et al. 1991; Holsti 1996). A policy assertion (a_p) has the structure (g(H+I)y), which means (the government should do something for y). Example 2.40 displays a policy assertion:

2.40 The government should not allow more immigrants.

In example 2.40 "the government" is (g), the predicator is "should not allow," which contains the obligation indicating word "should" and the infinitive "allow," while the direct object is " more immigrants."

Structure 3 can also be used with policies as example 2.41 illustrates:

2.41 The government has to resign.

In example 2.41 there is no direct object therefore structure 3 is applicable and the form is (g (H +I)). The only difference between norms and policies is that there is another subject. Norms are used for explaining the behavior of people (o) while policies indicate obligations for the government (g).

Rights, specifically requests for an answer dealing with civil right issues, are often queried in political science research (Sniderman et al. 1991). These perceived rights can be expressed using structure 1 where the subject is the matter at stake (e.g. abortion) and as subject complement (ri) an expression of permission such as "accepted," "allowed," or justified," which we will denoted by (xIri). An example of this type of concept is the following:

2.42 Abortion is permitted.

However, rights can also be expressed using structure 2. Then the assertion (a_{ri}) must contain a combination (oHRy), which means (someone has the right y). Example 2.43 illustrates our point:

2.43 Immigrants also have the right of social security.

The "immigrants" (o), "have the right of something" indicates the typical combination of the verb "have" and the direct object "the right of something"

(HR). The "of something," in this case "of social security" is a modifier within the direct object.

Action tendencies are often considered as the third component of an attitude (Ajzen and Fishbein 1980; Bradburn and Sudman 1988; Sudman and Bradburn 1983; Eagly and Chaiken 1993). An action tendency is what one intends to do in the future. The concept action tendency (a_t) can be represented in structure 2 or 3 where the predicator indicates a future deed of the respondent or (rFDy), which means r will do y. An example could be the following:

2.44 I want to go to the doctor.

Example 2.44 is a structure 3, where "I" is (r), "want to go" is the predicator (FD) indicating a future deed, and "to the doctor" is an adverbial. Structure 2 is also possible if the verb requires a direct object:

2.45 I will do my homework soon.

Example 2.45 uses structure 2 because there is a direct object (y) "my homework." It is followed by the adverbial "soon." In both cases (2.44 and 2.45) the most typical is the predicator which expresses a future deed of the respondent.

Expectations of future events (Graesser et al. 1994) are anticipations of events in which the respondent is not involved. The structure for an expectation (a_{ex}) is the same as in the case of action tendencies. The only difference is that the subject is not the respondent (r) but another grammatical subject (x). This means that the structure is xFD or xFDy. Examples are 2.46 and 2.47:

2.46 The storm will come.

2.47 The storm will destroy many houses.

So far all assertions have been clear about the concepts that they were supposed to represent. There are, however, also assertions used for which the meaning is not so clear. This is sometimes done intentionally but more often than not, by mistake. One of such types of assertions will be discussed below under the heading "evaluative beliefs."

Evaluative beliefs (a_{eb}) can be represented by many different types of assertions. Typically they have a *positive or negative connotation* (Oskamp 1991). Assertions presenting causal relationships are often used in this context. But because of their evaluative connotation, they indicate not only a causal relationship but also an evaluation of it. Therefore they are called "evaluative beliefs." These assertions are indicated by a_{eb}. In case of a causal relationship one structure is represented by (xCy_e). Example 2.48 illustrates this:

2.48 The budget reform has led to prosperity in the United States.

The "budget reform" is (x), "prosperity in the United States" is (y), and "has led to" is the causal predicator (C.) The noun "prosperity" referring to object (y) is clearly a word with a positive connotation (e), and therefore one can say that this statement also expresses an evaluation, besides the fact that it expresses a

relationship, which is typical to evaluative beliefs (y_e). A slightly different form of an evaluative belief is that the relationship predicator (C) contains a positive or negative connotation that is indicated by ($xC_e y$):

2.49 *The war destroyed a lot of buildings.*

In example 2.49 the subject (x) is "the war" which "destroyed" (C_e) "a lot of buildings" (y).

Behavioral assertions, which will be discussed in more detail in the paragraph on objective variables, can also become evaluative beliefs. Example 2.50 illustrates this:

2.50 *The Netherlands prospered in the 17th century.*

In this example the predicator "prospered" expresses a past deed with a positive connotation (D_e), which makes it an evaluative belief with the form ($xD_e y$).

These previous examples demonstrate that structures that do not contain explicitly evaluative terms can nevertheless indicate evaluative beliefs. In such a case, the assertion has to contain words with an evaluative connotation such as: to prosper, prosperity, succeed, success, flourish, fail, failure, miss, loss, destroy, spoil, kill.

Assertions indicating the concept "evaluative belief" can thus have the structure of several different assertions. Here we have mentioned only causal relations and behavior. What makes these assertions indicate an evaluative belief is the evaluative connotation of some words. Without this evaluative connotation the assertions cannot be seen as indicating "evaluative beliefs." Assertions, representing evaluative beliefs, have sometimes been used purposely by researchers to avoid socially desirable answers.

With this we conclude our introduction to the concepts-by-intuition that fall under the subjective variables category. These assertions are based on information that can be obtained only from respondents, whose views cannot be verified because they are personal views that represent subjective variables.

2.4.2 Objective variables

By *objective variables* we mean variables for which in principle information can also be obtained from a source other than the respondent. One could think of administrations of towns, hospitals, schools, and so on. Commonly these variables concern factual information such as behavior, events, time, place, quantities, procedures, demographic variables, and knowledge.

Behavior concerns present and past actions or activities of the respondent him/herself (Sudman and Bradburn 1983; Smith 1987). Structures 2 and 3 with an activity indicating predicator (D) can be used to specify the behavioral assertion (a_b). The typical form for structure 2 is (rDy), which means that the subject or respondent does or did y or with structure 3 it is (rD). It will be clear that the structure of this assertion is the same as the structure for an action tendency.

However, its content differs fundamentally from the latter. Action tendencies deal with subjective matters (likely future behavior) while behavior is factual and in principle controllable. Examples 2.51 and 2.52 show this structure:

2.51 *I am studying English.*
2.52 *I was cleaning.*

In example 2.51 "I" stands for (r), "am studying" is the action indicating predicator (D), and "English" is the direct object (y). In example 2.52 the subject "I" is again the respondent, while the action that indicates the predicator is "was cleaning." In this case there is no direct object. Therefore it is an example of structure 3, while example 2.51 employs structure 2.

The facts mentioned in these assertions can in principle be checked by observation as opposed to subjective variables such as, for example, a behavioral intention ("a person is planning to go to the hospital").

An *Event* represents another example of an objective variable. It pertains to other people's actions that are presently ongoing or had occurred in the past. The structure of this assertion (a_{ev}) is the same as the previous one except that the subject is not the respondent and therefore it is (xDy) or (xD). Examples of assertions characteristic to this concept are 2.53 through 2.55:

2.53 *My brother is studying.*
2.54 *My mother had washed the clothes.*
2.55 *The shopping center has been burglarized.*

In example 2.53 (x) is "my brother," "is studying" stands for the action predicator (D), and there is no direct object that makes it an example of structure 3. Example 2.54 belongs to structure 2. It has "my mother" as (x), the action predicator (D) is "had washed," and (y) is "the clothes." Example 2.55 belongs again to structure 3 with an adverbial as extension: (x) is "the shopping center," and "has been burglarized" represents the action predicator (D).

Demographic variables are used in nearly all surveys and are mentioned in all attempted classifications of data (Oppenheim 1966; Sudman and Bradburn 1983; Converse and Schuman 1984; Smith 1987; Bradburn and Sudman 1988). We represent demographic variables by the assertion (a_d). Structure 1 should be used for demographics (xId). The subject (x) is frequently the respondent or another person in his/her environment, but it differs from a judgment by the fact that the subject complement is limited to certain factual topics such as the respondent's gender, age, or occupation, summarized by (d). Examples 2.56 and 2.57 illustrate these assertions:

2.56 *I am 27 years old.*
2.57 *I am married.*

It will be clear that the structure of these assertions is the same as the one of an evaluation or a judgment. The only difference is the type of subject complement specified.

There are also assertions which relate to *knowledge* (a_k). They could ask, for example, who the 35th president of the United States was or which Russian leader had sent nuclear missiles to Cuba. The assertion to answer the first request would be structure 1, and the second would be structure 2. Examples 2.58 and 2.59 are examples of this type:

2.58 *Kennedy was the 35th president of the United States.*

2.59 *The Russian leader Khrushev had sent nuclear missiles to Cuba.*

The structure of these assertions requires historical or political knowledge of the respondent. These knowledge assertions can have any structure for objective variables. Our first example reads as follows: "Kennedy" is (x), "was" stands for the link verb predicator (I), and "the 35th president of the United States" is the subject complement (sc). Therefore the structure can be modeled as a_k=(xIsc).

The second example has the structure of an event: (x) is "the Russian leader Khrushev," the action predicator (D) is "had sent" and (y) is "nuclear missiles" while "to Cuba" is an adverbial.

Often information is requested in surveys about *time and place* of behavior or events. In an assertion this information is presented by adverbials indicating time/place-specific components. Examples 2.60 and 2.61 illustrate this:

2.60 *I worked yesterday.*

2.61 *I stayed in a hospital in Chicago.*

Thus, the focus shifts in these two examples from the act, to the specification of the time (2.60) or the place (2.61).

The first assertion is a time assertion a_{ti}=(rDti). It reads as follows: "I" is (r), "worked" is the behavioral connector (D), and "yesterday" is the time adverbial. The second example is a place assertion a_{pl}=(rDpl), where "I" is (r), (D) is "stayed," "in a hospital in Chicago" constitutes two place adverbials, indicated in the structure of the assertion by (pl). The reader may note that it is structure 3 that applies to time and place assertions.

Quantities can also be specified by structure 2. The assertion that can be formulated for quantities has the form (a_{qu}=rDqu). Example 2.62 illustrates this:

2.62 *I bought 2 packs of coffee.*

In example 2.62 "I" stands for (r), "bought" is (D), and "2 packs of coffee" is (y) the direct object. "2 packs" indicates the quantitative information (qu) and the modifier "of coffee" specifies the substance.

Assertions concerning *procedures* can be formulated similarly using structure 3 as (a_{pro}=(xDy, pro). An example is 2.63:

2.63 *I go to my work by public transport.*

"I" is (x), "go to" is (D), "my work" is a direct object (y), and "by public transport" is an adverbial that indicates the procedure (pro).

2.4.3 In summary

In this review most concepts-by-intuition used in the survey literature have been described. In these sections we have tried to make the structure of these assertions explicit. Table 2.1 summarizes them.

Table 2.1: The basic structures of simple assertions

Concepts		Structure 1 xIsc	Structure 2 xPy	Structure 3 xP
Subjective variables				
Evaluation	a_e	xIe	–	–
Importance	a_i	xIi	–	–
Values	a_i	vIi	–	–
Feelings	a_f	xIf	xFy or xPf	–
Cognitive judgment	a_j	xIc	–	–
Causal relationship	a_c	xIca	xCy	–
Similarity relationship	a_s	xIs	xSy	–
Preference	a_{pr}	xIpr	xPR y(z...)	–
Norms	a_n	–	o(H+I)y	o(H+I)
Policies	a_p	–	g(H+I)y	g(H+I)
Rights	a_{ri}	xIri	xHRy	–
Action tendencies	a_t	–	rFDy	rFD
Expectations of future events	a_{ex}	–	xFDy	xFD
Evaluative belief	a_{eb}	–	xP_ey or xPy_e	xP_e
Objective variables				
Behavior	a_b	–	rDy	rD
Events	a_{ev}	–	xDy	xD
Demographics	a_d	xId	–	–
Knowledge	a_k	xIsc	xPy	xP
Time	a_{ti}	–	–	xDti
Place	a_{pl}	–	–	xDpl
Quantities	a_{qu}	–	xDqu	–
Procedures	a_{pro}	–	–	xDy, pro

We are aware that these concepts can also be expressed in different ways, however the purpose of this exercise was *to suggest structures where there is no doubt that the generated assertions indicate the desired concepts-by-intuition.* Table 2.1 shows that some concepts can be presented in assertions with

different structures. Further research is required to determine whether there is a difference in the responses for different types of linguistic structures.

The table can also be used to detect which kind of concept has been used in assertions applied in practice. This is a more difficult task because there are different extensions of these simple sentences. Some of these extensions or variations in formulations will be discussed in the following sections. These extensions will make the coding of requests for answers more difficult than the production of proper assertions for the presented concepts.

2.5 ALTERNATIVE FORMULATIONS FOR THE SAME CONCEPT

Grammar provides a variety of different ways of expressing the same proposition; this is what some linguists call "allosentences," which are found in particular syntactic constructions and certain choices between related words (Lambrecht 1995: 5). We can select a form that is appropriate according to where we want to place the emphasis. Emphasis is placed mostly on new information in a sentence but it also might be desirable to place it on parts that are assumed to be known, or otherwise known as background information (Givon 1984: 251; Lambrecht 1995: 51). Some grammatical constructions that are syntactically different have the same content (Givon 1984; Huddlestone 1988; Lambrecht 1995), but they add emphasis to different parts of the sentence. The constructions studied in this section occur frequently in survey requests and are called *active/passive* and *existential*.[4] We begin with an example of the *active voice*:

2.64 *New laws have changed the position of black people.*

This assertion (2.64) means that the subject "new laws" is the so called "agent" and the direct object "the position of the black people" is the "patient" or "undergoer" of the change. If one reads this sentence the emphasis seems to be on "new laws." If we change this assertion into the passive voice, we obtain example 2.65:

2.65 *The position of black people was changed by new laws.*

In the passive voice (2.65) the emphasis is on the former patient "the position of black people" which becomes the grammatical subject while the agent becomes the adverbial "by new laws." To transform the passive assertion of example 2.65 into an existential construction, we need to put the word "there" at the beginning of the sentence and we obtain example 2.66:

---

[4] Linguists also mention the "cleft construction"; this means that a single sentence is divided in two parts (cleft), each with its own predicator while one is highlighted. To illustrate this we give an example: "It was new laws that changed the position of the black people;" or "it was the position of the black people that changed new laws." According to our experience, such constructions do not occur frequently in requests for answers, therefore we discuss them only briefly in Chapter 3.

2.66 *There has been a change in the position of black people due to new laws.*

In example 2.66 the subject "the position of black people" is substituted by "there," and the word "change" is highlighted.

The different formulations in examples 2.64 through 2.66 express the same concept, which is a relationship. But they emphasize different parts in the sentence. However, it is not clear how respondents react when they are confronted with these different forms. It can be that they pay attention only to the concept. On the other hand, they also can answer differently to the various grammatical forms. This is an issue that requires further empirical studies.

2.6 EXTENSIONS OF SIMPLE SENTENCES

Until now we have focused on the basic structure of assertions; however, in reality assertions have a lot of variation. They are expressed in sentences much longer than have been studied so far. Often indirect objects, modifiers, or adverbials are added to the simple sentences. In this section we will address this issue.

2.6.1 Adding indirect objects

An additional component that can be added to the simple sentences without changing the concept represented in the sentence is an indirect object. Examples 2.67 and 2.68, given previously, illustrate this:

2.67 *Honesty is very important to me.*

2.68 *He bought an apartment for his mother.*

These examples show that adding the indirect object component "to me" or "for his mother" does not change the concept the assertion refers to. The same holds true when modifiers are added to a sentence.

2.6.2 Adding modifiers

As we stated previously, a *modifier* gives a specification to a noun. The modifiers can be placed before and after the noun and be related to the subject and to the object. Previously some examples of this type were given as significant (2.14, 2.15, and 2.16). These examples demonstrated that normally modifiers are no complication for the assertions. Whether we say "Clinton" or "the popular Clinton" or "dirty clothes" instead of just "clothes" will rarely lead to serious interpretation problems for most respondents. However, the modifiers can be longer; for example, "the most famous president of the United States" can be written instead of just "president." If both the subject and the object have a modifier, the meaning of the sentence can become quite complicated. Therefore they should be used with moderation; they can be helpful but they can also lead to rather complex sentences.

2.6.3 Adding adverbials

In contrast to the previous additions to sentences discussed, adding an adverbial will *change the concept* most of the time. The reason is that adding such an adverbial implies providing specific information that becomes the focus of the attention (Givon 1990: 712). Structure 3 sentences often contain adverbial components or just an adverb. For example

2.69 *He worked full time.*

In this sentence the emphasis is not on whether he does or does not works, but, on the fact that he worked "full time" and implicitly not "part time." So this assertion expresses something about his work and is still an assertion expressing demographic information. But in the following example (2.70) a change in concept takes place:

2.70 *He worked hard.*

Adding the adverb "hard," the attention shifts from working or not working to "hard" or "lazy working," which expresses a cognitive judgment of one person about another. Take note that the concept has shifted from an objective variable to a subjective one. Examples 2.71 and 2.72 display concept shifts from objective to subjective variables, where the adverb has an evaluative (2.71), followed by an emotive (2.72) connotation:

2.71 *He worked very well.*

2.72 *He worked with pleasure.*

These sentences express an evaluative belief (2.71) and a feeling (2.72).

In section 2.4.2 we gave other examples of assertions for which the concept of intuition changed by adding adverbials with respect to time, place, quantity, or procedure.

2.7 USE OF COMPLEX SENTENCES

So far we have discussed only simple sentences or clauses, with only one subject and predicator. In contrast complex sentences consist of more subjects and predicators, because of additional clauses. Examples 2.73a–2.73d illustrate assertions with complex clauses (where subj.=subject and pred=predicator)

2.73a	*Immigrants*	*who come from Turkey*	*are generally friendly.*
	Subj. 1	Subj.2 Pred.2	Pred.1
2.73b	*Abortion is*	*permitted if a*	*woman is raped.*
	Subj.1	Pred.1	Subj.2 Pred.2
2.73c	*While driving home*	*he had*	*an accident.*
	Pred. 1	Subj.2	Pred.2
2.73d	*The Social Democrats*	*performed better than*	*the Conservatives.*
	Subj.1	Pred.1	Subj.2

Examples 2.73a and b display two subjects and two predicators, as the definition requires. The reader may note that example 2.73a displays a complex clause where the second clause "who came from Turkey," is embedded or nested in the first one. In the other examples the second clauses follow the first clause (2.73b–273d). There are thus two ways of joining sentences: linearly or embedded.

In example 2.73c the first subject is missing but implied, since it is the same as in the main clause "he." Example 2.73d omits the second predicator, and it seems to be implied since it has the same meaning as the first. The sentence would read correctly with "than the Conservatives *did/performed*."

Complex sentences can be built from coordinate clauses by linking them with coordinating conjunctions such as "and," "but," or, "neither," in which case they are considered the "same" level and called *main clauses*. Coordinate clauses can become rather problematic in survey research, as we will discuss in the following chapter, but from a linguistic perspective this type of complex sentence is clear and therefore we will concentrate on subordinate clauses in the next sections.

Examples 2.73a–2.73d expressed complex clauses consisting of a main clause and a subordinate clause. If the *subclauses* that are linked to the rest of the sentence by subordinating conjunctions ("who" 2.73a; "if" 2.73b; "when" 2.73c; "than" 2.73d) are omitted, then the remaining part is the main clause: "Immigrants are generally friendly" (2.73a), "Abortion is permitted" (2.73b), "He had an accident" (2.73c) or "The Social Democrats performed better in the elections" (2.73d).

At the beginning of this chapter we discussed the grammatical elements of simple clauses, which were the subject predicator, direct object, indirect object, object complement, and adverbial. All these parts of sentences *except the predicator* can also be expressed by a subordinate clause in complex sentences (Koning and Van der Voort 1996: 84–90).We will illustrate this by an example:

2.74a Problems in Turkey caused emigration to Europe.

2.74b Problems in Turkey caused that Turkish people emigrated to Europe.

Example 2.74a is a simple clause of structure 2 (subject + adverbial + predicator + direct object + adverbial). In example 2.74b the direct object + adverbial "emigration to Europe" are substituted by a subordinate clause "that Turkish people emigrated to Europe." It is thus characteristic of complex sentences that a component of a simple sentence is substituted by a subclause.

Having provided the necessary linguistic background to understand complex assertions, we will study them in more detail in the next sections.

2.7.1 Complex sentences with no shift in concept

The simple expression "emigration to Europe" (2.74a) has been substituted by the more elaborate subclause "that Turkish people emigrated to Europe" (2.74b). This example illustrates that the meaning of the two assertions is

similar but that the second formulation (2.74b) is much longer than the first. The subject (x) of assertion 2.74b is "problems in Turkey," which is followed by the causal predicator (C) "caused" and where the (y) is mentioned consisting of another assertion [a behavioral one (a_b)] which reads as follows: "Turkish people (x) emigrated (D) to Europe (y)." This interpretation of the assertion can be verified by asking: "what did the problems in Turkey cause?" Example 2.74 illustrates that the object in the previous assertion is substituted by another one. This complex assertion can be written more formally as (xRa_b). In this case both assertions, the simple one and the complex one, represent the same concept (a relationship), but the second assertion (2.74 b) is much more complex than the first (2.74a). Whether complexity of assertions makes a difference for the respondent is still an empirical question.

2.7.2 Complex sentences with a shift in concept

Substitutions of the sentence components y or x that represent different concepts can be employed for nearly all assertions discussed previously. Above we gave an example where the complex and the simple assertion represented the same concept (2.74a,b). There are, however cases where the two concepts present in the complex assertion are different. Below we provide several examples. A common example is the judgment of a relation. The relational assertion used (2.75) is one we have seen before:

2.75 Problems in Turkey caused emigration to Europe.

A judgment of this relation a_r is formulated in examples 2.76a and 2.76b:

2.76a That problems in Turkey caused emigration to Europe is quite certain.

2.76b It is quite certain that problems in Turkey caused emigration to Europe.

The equivalent meaning of the two linguistic variants of example 2.76 consists of the main sentence "(it) is quite certain" and the subordinate clause "that problems in Turkey caused emigration to Europe." However, the structure of these assertions (2.76a,b) is not (xIc) but (a_rIc). Therefore the assertion (a_r) "problems in Turkey caused emigration to Europe" takes the place of the subject x, the predicator is "is" and the subject complement is "quite certain." By asking oneself "what is quite certain?" (2.76b) we can conclude that the subject "it" can be substituted by "that problems in Turkey caused emigration to Europe." The phrasing of example 2.76a is a clearer example of this type of assertion, but 2.76b can be classified in the same category. Krosnick and Abelson (1991) discuss the use of such complex assertions, in particular the certainty about an opinion as a measure of *opinion strength*.

Evaluations can also be formulated with respect to assertions. Example 2.77 illustrates this point:

2.77 It is bad that the problems in Turkey caused emigration to Europe.

In 2.77 the structure is (a_rIe) and therefore this is an evaluation of an assertion.

In the same way, importance judgments can be formulated (2.78):

2.78 It is important to me that the Conservative Party continues to be strong.

While "that the Conservative Party continues to be strong" is an assertion on its own (a_e), in this statement an assertion concerning importance is formulated (a_eIi). Krosnick and Abelson (1991) discuss the requests using this type of complex assertion also as measures of "opinion strength."

Feelings can be formulated in the same way. Example 2.79 begins with the judgment (a_j):

2.79 Most immigrants are hard-working.

For this assertion (2.79) we can formulate an assertion for a feeling (2.80):

2.80 I am glad that most immigrants are hard-working.

In Example 2.80 the subject complement "glad" is extended by the subclause "that most immigrants are hard-working," which functions as an adverbial within the subject complement and could be paraphrased by "about the hard-working immigrants". The structure of 2.80 is (sIf a_j).

As a last example we show how a right is formulated on the basis of an evaluative belief in order to demonstrate the general approach. The evaluative belief a_{eb}=(xD_ey) is illustrated by example 2.81:

2.81 Immigrants exploit our social security system.

The assertion of a right (a_{eb} IR y) can then be formulated in example 2.82:

2.82 It is unacceptable that immigrants exploit our social security system.

These examples showed how this approach is used in general. Please keep in mind the complexity that can result. It is especially true when subject x and subject complement y are both substituted by assertions. Therefore, we do not recommend them for survey research, even though there is evidence that they are quite common in research practice. We did not include complex assertions in Table 2.1; however, the reader should be aware of that any x and y mentioned in Table 2.1 can be replaced by a complete assertion. We did not include this option in the table because the main clause will still indicate the same concept whatever the concept in the subclause may be.

2.7.3 Adding conditions to complex sentences

Another commonly used extension of an assertion is the *use of conditionals*. They express the circumstances under which something stated in the main clause can occur. They can express real or unreal things (Yule 1998: 123–152).

In survey requests both types of conditionals are used. Examples 2.83 and 2.84 show assertions with a real conditional:

2.83 *Abortion is permitted if a woman is raped.*

2.84 *If immigrants work harder, they will be as well off as our people.*

Example 2.83 clearly expresses a woman's right to abortion if she has been raped. Formally, it can be summarized as (xHRy |con) where "|con," indicates the condition. Example 2.84 indicates a future event depending on the prior occurrence of the "if" clause: ((xFDy) |con).

Also, sometimes unreal events are expressed in complex sentences. This is shown by examples 2.85 and 2.86:

2.85 *If immigrants worked harder, they could be as well off as our people.*

or

2.86 *If immigrants had worked harder, they could have been as well off as our people.*

Clearly, the evaluative state ("they could be as well off") in example 2.85 is unlikely because the "if" clause, describing the willingness of the immigrants to work harder, is in the past tense. In example 2.86 the evaluative state in the main clause ("they could have been as well off as our people") is impossible only because the "if" clause expressed in the past perfect implies that the condition was not fulfilled.

It is difficult to understand what concept is represented by these assertions (2.85 and 2.86). Our best guess is that they represent two concepts: a relationship suggesting that hard-working immigrants will be as well off as our people and the cognition that immigrants did not work hard, suggesting it is their own fault that they are in a worse situation. If researchers have difficulty in understanding what is being asserted by such assertions, it is very likely that the respondents will also be confused, which can lead to rather low reliability and validity scores. Nevertheless, assertions like this are not uncommon in survey research, as demonstrated in Table 1.1, item 3 (Chapter 1).

2.8 SUMMARY AND CONCLUSIONS

This chapter has discussed three basic assertion structures that can be used to represent most concepts-by-intuition from the social sciences. We also have indicated how the most commonly applied concepts-by-intuition in survey research can be expressed with certainty in assertions specifying these concepts. These rules are summarized in Table 2.1. The knowledge summarized in Table 2.1 can be used in two ways.

The table can be used to specify an assertion for a certain type of concept according to the criteria specified in Table 2.1. For example, if we want to specify an evaluation about immigrants, we know that the structure of the sentence should be (xIe). Therefore, we can formulate a statement such as

"immigrants are good people." If we want a feeling (xIf), we can write "immigrants are in general friendly." If we want a cognitive judgment (xIc), the statement is: "immigrants are in general hard-working." If we want to formulate a cognition concerning the reasons why immigrants come here, the structure is (xRy), and a possible assertion would be "Problems in their own country cause emigration to Europe." In the same way assertions can be formulated for any other concept.

Table 2.1 can also be used to detect which kind of concept has been used in assertions applied in practice. The elementary structures of the assertions refer in a simple way to the concepts mentioned. However, we have to say that the assertions can be made rather lengthy by use of complex sentences, subordinate clauses, time and place statements, and conditions. The use of such complicating possibilities can cause that the meaning of the assertions becomes much less intuitively clear than in the simple assertions. It is an interesting topic of further research to study what kinds of complications are possible without shifting the meaning of the request or assertion for the respondent.

EXERCISES

1. Formulate assertions concerning the Al Qaida network in terms of
   a. A cognition
   b. An evaluation
   c. A feeling
   d. A relationship
   e. A norm
   f. A policy
   g. A behavioral intention
   h. A future event
   j. A behavior
2. Guttman (1981, 1986) suggested the use of facets designs to create measurement instruments. The facet design presented in the table 2.2 below has been developed in discussions between the members of the International Research Group on Methodological and Comparative Survey Research (Saris 1996). The purpose of this table is to show that one can systematically formulate statements for different concepts-by-intuition mentioned above the columns. This can be done for the different aspects of life indicated in the rows of the table.
   a. Can you specify an assertion for each cell of the table using our procedure?
   b. Can the items in the rows be used to measure a concept-by-postulation?
   c. Can the items in the columns be used to measure a concept by postulation?

Table 2.2: Facet design for ethnocentrism

Aspects of life	Judgment	Evaluative belief	Evaluation out/ingroup	Norms	Policies
Way of life					
Religion					
Economic					
Political					
Personal					

3. Measurement instruments are not always carefully developed in research. Examples are the measurement instruments presented in Table 2.3.
   a. Indicate where the different items of Table 2.3 fit in the facet design presented in exercise 2.
   b. Can these items be used to form a concept-by-postulation?

Table 2.3: Operationalization of subtle and symbolic racism

Items	
1	Os living here should not push themselves where they are not wanted.
2	Many other groups have come here and overcame prejudice and worked their way up. Os should do the same without demanding special favors.
3	It is just a matter of some people not trying hard enough. If Os would only try harder, they could be as well off as our people.
4	Os living here teach their children values and skills different from those required to be successful here.
5	How often have you felt sympathy for Os?
6	How often have you felt admiration for Os?
7	How different or similar do you think Os living here are to other people like you?
7a	In the values that they teach their children
7b	In the religious beliefs and practices
7c	In their sexual values or practices
7d	In the language that they speak
8	Has there been much real change in the position of Os in the past few years?
9	Generations of slavery and discrimination have created conditions that make it difficult for Os to work their way out of the lower class.
10	Over the past few years Os have received less than they deserve.
11	Do Os get much more attention from the government than they deserve?
12	Government officials usually pay less attention to a request or complaint from an O person than from our people.

O stands for member(s) of the outgroup, which includes any minority group member(s).

4. For the ESS pilot study a proposal was made by Shalom Schwartz to measure basic human values. The suggestion for one of the items was as follows:

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Put an X in the box to the right that shows how much the person in the description is like you.

HOW MUCH LIKE YOU IS THIS PERSON?						
	Very much	*much*	*some what*	*a little*	*very little*	*not at all*
Thinking up new ideas and being creative is important to her. She likes to do things her own original way.						

   a. Specify the concepts that are present in this survey item.
   b. Check if these assertions represent the concepts they are supposed to represent.
   c. If needed, try to improve the survey item.
5. Check over your own questionnaire from Chapter 1 exercises to see
   a. What the concepts-by-intuition behind your requests are
   b. If your assertions indeed reflect these concepts-by-intuition

This Page Intentionally Left Blank

CHAPTER 3

The formulation of requests for an answer

So far we have discussed the distinction between concepts-by-postulation and concepts-by-intuition (Chapter 1). We also studied the way basic concepts-by-intuition used in survey research can be expressed in assertions (Chapter 2). In this chapter we will continue with the discussion of how assertions can be transformed into requests for an answer.

While the choice of the topic of requests and the selection of concepts are determined by the research goal of the study, the formulation of questions or requests for an answer, as we call them, provides much more freedom of choice for the designer of a questionnaire. A great deal of research has been done on the effect of different ways in which requests are formulated (Schuman and Presser 1981; Molenaar 1986; Billiet et al. 1986). Also a considerable part of the literature is devoted to devise rules of thumb for the wording of survey items (Dillman 2000; Converse and Presser 1986). On the other hand, relatively little attention is given to the linguistic procedures for the formulation of requests for answers in the survey literature.

Therefore, in this chapter we will discuss different procedures to transform the assertions, discussed in the last chapter, into requests for an answer. In doing so we make use of a large body of research in linguistics, especially Harris (1978), Givon (1990), Weber (1993), Graesser et al. (1994), Huddlestone (1994), Ginzburg (1996), and Groenendijk and Stokhof (1997). The rules for the formulation of requests for an answer in English will be presented in the text, but in general these formulation rules also apply in other languages such as German, Dutch, French, and Spanish. If they are different in one of the languages just mentioned, it will be indicated in the appropriate section by a note.

3.1 FROM CONCEPTS TO REQUESTS FOR AN ANSWER

The term "request for an answer" is employed, because the social science research practice and the linguistic literature (Harris 1978; Givon 1990; Weber 1993; Graesser, et al. 1994; Huddlestone 1994; Ginzburg 1996; Groenendijk and Stokhof 1997; Tourangeau et al. 2000) indicate that requests for an answer are formulated not only as requests (interrogative form) but also as orders or

instructions (imperative form), as well as assertions (declarative form) that require an answer. Even in the case where no request is asked, and an instruction is given or a statement is made, the text implies that the respondent is expected to give an answer. Thus the common feature of the formulation is not that a request is asked but that an answer is requested.

If an assertion is specified for a concept, the simplest way to transform it into a request for an answer is to add a prerequest in front of the assertion. This procedure can be applied to any concept and assertion. Imagine that we want to know the degree of importance that the respondents place on the value "honesty" as in examples 3.1a–3.1d:

3.1a *Honesty is very important.*
3.1b *Honesty is important.*
3.1c *Honesty is unimportant.*
3.1d *Honesty is very unimportant.*

To make a request from these assertions prerequests can be added in front of them, as for example:

3.2a *Do you think that honesty is very important?*
3.2b *Do you think that honesty is important?*
3.2c *Do you think that honesty is unimportant?*
3.2d *Do you think that honesty is very unimportant?*

Using such a prerequest followed by the conjunction "that" and the original assertion creates a request called an *indirect request.* The choice of one of these possible requests for a questionnaire seems rather arbitrary or even incorrect as this specific choice of the request can lead the respondent in that direction. Therefore a more balanced approach has been suggested:

3.2e *Do you think that honesty is very important, important, unimportant or very unimportant?*

In order to avoid such an awkward sentence with too many adjectives it is advisable to substitute them with a so called WH word like "how", as in the example below:

3.2f *Can you specify how important honesty is ?*

This is also an indirect request with a prerequest and a subclause that started with a WH word and allows for all the assertions specified above (3.1a–3.1d) as an answer and other variations thereof.

Instead of indirect requests *direct requests* can also be used; the most common form is an interrogative sentence. This type of request is normally called a "request" or a "direct request." In this case the request can be created from an assertion by the inversion of the (auxiliary) verb with the subject component. The construction of direct requests by the inversion of the verb and

subject component is quite common in many languages but also other forms can be used.[1]

Let us illustrate this with another example "vote intention," which is a behavioral intention. It can be formulated in an assertion of structure 2 (Chapter 2) with an auxiliary verb indicating that the action will be in the future. This leads to the following possible assertions:

3.3a *I am going to vote for the Social Democrats.*
3.3b *I am going to vote for the Republicans.*

One can transform these assertions into direct requests by inverting the auxiliary verb and the subject, while a simultaneous change from the first to the second person for the subject is also necessary. It leads to examples 3.4a and 3.4b:

3.4a *Are you going to vote for the Social Democrats?*
3.4b *Are you going to vote for the Republicans?*

Here the requests can be seen as "leading" or "unbalanced" because they have only one possible answer option. It could be expected that a high percentage of respondents would choose this option for this reason. Therefore, the requests can be reformulated as follows:

3.5 *Are you going to vote for the Social Democrats or the Republicans?*

A different way to formulate a direct request is also possible. We have seen that the point of interest is the party preference. Therefore one can also omit the names of the parties in the request and place a "WH word" in front of the request. In this case one is interested in the party preference that people intend to vote for. Hence, the words "Social Democrats" and/or "Republicans" are omitted and the WH word "What" followed by the more general term "party" is placed in front of the request, which leads to the following request for an answer:

3.6 *What party are you going to vote for?*

The advantage of this format is that it is not biased to a political party, by mentioning only one possibility or giving first place to a party in the sentence word order.

This overview shows that two basic choices have to be made for formulating a request for an answer: the use of direct or indirect requests and whether to use WH words. The combination of these two choices leads to four different

[1] In French it is also possible to place the question formula "Est-ce que" in front of a declarative sentence to indicate the interrogative form. Spanish, for instance, constitutes an exception since one does not have to use the inversion, as rising intonation of the declarative form is already enough. Interrogatives are indicated by two question marks, one in front of the clause (¿) and the other at the end of the clause (?).

types of requests, which we will describe below; however, before doing so, we will discuss one other distinction.

Besides the interrogative form, two other grammatical forms of a request for an answer are also possible, the second of which is the imperative form. In its basic form the request consists only of an instruction to the respondent, as for example:

3.7 *Indicate the party you are going to vote for.*
1. Republicans
2. Social Democrats
3. Independents
4. don't know

Example 3.7 illustrates that requests for an answer can also have another grammatical form than an interrogative one. Example 3.7 is colloquially known as an instruction or in grammatical terms it is referred to as an "imperative." This is another example of a direct request for an answer.

The third grammatical form, a declarative request, is not possible as a direct request for an answer but only as an indirect request. Illustrations are examples 3.8 and 3.9. Both examples have a declarative prerequest, followed by a WH word and an embedded interrogative query, and example 3.9 displays an interrogative prerequest with an embedded declarative query:

3.8 *I would like to ask you what party you are going to vote for.*
3.9 *Next we ask you whether you are going to vote for the Republicans or the Democrats.*

Although these are statements from a grammatical perspective, it is commonly understood that an answer to the embedded interrogative part of the sentence is required.

This overview shows that many different requests for an answer can be formulated to measure concepts like "the importance of the value of honesty" or "vote intention." However, it is important to note that whatever the request form used, there is no doubt that all these requests measure what they are supposed to measure. Therefore there is no real difficulty with making an appropriate request for a concept if the assertions represent the concept of interest well. It only points further toward the importance of the previous chapter in the whole process of designing requests.

3.2 DIFFERENT TYPES OF REQUESTS FOR AN ANSWER

After the introduction of the basic forms of requests for an answer we will now examine how the different requests can be formulated in more detail. Table 3.1 summarizes the different types of requests for answers occurring in survey interviews according to their grammatical form and use in survey research. The table shows that not all theoretically possible combinations can be formulated; direct instructions with WH words are impossible because they automatically

become indirect requests. Indirect requests with embedded declarative statements are also only possible without WH words because these subclauses have to begin with the conjunction "that" to be considered as declarative. We will discuss and illustrate the remaining options starting with direct requests.

Table 3.1: Different types of requests for answers

WH words	Direct request	Indirect request	
WH word not present	Direct instruction	Imperative + interrogative	-
	Direct request	Interrogative + interrogative	Interrogative + declarative
	-	Declarative + interrogative	-
WH word present	-	Imperative + interrogative	-
	Direct request	Interrogative + interrogative	-
	-	Declarative + interrogative	-

3.2.1 Direct request

We have already given several examples of direct requests. Therefore we will be relatively brief about this type of request. We start with the direct instructions.

3.2.1.1 The direct instruction

As was mentioned above, the direct instruction consists of a sentence in the imperative mood. This form is not so common in colloquial language but is quite common in written questionnaires or other written formats that one has to fill out for the government and other agencies. In this case no request is asked but just an instruction is given at what one has to do. Very common examples in very short form appear on application forms starting with

3.10 *First name:* .....
3.11 *Family name:* .....
3.12 *Date of birth:* .....

Most people understand these instructions and write the requested information on the dots. In the case where less elementary information is asked for, more instruction has to be given and full sentences are used. Very common examples include the following:

3.13 *Select from the specified possibilities the one that fits your family situation best:*

1. *Single*
2. *Single with child(ren)*
3. *Married without child(ren)*
4. *Married with child(ren)*

Another example taken from social research is

3.14 *Indicate your opinion about the presidency of Clinton by a number between 0 and 10, where 0 means very bad and 10 very good.*

Another example from mail questionnaires is also

3.15 *Indicate your opinion about the presidency of Clinton by a cross on the line below:*

Very bad *Very good*

I_____I_____I_____I_____I_____I_____I

In interviewer-administered questionnaires these instructions are very uncommon and a more polite request form is preferred.

3.2.1.2 The direct request

As was mentioned before, a direct request contains an inversion of the (auxiliary) verb and the subject. The word order thus is changed. To illustrate we start with the following assertions:

3.16 *The position of blacks has improved recently.*

3.17 *I prefer the Republican Party above the Democratic Party.*

From the first example a direct request can be formed by putting the auxiliary verb "to have" in front of the subject:

3.18 *Has the position of blacks improved recently?*

This is called an *inversion* because the position of the auxiliary verb and the subject are reverted. Example 3.17 in request form could be

3.19 *Do you prefer the Republican Party above the Democratic Party?*

Here the auxiliary is not the verb " to be " but "to do" and the subject is "you."

Direct requests can also be formulated as we have seen using WH words. So let us look at this approach a bit more carefully here. The linguistic literature (Lehnert 1978; Harris 1978; Chisholm el al. 1984; Givon 1990; Huddleston 1994)

treats requests introduced by words like "who," " whose," " whom," " why," "where," " when," "which," "what," and "how" as a specific type of request. In English they are also called *WH-interrogatives*. According to Givon (1990: 739) a specific request is used when the researcher and the respondent share some knowledge but a particular element is unknown to the former and this component is the asked for element. This element is replaced by an interrogative word and constitutes the focus of the request, which means that the researcher wishes to draw special attention to it. In the previous section we have seen that one reason to use WH words in front of direct requests is to avoid leading or unbalanced requests. The example was:

3.20 What party are you going to vote for ?

The advantage of this form is that one cannot be blamed for giving an advantage to one of the two parties by mentioning only one or mentioning one party as the first in the request. However, there are other reasons to use WH-fronted requests. Fronting the word "when" realizes the request to ask for the *time* when the change occurred:

3.21 When did this change occur?

By asking a "where" request one can determine the *place* where the change occurred:

3.22 Where did this change occur?

Finally by asking a "why" request one can determine the *cause* or *motives* of the change:

3.23 Why did this change happen ?

These examples show that WH requests are used to ask a specific element. Grammatically the WH word stands in the beginning and mostly also a switching of the subject auxiliary occurs. The nature of the WH word determines the missing element. This fronting of the request word occurs in many languages with slight variations.[2]

There are more WH words that can be used. We shall return to this issue in the following sections.

3.2.2 Indirect request

Indirect requests for an answer necessitate further discussion because they can come in many different forms as indicated in Table 3.1. We will discuss the different forms in sequence. Because it is rather natural to use in these indirect requests WH words like "whether" or "what" or "which" we will not completely separate the two types of requests but give examples with and without WH words.

2 In French additionally the interrogative form "est-ce que" might be put after the specific question word.

3.2.2.1 Imperative – interrogative requests

As stated previously, an indirect request consists at least of two parts. The first part is the main clause and contains mostly a prerequest by which the researcher indicates a desire to obtain information about something in a neutral or polite way. The queried topic is then embedded and frequently presented by the second part, a subordinate clause. When a neutral prerequest is formulated as an order in the imperative mood, words like "tell," "specify," "indicate," show," and "encircle," are characteristic. The researcher signals by the use of these words to the respondent to inform him about something. The topic the researcher wants to know can, for instance, be specified by another main clause that is formulated as direct request. Examples 3.24 and 3.25 serve as an illustration:

3.24 *Tell me, did you leave school before your final exam.*

3.25 *Specify, what were your reasons for leaving school before your final exam.*

In the first example (3.24) the imperative is followed by a direct query characterized by the inversion of the auxiliary verb and the subject "did you." In the second example (3.25) the imperative is followed by a direct specific query initiated by the WH word "what" and the inversion of the auxiliary and subject "are the reasons." Note that in both requests the requests are main clauses.

Requests can also be formulated using subordinate clauses. Some examples (3.26 and 3.27) are provided below:

3.26 *Tell me, if you left school before your final exam.*

3.27 *Specify what were your reasons for leaving school before your final exam.*

Both examples show that the requests for answers are formulated as subordinate clauses and there is also no inversion present as is the case of direct requests. Example 3.26 has "if" as the conjunction of the subordinated clause. Since the subordinate clauses after "if, which, what, whether, who etc." function as indirect or embedded interrogatives we call this kind of requests (3.24, 3.25, 3.26 and 3.27) of the form imperative + interrogative.

Since the researcher wants to elicit information from the respondent the communication requires some politeness in the interaction. In order to make the prerequests in *imperative mood* more polite researchers frequently add the word "please" as in 3.28 and 3.29:

3.28 *Please, tell me, did you leave school before your final exam?*

3.29 *Please, specify what were your reasons for leaving school before your final exam.*

These examples demonstrate that the grammatical form has not changed, only the utterance is in a more polite tone. We have also shown that indirect requests can be formulated with and without WH words as is generally the

case. Therefore we will not emphasize this issue any further but discuss the use of the WH words in a separate section below.

3.2.2.2 Interrogative–interrogative requests

Another way to make prerequests more polite is to change the grammatical form, namely, replace the imperative mood with the *interrogative mood* while formulating prerequests. Research has shown that, there seems to be a linguistic continuum where prototypes of imperative forms gradually shade into polite interrogative forms (Chisholm et al. 1984, Givon 1990). Below we demonstrate how imperative prerequests in survey research can gradually change into more and more deferent prerequests in interrogative form. Examples could be:

3.30a Tell me whether you are going to vote in the next election.
3.30b Please tell me whether you are going to vote in the next election.
3.30c Will you tell me whether you are going to vote in the next election?
3.30d Can you tell me whether you are going to vote in the next election?
3.30e Can you please tell me whether you are going to vote in the next election?
3.30f Could you tell me whether you are going to vote in the next election?
3.30g Could you please tell me whether you are going to vote in the next election?
3.30h Would you tell me whether you are going to vote in the next election?
3.30i Would you please tell me whether you are going to vote in the next election?
3.30j Would you like to tell me whether you are going to vote in the next election?
3.30k Would you mind telling me whether you are going to vote in the next election?
3.30l Would you be so kind as to tell me whether you are going to vote in the next election?

The first two examples 3.30a and 3.30b, are in the imperative mood. The remaining examples 3.30c–3.30l, switch to the interrogative mood, characterized by the inversion of the auxiliary verb and the subject. They use different combinations of the modal auxiliaries such as "will," "can," "could," and "would," indicating that they are asking for permission to ask for something. They start with asking permission by "will," which is gradually more polite than the imperative and are followed by the use of "can," which is a bit more hesitant, where the addition of "please" increases the relative politeness of the sentence. Thereafter the more polite and more distant form of "could" is introduced, which is again combined with "please" to increase politeness mood. Examples 3.30h to 3.30l use the form "would," which is even less forward

than the previous forms and therefore adds to the polite feeling. These examples show some gradations of politeness within its uses by adding "please" or combining it with "would you like," "mind," or "be so kind."

The reader may have noticed that logically the answer "yes" to any of these polite interrogative requests signifies that people are either willing to or can give the answer since it is formally related to the prerequest. Even in the polite form respondents in general will suppose that they are asked to appraise the embedded request presented to them and therefore will answer "yes," meaning that they are going to vote in the next election, or "no," meaning that they are not going to vote. If it is anticipated that polite requests lead to confusion, it is better to avoid using them. An unusual variation of the above two types is illustrated below:

3.30m Which issue, tell me, will mostly influence your vote?
3.30n Which issue, would you say, will mostly influence your vote?

These examples show that the prerequests are placed within the clause that would normally be considered as an embedded sub clause.

3.2.2.3 Declarative – interrogative requests

It is also possible to use *polite declarative* prerequests. They are presented in examples 3.31a and 3.31b:

3.31a I ask you whether you are going to vote during the next elections.
3.31b I would like to ask you whether you are going to vote during the next elections.

It is interesting to note that in examples 3.31a and 3.31b no actual request is presented. Formally the two texts are statements. As with the case of polite interrogative requests, research practice and conversational custom make it informally understood to listeners that they have to provide an answer to the embedded part in the sentence.

3.2.2.4 Interrogative – declarative requests

Finally it frequently happens that in survey research prerequests are formulated in the interrogative mood and the embedded request, in the declarative form. Examples 3.32a and 3.32.b illustrate this:

3.32a Do you think that the Republicans will win the elections?
3.32b Do you believe that abortion should be forbidden?

These examples show that the request is introduced by the declarative conjunction "that." The most common form of this type of request for an answer is illustrated by the next example:

3.32c Do you agree or disagree that women should have the right to abortion?

This is a popular form because any assertion can be transformed directly into a request for an answer by adding a prerequest (e.g., "Do you agree or disagree" or "How much do you agree or disagree") in front of the statement. Respondents are often provided with whole series of assertions of this type.

3.2.2.5 *More than two parts*

We also stated that a request for an answer consists at minimum of two parts, which in practice can mean that more than one prerequest occurs, which can take all kinds of grammatical forms. Semantically, one of them can be either neutral or polite, and the other may convey a concept-by-intuition where the proper request follows. Examples 3.33 and 3.34 illustrate this:

3.33 *Please, tell me whether you think that homosexuals should be given the same rights as heterosexuals.*

3.34 *I would like to ask you whether you can tell if you think that homosexuals should be given the same rights as heterosexuals.*

In example 3.33 the prerequest "please tell me" is a polite imperative while the second prerequest, "whether you think," introduces a cognitive judgment in an embedded interrogative form followed by a declarative mood "that homosexuals should be given the same rights as heterosexuals," conveying a specific policy.

Example 3.34 illustrates a chain of three prerequests. The first, "I would like to ask you," is a polite declarative statement. The second, "whether you can tell," is a neutral prerequest in interrogative form and the third, "if you think," relates again to an interrogative constituting a cognitive judgment. The main request is initiated by "that" and conveys a policy.

Here it is important to state that while the sentences are becoming quite long, the main risk is that the proper request will fall to the background. In the last section we will formulate some hypotheses concerning the possible effects of the consequences of length and complexity of sentences on the response.

3.3 THE MEANING OF REQUESTS FOR AN ANSWER WITH WH REQUEST WORDS

In all forms of requests for answers WH request words can be used as we have explored in Section 3.2.1.2, which studied direct requests with a specific introductory word. In that section it was also mentioned that these words refer to a specific aspect of an issue assuming that the basic idea is known. The example given previously was a request referring to the change of the position of black people in the United States:

3.35 *When did this change occur?*

This request for an answer presupposes that the respondent agrees that a change has occurred; otherwise the request has no meaning:

3.36 *Has the position of the blacks in the United States changed?*

It is clear that example 3.35 asks for the objective concept-by-intuition "time" while example 3.36 measures the subjective "judgment" of possible change. Here we see that a change in meaning similar to that described in the last section occurs. The difference is that now the change in concept is not due to a prerequest but to a WH request word and in general the concept referred to by the WH word is clear, even though some of these words can imply many concepts in a request for an answer. The different meanings of the requests for an answer using the different WH words are the topic of this section, which will be discussed in sequence from simple to complex.

3.3.1 "When," " where," and "why" requests

The most simple WH requests are the requests starting with the word "when," "where" or "why." These requests are simple because the words indicate only one specific concept. It is common knowledge that "when" asks for a time reference; "where" asks for a location, and "why" asks for a reason. Please refer to Section 3.2.1.2 for examples.

3.3.2 "Who" requests

"Who," "whose," and "whom" are used for asking information about a person or several people. "Who" and "whom" are pronouns that substitute for a noun, while "whose" can also be a determiner that occurs together with nouns, like "whose house." "Who" queries the personal subject. Examples of "who" that queries the personal subject are:

3.37 *Who is the new president of the United States?*

3.38 *Who is the most powerful person in the European Union?*

Using *"whose"* signifies requests asking for ownership:

3.39 *Whose house is this?*

On the other hand *"whom"* requests information about a personal object:

3.40 *To whom did you sell the house?*

3.3.3 "Which" requests

The request word "which" is used for *preference* requests such as

3.41 *Which party do you prefer?*

or

3.42 *Which car do you like the most?*

It can also be used as an alternative for "who." In combination with "which one" in example 3.43, it refers to a definite set of persons (Givon 1990: 794):

3.43 *Which one did it?*

"Which" also can be used as an alternative for "why," "where," or "when, if it is used in combination with nouns like "reason," "country," or "period," For

example, instead of "why," one can use "which" to ask about *relations*:

3.44 Which was the reason for the changes?

Instead of "where," one can use "which" to ask about *places*:

3.45 In which area did the change take place?

Instead of "when" one can use "which" to ask about *time*:

3.46 In which period did the change take place?

And instead of "how" requests (which will be discussed later), "which" requests can also be used to ask about *procedures*. For example:

3.47 In which way do you solve your financial problems?

The reader should be aware that instead of "which," one can in these cases also use "what." This request word is the topic of the next section.

3.3.4 "What" requests

"What" can be used in even more requests as it asks for the subject or the object. One very common use of "what" is in *demographic* requests such as

3.48 What is your family name?
3.49 What is your highest education?
3.50 What is your age?

It is also used in consumer research to ask for a specific aspect of the *behavior* of customers:

3.51a What did you buy?
3.51b What did you pay?

In time budget research or studies of leisure time a more open request type of "what" is used to ask for *behavior*:

3.52 What did you do (after 6 o'clock)?

"What" in combination with verbs like "cause" or nouns like "motives," or "goals" can also indicate a *relation*:

3.53 What caused the outbreak of World War I?

"What" can also be used to formulate requests about subjective variables. For example:

3.54a What do you think of Clinton's quality as a president?

or

3.54b What do you think of President Clinton?

Note that example 3.54a asks for an evaluation. However, it is not clear what concept is measured in example 3.54b. This depends on the answer alterna-

tives. If they are preset, they could be formulated in terms of various concepts. If they are not preset, it depends on what comes to the mind of the respondent at the moment of requesting.

3.3.5 "How" requests

Special attention has to be given to requests using the term "how." This term can be used in different contexts and in still more ways. The following different uses of the request word "how" will be discussed:

Measure of a procedure
Measure of a relationship
Measure of an opinion
Measure of quantity
Measure of extremity
Measure of intensity

We start with the use of "how" when asking about *procedures*. The request word "how" can first be used to ask about *procedures* that people use to accomplish a certain task. Typical examples include

3.55a How do you go to your work?

or

3.55b How did you solve your financial problems?

Examples 3.55a and 3.55b use the word "how" specifically and similar to the way words like "who," "where" and "when" are used in the previous sections.

A second application of the request word "how" is in requests about *relations* such as

3.56 How did it happen that the position of black people changed ?

In this case the request asks about the *cause* of the event mentioned. This request is rather close to the procedure request but the former one asks for a "tool" while the later one asks for a "cause."

The third application of the "how" request is an *open opinion* request such as

3.57 How do you see your future?

This request is similar to the open request we mentioned before when we discussed the "what" requests'. In fact often one can substitute "what" for "how.

The fourth use of the request word "how" is in requests about *quantities* and *frequencies* such as

3.58a How often do you go to the church ?

or

3.58b How many glasses of beer did you drink ?

or

3.58c How many hours a day do you watch television?

We have put this use of the request word "how" in a separate category because the answer is specific, as in our examples the expected answer is a number. The following applications of "how" are similar, but with different answers.

A fifth application of the "how" request form relates to requests that ask about the *extremity* of an opinion. They modify the request word by an adjective or past participle. Typical examples are

3.59a How good is Mr. Bush as president: very good, good, neither good nor bad, bad, or very bad?

or

3.59b How interested are you in politics: very interested, interested, a bit interested, or not at all interested?

In requests 3.59a and 3.59b respondents are asked to give more details about their *opinion*. The "how" request form indactes *extremity*. This form can also be applied to objective variables. An example is below:

3.60 How many kilos do you weigh: under 50 kilograms, between 50 and 60 kilograms, between 61 and 70 kilograms, or above 70 kilograms?

We should mention that this can also be done by a direct request with answer categories. For example we can ask

3.61a Is Bush a very good, good, neither good nor bad, bad or very bad president?

or

3.61b Are you very interested, rather interested, a bit interested, or not at all interested in politics?"

or

3.61c Do you weigh under 50 kilograms, between 50 and 60 kilograms, between 61 and 70 kilograms or above 70 kilograms?

It is unknown whether the direct request or the "how" request is better. However, some experiments have shown that requests with labels as responses are preferable if frequencies are asked for (Saris and Gallhofer 2004).

The sixth application of the "how" request asks for the *intensity of an opinion*. This type looks similar to the previous one, but an argument can be made that it represents a different request form [Krosnick and Fabrigar, (forthcoming)]. If this is the case we do not ask how extreme an opinion is but how strongly people agree with an assertion. For example

3.62a How strongly do you agree with the statement that Clinton was a good president?

or

3.62b How strongly do you believe that you will get a new job next year?

In such requests the gradation is not asked with respect to the quality of the president or the likelihood of an event but with respect to the strength of an opinion. Therefore it is called the *intensity* of an opinion.

Most of the specific requests have an equivalent translation in other languages. However, the word "how" has different meanings in romance languages like French and Spanish.[3]

3.4 SUMMARY AND DISCUSSION

In this chapter we focused on different linguistic possibilities to formulate a request for an answer. We called it a "request for an answer" because not only interrogative forms (requests) are used to obtain information from the respondents; imperative and declarative statements are also commonly employed. What the three request types share in common is that they ask the respondent to make a choice from a set of possible answers.

We have discussed several procedures. The first distinction we made was between direct and indirect requests. Direct requests consist of only one sentence, a request or an imperative, while indirect requests consist of a prerequest in the form of an interrogative, imperative, or declarative sentence with an embedded sentence that contains the real request. We also discussed specific requests introduced by particular request words such as "when," "where," " why," " which," " who," " what," and "how." These request words are used when the researcher wants to get specific information from the respondent about, for example, the time, place, or reason(s) of event. These possibilities are summarized in Table 2.1.

The most important result of this linguistic analysis is that one can formulate very different requests for an answer while the concept, the topic of research and the set of possible responses is the same. Logically that would suggest that the requests provide the respondents with the same choice and therefore the requests can be seen as equivalent forms. However, we have to warn the reader that the possibility cannot be excluded that differences will nevertheless be found in the responses for the different forms because the difference in politeness of the forms may have an effect on the response. In Chapter 4 we will demonstrate how to formulate requests for answers that are linguistically very similar but measure different concepts.

---

[3] In French, for instance, "how" in procedure, relationship, and opinion requests is translated as "comment." For "how" in frequency requests, "combien," or "avec quelle frequence," or "est-ce souvent que", is used. The extremity and intensity are expressed by "de quelle qualité est" and "dans quelle mesure vous etes d'accord," and so on. In Spanish, "how" in procedure, relationship, and opinion requests is translated by "como." For "how" in frequency requests, "cuanto" is used. The extremity and intensity are expressed by "hasta que punto" or "hasta que grado" and "en que medida."

By specifying all these different forms[4] we tried to indicate the diversity of the possibilities to formulate requests for answers in survey research. Although linguists suggest that in many cases the meaning of the requests is the same, this does not mean that respondents will perceive these requests as identical and that they will reply in the same way.

Without claiming that all requests for answers fit in the system developed in this chapter, we think that it is useful to keep these possibilities in mind when formulating requests for answers and analyzing requests for answers to clearly grasp the diverse grammatical forms and the potential differences in meaning of the requests.

EXERCISES

1. For the following two concepts-by-intuition derive assertions representing these concepts and transform these assertions into different requests for an answer.
   - Trust of the government
   - The occupation of the respondent
2. Two requests for an answer have been mentioned below.
   - *Is it the position of black people that has changed?*
   - *Is it the position of black people that has changed by new laws?*

   a. What are the potential answers to these requests?
   b. Do these answers mean the same?
   c. Why is there a difference?
3. How would you formulate a request about
   a. A perception if women have the right of abortion
   b. The norm that women should have this right
   c. The evaluation of this right
   d. An importance judgment of this right
4. Finally, check for your own questionnaire whether the transformation of the concepts-by-intuition in requests for an answer was done in the proper way. Should you change some requests?

[4] Linguists (Chisholm et al. 1984; Givon 1990; Huddlestone 1988, 1994) also discern some other types of questions that are, in our opinion, typical for normal conversation but not for requests for answers in survey research. To these questions, for instance, belong so-called "echo questions" which repeat what has been said before because the listener is uncertain about having understood the question well. An example could be: "Am I leaving tomorrow?" "Multiple questions" are also used frequently in conversation (Givon 1990: 799) such as "who said what to whom?" In English interrogative tags also are quite common such as "he left alone, didn't he." It is clear that such constructions are too informal and therefore are preferably avoided in survey research.

This Page Intentionally Left Blank

PART II

Choices involved in questionnaire design

Part I discussed the basic steps needed to formulate requests for an answer in order to operationalize the measurement of the desired concepts. Part II will show that in survey research many more choices have to be made to design a questionnaire. The following issues will be discussed in sequence:

1. The different ways requests for an answer can be formulated (Chapter 4)
2. The choice of the response alternatives (Chapter 5)
3. The structure of open-ended and closed survey items (Chapter 6)
4. The structure of batteries of survey items (Chapter 7)
5. Other choices in survey design such as the order and layout of the questions and the choice of the data collection method (Chapter 8)

This Page Intentionally Left Blank

CHAPTER 4

Specific survey research features of requests for an answer

Chapter 3 examined the various linguistic structures of requests for answers. In this chapter we will discuss features of requests for an answer that are important with respect to their consequences for survey research. Hence, we will first look at the characteristics of requests that cannot be changed by the researcher because they are connected with the research topic. Then we will discuss some features that the researcher has influence over, such as the choice of the prerequest and the use of batteries of requests for an answer with the same format. So far we have discussed only single requests, but if batteries are used, the form of the requests changes significantly.

Other issues that social scientists are concerned with include whether the request is balanced in the sense that equal attention is given to positive and negative responses in the request and whether absolute or relative judgments are asked, as well as whether a condition should be specified within the request. Finally, the request for an answer can include "opinions of others," or "stimuli to answer," or emphasize that a "personal opinion" is asked. In the following sections these characteristics will be discussed in detail.

4.1 Select requests from databases

So far we have suggested the following method to develop a request. First, it is crucial to determine what needs to be studied; for example, "the satisfaction with the work of the present government" or "the amount of hours people work normally." The first concept is a feeling about the government and the second is a factual request about the work. Next typical assertions for these concepts (Chapter 2) need to be specified like the two examples below:

4.1a *I am (very) (dis)satisfied with the work of the present government.*

4.2a *Normally I work x hours.*

The last step is to transform these assertions in requests for an answer (Chapter 3), for example

4.1b *Are you satisfied or dissatisfied with the work of the present government?*

This process has little margin for error. The requests measure what was planned to be measured. However, there are other ways of obtaining requests.

For many topics requests already exist in archives such as the one in Cologne (Germany), Essex (United Kingdom), or Ann Arbor (United States) or "question banks" such as the one of CASS in Southampton. Mostly the requests are ordered in some type of classification. However, beware that the classification has to be very detailed in order to find the proper requests. For example, the following classification can be used as a first step:

1. *National politics*
2. *International politics*
3. *Consumption*
4. *Work*
5. *Leisure*
6. *Family*
7. *Personal relations*
8. *Race*
9. *Living conditions*
10. *Background variables*
11. *Health*
12. *Life in general*
13. *Other subjective variables*

This first step in classification is not detailed enough because a large number of requests concerning national politics (the first topic) and concerning work (the fourth topic) exist. Therefore be prepared to invest some time in searching the exact measure of the desired concept. The criterion to evaluate whether a request measures what was intended to be measured is the same as was discussed in the first three chapters. If a concept-by-intuition is studied a direct request is possible, and Chapters 2 and 3 are applicable. If a concept by postulation is being studied, first determine what concepts-by-intuition form the basis for this more abstract concept and then find their direct measures as discussed in the previous two chapters. The most important criterion is, of course, that the possible answers represent assertions that are obvious assertions for the chosen concepts-by-intuition. Chapter 2 provides ample suggestions for this type of check.

4.2 OTHER FEATURES CONNECTED WITH THE RESEARCH GOAL

Directly connected with the research goal and consequently with the choice of concept are some other characteristics of the survey items: the *time reference*, *social desirability*, and *saliency* or *centrality*. We start with the time reference.

Requests can be asked about the present situation: feelings at the moment or satisfaction with different aspects of life or opinions about policies, norms, or rights. Requests can also be directed to future events or intended behavior. One can ask whether one will buy some goods in the future or will support some

activity or expect changes or events, for instance. Finally, survey items can be directed to the past asking whether one has bought some thing last week or whether one has been to a physician, dentist, and hospital during the last year. It will be clear that the time period mentioned in the request – past, present, or future – is completely determined by the goal of the research, and the designer of the study normally has no possibility to change this time period. Only the requests about the past give a bit more freedom to the researcher. Let us look at this issue a bit more closely.

The time period indicated in requests about the past is called the *reference period*. It will be clear that the longer the reference period is, the more unlikely it is that one can reproduce the requested information from memory. This holds especially for activities that occur very frequently such as, for example, media use. For that reason researchers use as an alternative requests about yesterday. Hence, instead of the request in example 4.3a they ask example 4.3b:

4.3a *How much time did you spend watching programs on politics or actuality last week?*

4.3b *How much time did you spend watching programs on politics or actuality yesterday?*

But because requests like 4.3b lead to unusual results for at least some people, one also asks the request of example 4.3c:

4.3.c *How much time did you spend watching programs on politics or actuality on a normal day?*

It is unclear what time period is used in this request. One could say that the respondent is asked for his/her normal behavior at present. Such a shift in time is of course only possible if the research goal allows it.

One more problem should be mentioned concerning requests referring to the past. It is well known from research that people have a tendency to see events as closer to the date of the interview than is true in reality. This phenomenon is called *telescoping* (Schuman and Presser 1981). A typical request that reflects this problem is shown in example 4.4:

4.4 *Have you experienced robbery or theft during the last year?*

Respondents are inclined to mention many more cases than should be reported. Scherpenzeel (1995) found that the reported number of cases is twice as high using this request (4.4) than when one asks two requests illustrated by examples 4.5a and 4.5b:

4.5a *Have you experienced robbery or theft during the last 5 years?*

4.5b *How about the last year?*

It seems that people can better estimate the point in time if first a larger reference period is mentioned (4.5a) than in a one-step procedure like 4.4.

In general the designer of a questionnaire has little flexibility with respect to the specification of the time period mentioned in the requests. Basically he/she has only a choice with respect to the reference period that will be mentioned.

A second characteristic that is directly connected with the choice of the concept is the *social desirability* of some responses. As an example, we can mention that using the direct request about political interest, it is socially desirable for some people to answer that they are interested even if they are not. This happens because the respondents want to make a good impression on the interviewer. This means that differences in responses can be expected between surveys using interviewers and studies that do not use interviewers. So, for sensitive requests differences are expected between personal or telephone interviews and mail surveys and other self-completion procedures. For requests about criminal and sexual behavior, very large social desirability effects have been found in this way (Aquilino 1993, 1994; Turner et al. 1998). This suggests that in a study where social desirability can play an important role, one should consider using a data collection method that reduces the effect of social desirability as much as possible.

The third characteristic that is directly connected with the choice of a concept is the *centrality* or *saliency* of the necessary information to answer the requests. In the past the idea was that people have an opinion about many issues stored in memory that they just had to express in one of the presented response alternatives. Nowadays, researchers have a different view on this process, thanks to the important work of Converse (1964), Zaller (1992), Tourangeau et al. (2000). It is more likely in many situations that people create their answers on the spot when they are asked a request. They will do that on the basis of all kinds of information that they have stored in memory, and it depends on the context of the request, recent events, and their mood which information will be used and therefore what answer will be given. As a consequence, one can expect quite a lot of variation in answers to the same request at different points in time (Converse 1964; Van der Veld and Saris 2003).

However, one should not exaggerate this point of view. There are requests where most people give more or less the same answer all the time, for example, requests about their personal lives, backgrounds, and living conditions. There are also topics about which some people have rather stable opinions and others do not. For example, with respect to political issues, some people who are very interested and follow what is going on have a clear opinion; there are, of course, also people who are not at all interested in politics and are, therefore, more likely to provide different answers if they are forced to answer requests about these issues. This does not mean that this division will always be the same. It may be that the people, who know nothing about politics, know a lot about consumer goods and education where the political interested respondents do not know much about these issues. So the saliency of opinions depends on the topic asked and the interest people have in the specific domain of the survey items (Saris and Sniderman 2004).

4.3 SOME PROBLEMATIC REQUESTS

Besides the problems unavoidably connected with the research topic, there are also problems that can be avoided such as the so called "double-barreled requests" and assertions with more than one component. We will also indicate how to correct them in order to improve the comprehension of the respondents. These complications are also mentioned by Daniel (2000) in his request taxonomy.

4.3.1 Double-barreled requests

In the literature about survey research the problem of requests with several concepts has been extensively discussed (Converse and Presser 1986; Fowler and Maggione 1990; Lessler and Fortsyth 1996; Graesser et al. 2000a,b). An example of such a so called double-barreled request could be

4.6a *How do you evaluate the work of the European Parliament and the Commission?*

The problem with such a request with two object complements in 4.6a (the work of the European Parliament and the Commission) is that two simultaneously opposing opinions are possible: a positive opinion about the Parliament and a negative opinion about the Commission. This leads to confusion about how to answer the request. Linguistically this is a complex sentence built up with the coordinate conjunction "and," and as we stated in Chapter 2, in this case with two different subjects it can become problematic. To avoid this problem, two requests, each containing one of the object complements, is a solution:

4.6b *How do you evaluate the work of the European Parliament?*

4.6c *How do you evaluate the work of the European Commission?*

Another example of two concepts in one request is the following:

4.7a *Do you agree with the statement that the asylum seekers should be allowed into our country, but should adjust themselves to our culture?*

Although such a statement is not unusual in colloquial speech it can create problems for clear answers in surveys. The reason is that the first part of this statement is a right but the second part is a norm. It is again quite possible that a person is opposed to immigration but thinks that immigrants should integrate once they have entered a country. Again, this respondent can be perplexed about what answer to provide to this request. Splitting this statement into two separate requests creates clarity:

4.7b *Do you agree with the statement that asylum seekers should be allowed into our country?*

4.7c *Do you agree with the statement that if asylum seekers come to our country, they have to adjust themselves tc our culture?*

The previous examples 4.6a–4.7c showed the problem of double-barreled requests. There are also double-barreled requests that work as intended, as a study for some items of the human value scale of Schwartz (1997) demonstrated. The items are formed by a combination of a value and a norm or a feeling. An example is the following request:

4.8 How much are you like this person?
Looking after the environment is important to him/her. He/she strongly believes that people should care for nature.

In this case the importance of a value and a norm are combined in a complex assertion of similarity. This is in principle a typical example of a double-barreled request, but if we ask the two assertions separately with the same prerequest the correlation between the answers (after correcting for random errors) is so high (.95) that one can assume that these two assertions measure the same (Saris and Gallhofer 2004).

The above is an interesting example showing that double-barreled requests do not always have to be problematic. However, one should be aware that they can cause problems and should be used only after a careful study of the consequences. In general such requests can be very confusing for respondents.

4.3.2 Requests with implicit assumptions

There are also requests for answers that assume a first component that is not literally asked but is implicitly true in order to respond to the second component. An example could be

4.9a What is the best book you read last year?

Here the hidden assumption is that the respondents actually read books. People who do not read books can be unsure about how to answer this request. If the hidden component is made explicit in a separate request, the problem is resolved:

4.9b Did you read books last year?

If yes:

4.9c What is the best book you read last year?

Sometimes the previously discussed hidden assumption in the first component, is stated explicitly in the request but the focus for answering is on the second component (Emans 1990) such as in example 4.10:

4.10 Did you read books last year and what is the best book you read?

Again, respondents who do not read books will be confused about how to answer the request. Again, the remedy is to split these two requests into two separate requests.

4.4 SOME PREREQUESTS CHANGE THE CONCEPT-BY-INTUITION

Although it is possible to transform assertions in many different ways into requests for answers, it is not always risk-free. In the previous chapter we have discussed prerequests referring to words such as "saying," "telling," " asking," and " stating," which were used to indicate a simple transfer of information. They did not refer to specific concepts-by-intuition as described in Chapter 2, which might differ from the concept used in the request for an answer. Hence it can be concluded that using these verbs will not change the concept-by-intuition, as this is shown in four different assertions in direct request format below:

4.11a Has the position of black people changed in the last 30 years?
4.11b Was Clinton a good president?
4.11c Should women have the right to abortion?
4.11d Did you live with your parents when you were 14 years old?

In sequence these requests represent a judgment (4.11a), an evaluation (4.11b), a right (4.11c) and a behavior (4.11d). If at the beginning of the request "tell me," "may I ask," or any other prerequest is combined with any of the abovementioned neutral verbs the concept measured will not change.

Prerequests of survey items such as "think," "believe," " remember," "consider," " find," " judge," " agree," " accept," " understand," and " object," refer to a cognitive judgment. Linguists like Quirk et al. (1985: 1180–1183) independently classified these verbs in a similar way. One would think that using such verbs in the prerequests would change the concept measured, but it doesn't always happen, as can be seen in the next three examples.

4.12a Do you think that the position of black people has changed in the last 30 years?
4.12b Do you think that Clinton was a good president?
4.12c Do you think that women should have the right to abortion?

There are also verbs which measure feelings such as "like," and "enjoy." If such verbs are used in prerequests in the same way, the concept may change to a feeling about a concept. Examples 4.13a–4.13c illustrate this:

4.13a Do you like that the position of black people has changed in the last 30 years?
4.13b Do you like that Clinton was a good president?
4.13c Do you like that women should have the right to abortion?

The structure of the requests is exactly the same, only the meaning of the verb is changed from "think" to "like" (4.12–4.13).

The same effect occurs with adjectives that refer to other concepts like "importance" or "certainty." In the examples below we see that the concepts asked in the indirect requests are different from the concepts in the direct requests mentioned so far.

4.14a Is it important for you that the position of black people has changed in the last 30 years?
4.14b Is it important for you that Clinton was a good president?
4.14c Is it important for you that women should have the right to abortion?

These examples clearly indicate that one has to be careful with a change from a direct request to an indirect request for substantive reasons. By selecting an indirect request the concept-by-intuition measured in the request, can change in agreement with the concept expressed in the verb or adjective of the prerequest. That is not the case with the neutral terms that we have used in the previous sections, but this occurs with less neutral terms and not always as we saw in the changed verb examples (4.12a–4.12c) "think," " believe," and similar which measure judgments. This is still an area where further research is needed to investigate when the concept measured changes and when it does not.

In Chapter 2 we mentioned that terms added to an assertion can change the concept. Thus, using a prerequest that is introducing a different concept-by-intuition than the concept connected to the embedded query is referred to as a *complex assertion*. As was stated before, complex concepts seem to confuse people leading to lower reliability of responses (Saris and Gallhofer 2004) and should be avoided if possible.

4.5 BATTERIES OF REQUESTS FOR ANSWERS

In survey research many requests are asked, one after the other in series. If they are in similar form or can be made similar, then the whole process can be simplified by the use of *batteries of requests*. In batteries the entire request and answer categories including the introduction, the request in the broadest sense, and the eventual components after the request such as instructions are mentioned before the first stimulus or statement. Subsequently, one stimulus or statement after the other follows without repeating the request and the answer categories, since it is assumed that the respondent already knows them. Written questionnaires present stimuli and statements often in table format where the stimuli or statements are presented in rows and the answer categories or rating scales, in columns. We will call this kind of structure a "battery of requests for answers." The difference between stimuli and statements is that statements are complete sentences while stimuli do not consist of complete sentences. They can contain a noun, a combination of nouns, or another part of a sentence or a subordinate clause.

From the above one can conclude that requests for answers with stimuli or statements are quite different from the requests for answers studied in Chapter 3 because they occur in series. The consequences of this approach, which is typical for survey research, will be discussed in later chapters. Here we want to present the structure of batteries and to discuss some of the choices that have to be made to construct batteries. We start with the use of stimuli.

4.5.1 The use of batteries of stimuli

Example 4.15 presents a possible formulation of a battery of stimuli:

4.15 *There are different ways of attempting to bring about improvements or counteract deterioration in society. During the last 12 months, have you done any of the following?*
Please mark either "yes" or "no".

	Yes 1	No 2
A. Contacted a politician	☐	☐
B. Contacted an association or organization	☐	☐
C. Contacted a national, regional or local civil servant	☐	☐
D. Worked in a political party	☐	☐
E. Worked in a political action group	☐	☐
F. Worked in another organization or association	☐	☐
G. Worn or displayed campaign badge/sticker	☐	☐
H. Signed a petition	☐	☐
I. Taken part in a public demonstration	☐	☐
J. Taken part in a strike	☐	☐
K. Boycotted certain products	☐	☐

In this example "any of the following" stands for the so-called stimulus, which could be a single action such as "contacted a politician" or "taken part in a strike." Such stimuli batteries can also consist of nouns or combinations of nouns. Example 4.16 illustrates this:

4.16 *How satisfied are you with the following aspects of life:*
1. Your income
2. Your house
3. Your social contacts
...

Another possibility is that a stimulus consists of a part of a verb phrase such as in example 4.17:

4.17 *Did you do any of the following?*
Shopping
Cleaning
Washing
...

The reader should be aware that stimuli also could occur in all kinds of combinations of requests for answers such as example 4.18 illustrates:

4.18 *Please tell me, whether or not you are satisfied with the following aspects of life:*

One reason to use batteries of stimuli is that the requests and the response categories do not have to be repeated each time. This is very efficient for the questionnaire designer, and the printing of the questionnaires and the interviewer, since they have less to write, print, and read. So far we have not seen any convincing evidence that this approach has a negative effect on the answers of the respondents, although one can expect that they will not answer the requests independently of each other. It is more likely that they make use of their previous answer to judge the next stimulus in case of evaluations on scales. This would lead to correlated errors between the responses; however, Saris and Aalberts (2003) did not find strong evidence for this in their research.

4.5.2 The use of batteries of statements

Very popular in survey research is the indirect request with an interrogative prerequest using the verb "agree" followed by assertions discussed in Chapter 2, often called "statements." A typical example[1] of such a battery of agree/disagree requests is given below. Example 4.19 is taken from a study of Vetter (1997), but the concept "political efficacy," which is measured here has already been questioned in a similar way in 1960 in *the American Voter* (Campbell et al. 1960):

4.19 How far do you agree or disagree with the following statements (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5) strongly agree ?

Statements	*Possible responses*				
	1	*2*	*3*	*4*	*5*
A. I think I can take an active role in a group that is focused on political issues					
B. I understand and judge important political requests very well					
C Sometimes politics and government seem so complicated that a person like me cannot really understand what is going on.					

Typical for such a battery of statements are the following characteristics:

1. The request for an answer is formulated only once before the first statement;
2. Also the response categories are mentioned only one time;
3. The formulation of the request for an answer is rather abstract by use of the term "statement" at the place where normally the statement itself follows.

[1] These requests for an answer were originally formulated in German. The authors of this text have translated them into English. These requests are not given as examples of very good requests for this section.

If we abide by the rules we have seen in Chapter 3, the following formulations could also be an alternative:

4.20a How far do you agree or disagree that you can take an active role in a group that focused on political issues: (1) disagree strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5) strongly agree?

4.20b How far do you agree or disagree that you understand and judge important political requests very well: (1) disagree strongly, (2) disagree,(3) neither agree nor disagree, (4) agree, (5) strongly agree?

4.20c How far do you agree or disagree that sometimes politics and government seem so complicated that a person like you cannot really understand what is going on: (1) disagree strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5) strongly agree?

This transformation to a standard indirect request with an interrogative agree/disagree prerequest and a different embedded declarative assertion of each request makes it clear that the battery form is far more efficient.

Krosnick and Fabrigar (forthcoming) make a comparison with direct requests for an answer. They suggest that the popularity of the use of agree/disagree batteries lies in the fact that it reduces the amount of work as we have mentioned above and maybe even more importantly, this approach can be applied to nearly all possible assertions in the same way.

If direct requests for an answer are more desirable a different form for each assertion is needed, as is illustrated for the same assertions in examples 4.21a–4.21c. The transformation of the battery mentioned above to three direct requests leads to the following result:

4.21a Could you take a very active, quite active, limited role or no role at all in a group that is focused on political action ?

4.21b Can you understand and judge important political issues very well, well, neither good nor bad, bad , very bad ?

4.21c How often does it seem to you that politics and government are so complicated that a person like you cannot really understand what is going on: very often, quite often, sometimes, seldom, or never ?

This transformation again indicates the efficiency of the battery format for the researcher and the interviewers. They do not have to specify and read a different response scale for each separate assertion. Whether the efficiency for the researcher and the interviewer goes together with efficiency for the respondent and with better data is another matter. Saris and Krosnick (forthcoming) have the following opinion on the matter:

> The goal of agree/disagree requests is usually to place respondents on a continuum. For example, an assertion saying "I am usually happy" is intended to gauge how happy the respondent usually is, on a dimension from "never" to "always." An assertion saying "I like hot dogs a lot" is intended to gauge how much the respondent likes hot dogs, on a dimension from "dislike a lot" to "like a lot." And a statement saying "Ronald Reagan was a superb President" is intended to gauge respondents' evaluations of Reagan's performance, on a dimension ranging from "superb" to "awful."
>
> To answer requests with such statements requires four cognitive steps of respondents (Carpenter and Just 1975; Clark and Clark 1977; Trabasso et al. 1971). First, they must read the statement and understand its literal meaning. Then, they must look deeper into the statement to discern the underlying dimension of interest to the researcher. This is presumably done by identifying the variable quantity in the statement. In the first example above, the variable is identified by the word "usually" it is frequency of happiness. In the second example above, the variable is quantity, identified by the phrase "a lot." And in the third example, the variable is quality, identified by the word "superb." Having identified their dimension, respondents must then place themselves on the dimension of interest. For example, the statement, "I am usually happy," asks respondents first to decide how happy a person they are. Then, they must translate this judgment into the agree/disagree response options appropriately, depending upon the valence of the stem. Obviously, it would be simpler to skip this latter step altogether and simply ask respondents directly for their judgments of how happy they are.

It is self-evident here that answering batteries of statements is not a simple task for the respondent. Moreover, hundreds of papers have been written about the issue that respondents may have a tendency to simplify their task and to answer all requests in a battery in a same way. This phenomenon is called *response set* or *acquiescence*. The response set will increase the correlation between the answers in the batteries but this extra correlation is a method effect and has nothing to do with the substance of the requests. Krosnick and Fabrigar (forthcoming) and Billiet and McClendon (2000) have discussed this problem extensively. It is also one of the possible reasons why method effects are found in multitrait-multimethod studies (Andrews 1984; Költringer 1995; Scherpenzeel and Saris 1997; Saris and Aalberts 2003).

Finally, Krosnick and Fabrigar (forthcoming) have made the argument, mentioned in Chapter 2, that the requests asking "How far do you agree " does not estimate the extremity of an opinion but the intensity, which is a different aspect of measurement. The latter aims at the strength of the agreement with the statement and this is not the same as the extremity of an opinion in the

former. If one says "I like ice cream very much," that is not the same as "I very strongly agree with the statement: I like ice cream."

We would like to mention one more complication for this method. As was mentioned above the respondents have to place themselves in the dimension of interest. After careful examination of statement 4.19c, it was suggested that the purpose of the item was to evaluate how often people had the impression that politics and government were too complicated. This was formulated in example 4.21c which is repeated here in example 4.22.

4.22 How often does it seem to you that politics and government are so complicated that a person like you cannot really understand what is going on: very often, quite often, sometimes, seldom, or never?

It is very clear what a choice of one of the answer categories means; however, this does not mean that no errors will be made (Hippler and Schwarz 1987) or that people have a clear opinion in their mind of what they should say (Tourangeau et. al. 2000).

However, several alternatives for this request are available if an agree/disagree format is used, as we show in questions 4.22a–4.22e:

4.22a How far do you agree or disagree that politics and government ***very often*** *seem so complicated that a person like me cannot really understand what is going on: (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree,(5)strongly agree*

4.22b How far do you agree or disagree that politics and government ***quite often*** *seem so complicated that a person like me cannot really understand what is going on: (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5)strongly agree?*

4.22c How far do you agree or disagree that politics and government ***sometimes*** *seem so complicated that a person like me cannot really understand what is going on: (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5)strongly agree?*

4.22d How far do you agree or disagree that politics and government ***seldom*** *seem so complicated that a person like me cannot really understand what is going on: (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5) strongly agree?*

4.22e How far do you agree or disagree that politics and government ***never seem*** *so complicated that a person like me cannot really understand what is going on: (1) disagree very strongly, (2) disagree, (3) neither agree nor disagree, (4) agree, (5) strongly agree?*

These statements differ only by the word indicating the frequency of the occurrence of the event of interest. Logically all these possibilities (and many others) can be employed, and there is seemingly no reason to prefer one over another. But are there practical reasons to prefer one request above the other? In order to check this, let us perform a small thought experiment.

Imagine that you have the idea *very often* that politics is too complicated for you. Now an interviewer comes with the request 4.22a and if you have the idea *very often*, then your answer is simple: strongly agree. Imagine now that you have the idea *often* and you are confronted with the same request 4.22a: both agree and disagree could be chosen. Formally disagree is better but with a bit of flexibility you could as a respondent also say agree. Suppose now that you have the idea *sometimes* and you are confronted with the same request: most likely you would choose disagree.

Now, imagine again that you have the idea *very often* but the request is asked if you have this idea sometimes as in 4.22c. You may be confused as to what to answer because you can say disagree since you have these ideas often but you can also agree as you have them more than sometimes. Suppose now that you *never* have these ideas and the interviewer uses request 4.22c with the term sometimes. You could say "disagree" since you never have these ideas or you can agree depending on your perception of whether "sometimes" is rather close to never.

Our thought experiment shows that the statements in the middle of the scale encounter the problem that people at both sides of the spectrum can give the same answer, which makes further analysis rather problematic. Extreme statements have a lesser issue with this particular problem, but these statements have the problem that people with a different opinion than stated in the request can all choose the same response of disagree. This effect will be even stronger when the extreme statement is very extreme.

The conclusion on the basis of our practical analysis is that, if one really wants to use statements, one should choose a statement that represents an extreme position but that is not too far from the opinions of the people; otherwise no variation will be obtained. This analysis also shows that the choice of the formulation of item 3 in the political efficacy request is definitely incorrect.

Given all the complications of batteries with statements it is very questionable why this type of formulation is so popular. Further research is required, but we recommend avoiding this approach and using direct requests. It is more work for the researcher and the interviewer but it simplifies the task of the respondents and probably increases the quality of the answers.

4.6 OTHER FEATURES OF SURVEY REQUESTS

The possible consequences of other features of requests are discussed in the next sections.

4.6.1 The formulation of comparative or absolute requests for answers

We move now to a quite different aspect of the formulation of requests for an answer, namely, the use of comparative or absolute judgments. Comparative requests for answers ask about the similarity or dissimilarity of two objects, and they also ask for degrees of similarity. Examples of this type include

4.23a Are you more satisfied with your new job than with the old one?
4.23b Do you earn less money in your new job?
4.23c How much better is your daughter in languages than your son?
4.23d How much do you prefer languages above science?
4.23e Which political party do you prefer?

As the first two examples 4.23a and 4.23b illustrate, the inequality can be expressed by "more...than" or "less...than" where the comparison "than" can be implicit as in the second example. But it also can be expressed by comparative adjectives or adverbs such as "much better than" (4.23c) or by words that indicate a preference as shown in the last two examples.

Requests for an answer that ask for an absolute judgment, in contrast, do not express a comparison in terms of more or less than from a reference object. Absolute judgments are very frequently used in survey research. Examples are as follows:

4.24a Are you satisfied with your job?
4.24b How satisfied are you with your job?
4.24c How good are you at mathematics?

Although absolute judgments are very popular in survey research, it is questionable whether people are very good in making such judgments. In psychophysics this phenomenon has also been observed by Poulton (1968). Similar results have been found by Saris (1988) in survey research. A famous experiment by Schwarz and Hippler (1987) showed the same results. They asked for the amount of time people spent watching TV and showed that even in such cases many people gave relative judgments, relative to watching patterns of other people, suggested by the specified response categories, and not absolute judgments. We will come back to this example in the next chapter.

4.6.2 Conditional clauses specified in requests for answers

Sometimes in requests for answers clauses are included that refer to something that must happen first so that something else can happen. This is called a "condition" in the narrowest sense, or an event is mentioned that is qualified as uncertain. Such clauses are called *conditional* (Swan 1995: 245,252), and they restrict the content of the request to this specific condition or event. The following examples can illustrate conditional clauses:

4.25a Do you think it is acceptable that a woman has an abortion if she has been violated?
4.25b If the present government is reelected, do you believe that they will realize what they had promised before the elections?

4.25c Should refugees be allowed to work in our country, provided they take simple jobs?

4.25d Should Muslims be allowed to build mosques in our country as long as they are not subsidized by the government?

4.25e If you finish your studies in some years, are you planning to work in the field of study?

4.25f Suppose that the government increases the income tax next year; would you have to change your lifestyle?

4.25g Imagining, you were the president, which of the following measures for our country would you take first?

Examples 4.25a–4.25d illustrate conditions in the narrowest sense. The first two are specified by an "if" clause, while the third and the fourth use the expressions "provided, and "as long as," which means that the event mentioned in this clause should occur first before the main clause can be appraised. Examples 4.25e–4.25g refer to uncertain or hypothetical events. Request 4.25e is again formulated with the word "if" and expresses just an uncertain event in the future. Often reality is too complex to be asked without condition, like requests about abortion.

Request 4.25f uses the word "suppose" and indicates in this example again an uncertain event in the future, while the last example expressed by "imagine" refers – because of the use of the past tense – to a very unlikely event in the future. Respondents may have never thought about these specific hypothetical situations. In that case they have not premeditated their answer, and it is questionable if these responses have any stability (Tourangeau et al. 2000).

4.6.3 Balanced or unbalanced requests for answers

A *balanced request* for an answer means that it is made formally explicit that both negative and affirmative answers are possible (Schuman and Presser 1981:180, Billiet et al.1986: 129). If only one answer direction is provided the request for an answer is called *unbalanced*. An example of a balanced request could be:

4.26 To which extend do you favor or oppose euthanasia?

This request is balanced as it explicitly specifies both answer directions: in favor of and in opposition to. Sometimes this seems to be a bit exaggerated. For example one could also have asked:

4.27 Do you strongly favor, favor, neither favor nor oppose, oppose, or strongly oppose euthanasia?

Such requests are formulated because the researcher tries to prevent more attention being given to one side of the scale than to the other. In general it is supposed that a bias in the response will occur in the answer direction that is indicated in the request even though there is no research evidence supporting

this assumption. The reason that no errors have been found may be that people are very much familiar with one-sided formulations and are very well able to fill in the missing alternatives themselves (Gallhofer and Saris 1995).

The following example is balanced although the request indicates none of the answer directions:

4.28 What do you think about euthanasia?

A request that does not specify the different sides is also considered as balanced in our research, although this is a rather arbitrary decision. Examples of unbalanced requests for an answer could be:

4.29a To what extent do you favor euthanasia?
4.29b To what extent do you oppose euthanasia?
4.29c Some people think that euthanasia should be legalized. In principle, what is your opinion about euthanasia?

Example 4.29a only mentions the positive answer direction, while the negative one should be guessed by the respondent. In example 4.29b only the negative direction is indicated and in example 4.29c only a favorable opinion is mentioned in the survey item.

In the case where the response possibilities go from zero to positive or from zero to negative (unipolar scales, Chapter 5), the notion of balance is not applicable because there exists only one direction. An example might illustrate this:

4.30 How often do you go to church?

Here "often" is mentioned in the request, but the request is nevertheless unbiased because this is a unipolar request, as there is only one side. The following request for an answer, however, is more complicated:

4.31 To what extent do you favor euthanasia?

This question is unbalanced because only one side of the scale is indicated. However, the unbalanced question can be unbiased if it is posed only to respondents in favor of euthanasia. Otherwise this request is a "leading" request and that is an extreme form of bias.

4.7 SPECIAL COMPONENTS WITHIN THE REQUEST

Sometimes other components, not necessarily belonging to the request for an answer are placed in the request. We shall discuss two different components: remarks to stimulate the respondent to answer and remarks that emphasize that the subjective opinion of the respondents are requested and not a general statement. We start with the remarks that are intended to stimulate the response.

4.7.1 Requests for answers with stimulation for an answer

A special stimulation to elicit an answer from the respondent can be included in the requests for answers. They can be in either imperative or interrogative prerequests with all kinds of gradation of politeness as already mentioned in Chapter 3 in connection with procedures to formulate requests for answers. Some examples of a stimulation to answer within requests for answers could be

4.32a Tell me, are you going to vote?
4.32b Would you be so kind as to tell us what you did before studying at the university?
4.32c Could you tell us who is the president of the EU?

Sometimes a stimulation for an answer also occurs in other parts of survey items such as introductions or motivations of the researchers, which are discussed in Chapter 6.

The presence or absence of a stimulation to answer requires attention because their presence might make a difference in the readiness of the respondent to comply. If a stimulation is formulated very politely, it might be that the respondent is more inclined to answer, even if this person has no specific opinion and might just give a random opinion because of the extra encouragement to give an answer.

4.7.2 Emphasizing the subjective opinion of the respondent

Like stimulation for an answer, a stimulus for the respondent to give his/her own opinion can occur within requests and encourage the subjects to give an opinion even if he/she hardly thought about the issue. However, this procedure has an effect and will be studied later. Some examples of stimulation of respondent opinion might be:

4.33a According to you, what is the most important issue in this election?
4.33b In your opinion who is responsible for the economic recession in our country?
4.33c What do you believe/think is the main reason for the economic recession?
4.33d We would like to know whether you personally think that the death penalty should be implemented.

The first two examples relate to specific direct requests where the expressions "according to you" or "in your opinion" stress that a personal appraisal is desired. In the third example the interrogative clause "do you think" emphasizes the subjective opinion and in the fourth example the clause "whether you personally think..." functions in a similar way. Emphasis on the subjective opinion also can occur in other parts of the survey item such as in the introduction (see Chapter 6).

Summary

In this chapter several decisions in developing a request for an answer are discussed again, but now from the perspective of a survey researcher. The choice of the research topic brings with it some unavoidable consequences. For example, given the research goal, the decision of whether the requests are directed to the past, present, or future is predetermined. The research goal also determines the social desirability of the possible response alternatives and the salience of the topic. However, the format of the question can be chosen while doublebarreled requests, requests with an implicit assumption, and prerequests that change the concept can be easily avoided.

In this chapter we suggest why the use of batteries is so popular in survey research. The reason is mainly the efficiency of the formulation because of the request and the answer categories have to be mentioned only once. To our knowledge batteries with stimuli do not create problems, but batteries with statements have been criticized heavily by different authors. One reason is the possibility of response set or acquiescence that can generate correlations that are due to the method (use of a battery) and have no substantive meaning. Another problem is that the choice of the statements is rather arbitrary, but the choice will certainly have an effect on the response distributions and most likely also on the correlations with other variables.

Furthermore, several characteristics of requests for an answer have been discussed which may play a role in the quality of an item. First, the choice between absolute and comparative judgments has been discussed, followed with considerations for the choice between balanced and unbalanced requests. Whether we can say that one characteristic is indeed better than another requires further research. But as far as we know, the balancing of the requests by survey researchers does not seem to be based on empirical evidence while at the same time balancing the requests makes the formulations much more complex.

Finally, it was mentioned that sometimes researchers include in the texts requests to stimulate respondents to give answers or to give their own opinions. These choices also require further research to determine whether adding them to texts has a positive effect on the results.

Exercises

1. Look at the following request for an answer:
   Do you have the feeling that homosexuals should have the same rights with respect to marriage and raising children?
   a. What do you think that the researcher wants to measure?
   b. What went wrong in the formulation of this question?
2. Formulate a battery for human values using the following value stimuli: honesty, love, safety, and power.
3. Formulate a battery for human values using the same values mentioned in exercise 1, but this time make a statement for each of these values.

4. Several alternative statements can be formulated; indicate how these different statements can be created for the value "honesty."
5. Which of the statements you created in exercise 3 is the best?
6. How would a human value request be formulated for the value "honesty" using direct requests?
7. Is it possible to formulate this request in an absolute and in a comparative way?
8. Is your request balanced? If so, when could it be considered unbalanced? If not, how could it be balanced?
9. Can you also add texts to the last statement to stimulate a response and to emphasize that a personal opinion is asked?
10. With respect to your own questionnaire, discuss whether you have made the best choices while considering the abovementioned options? If so, why?

CHAPTER 5

Response alternatives

So far, we have been discussing requests for answers. As was indicated in Chapter 3, the requests can have many different forms, which in turn can create the same response alternatives for the respondent. However, the fact that the same response possibilities are present does not mean that the requests for an answer measure the same thing. Along the same line, it is not immediately clear whether requests for an answer that are identical but differ in the set of possible responses measure different variables. This is an empirical question which has to be answered for different measures. Saris (1981) showed that at least some sets of response scales, although different, will give responses that are identical, except for a linear transformation suggesting that roughly speaking, these measures are indeed identical.

Another issue studied by many people is whether it makes sense to present the respondents with more than only a few categories. Most textbooks suggest, in reference to Miller (1956), that people can not use more than approximately 7 categories. Cox (1980) has argued that Miller's rule does not apply at all to this problem. He suggests that more information can be obtained if more categories are used. This opinion is shared by a few more researchers (Saris et al. 1977; Andrews 1984; Alwin 1997; Költringer 1995).

Finally there are people who suggest that it would be advisable for certain problems [Krosnick and Fabrigar (forthcoming)] or in general in qualitative research, not to use explicit response alternatives. They suggest that requests with open answer categories are the best because they do not force the respondents in the frame of reference of the researcher.

All these options will be discussed below. The arguments pro and con will be mentioned and an empirical evaluation of the effects on data quality of the different possibilities will be given in Part III of this book.

5.1 Open requests for an answer

As has been mentioned above, some people argue that requests with open answer categories are better than requests with closed categories because people can follow their own thoughts and are not forced in the frame of refer-

ence of the researcher. A request that is exemplar for this dilemma and which has been studied frequently is as follows:

5.1 *What is the most important problem that our country is confronted with nowadays?*

This request can be asked as an open request as indicated above or with possible responses, chosen on the basis of prior research based on the open request. A comparison between these two requests has been studied several times by Schuman and his colleagues. Schuman and Presser (1981) reported that the results from the two requests are very different. The open request seems to be influenced by events that were recently discussed in the media, while the request with response categories provides a frame of reference indicating what is expected from the respondent. The option of "other" category along with a set of responses can be introduced but it turns out that this option is not chosen as frequently as expected. Hence, the authors concluded that the given response categories of a request guide respondents in their answer choices.

Subsequent research by Krosnick and Schuman (1988) suggests that there is more consistency across the open and closed request results if the coding of the answers of the open request is more in line with the categories used by less-educated people. This brought Krosnick and Fabregar (1997) to conclude that open requests are preferable because the effect of the researcher on the result is avoided.

The last statement may be correct for the abovementioned type of request, where a choice out of a multitude of nominal categories is requested, however, the findings need to be investigated further to determine whether they are also true for other open requests for an answer. Therefore, let us explore some other possibilities.

Krosnick and Fabrigar (forthcoming) indicate in another chapter of their book that not all open requests can be trusted at face value. They discuss the open "WHY request and the validity of introspection." In psychology, introspection has been discussed at length by the different schools of thought where some scholars think that only people can know why they do things, and therefore they should be asked. Other scholars argue that answers based on introspection cannot be trusted. One of the reasons provided is quick memory loss of thoughts concerning the choices made. Therefore a "think aloud" procedure is suggested, but if one asks for arguments before or while people are making choices, this in itself can influence the process (Ericson and Simon 1984; Wilson and Dunn 1986) and most of the time rationalizations of the answer choice are provided. This is not only the view of the behaviorists like Skinner (1953), but also of scholars with a less extreme point of view (Nisbett and Wilson, 1977).

Krosnick and Fabrigar (1997), while applying this bulk of research on survey research, comment "... if results based on introspection requests seem sensible on their surface, we would all be inclined to view them as valid. And yet, as Nisbett and Wilson (1977) and Wilson and Dunn (1986) have made clear, this

apparent sensibility may well be the result of people's desire to appear rational, rather than the result of actual validity of introspection." Therefore Krosnick and Fabrigar (1997) clearly indicate their reservations with the use of introspection procedure with the open request for an answer method. One should, however, also remark that formulating alternative procedures for introspection is not very easily done.

Wouters (2001), in her research, has specified open requests for all kinds of combinations of concepts and request forms that have been mentioned in Chapters 2, 3, and 4. For example one could ask:

5.2 *How would you evaluate the presidency of Clinton?*

It is clear that an evaluation is asked but the possible responses are not specified. So, this is an open request, and the respondent can give an answer in many different ways. In a similar way Wouters (2001) was able to transform nearly all possible closed requests into open-ended requests for answers. Hence the pertinent question is which of the two forms is better. To answer this question, a lot of research is still needed. Presently, we can say only that closed requests are more efficient than open requests because the former do not require an extra coding phase.

The analysis of Wouters (2001) also showed that it is not always simple to formulate a closed form for all open requests. We will demonstrate our point with the following example:

5.3a *What do you think about the presidency of Clinton?*

Example 5.3a is an open-ended request, however what is special about this request is that it does not measure a specific concept because respondents can answer with an evaluation (good or bad) but also with a cognition (that Clinton's government was the first to balance the budget) or a relationship (that Clinton's presidency led to an impeachment procedure) as just a few examples of possible answers. Not only is the answer open-ended but also the concept itself that is measured. Our hypothesis is that such requests are used to determine what aspect of the object the respondents consider the most important from which are derived further requests about this aspect. If that is true, an alternative in closed form to the open-ended request could be

5.3b *What is for you the most important aspect of the presidency of Clinton?*

1. *His foreign policy*
2. *His national policy*
3. *His economic policies*
4. *His personal conduct*
5. *Others*

Another type of open request that is hard to formulate in closed form concerns the enumeration of different events of actions. An example is

5.4 *Can you describe the different events that took place before the demonstration changed into a violent action?*

Here the respondent has to provide a series of events that have occurred in sequence.

From example 5.4 it can be inferred that asking an equivalent request in closed format would require a very different and complex series of requests.

Another type of request for an answer that requires special attention is a request for a frequency or an amount. Examples are found below:

5.5a *How many hours did you watch TV last night?*

5.5b *How much did you pay for your car?*

These requests are in some sense the opposite of the open requests we have discussed above, because now it is very clear how the respondents have to answer. The first request asks for a number what indicates the number of hours they have watched TV, and the second asks for a monetary amount. So people know quite well how they should answer, but nevertheless the answer is open because no response options have been provided to them (Tourangeau et al. 2000). For these requests that ask numeric answers, closed alternatives have been formulated. They will be discussed in further detail in the section on vague quantifiers.

It might depend on what request type we are about to use whether we choose an open or closed form. For most open requests alternatives in closed form exist; for others, alternative closed requests are difficult to formulate. For those requests that can be asked in a variety of ways, different aspects should be considered. First, it is important to consider whether more information is obtained through using the open request format. If that is not the case, then it is better to choose the closed form because the processing of the information is much easier. A second issue is, whether open and closed requests lead to different response distributions and relationships with other variables. If that is the case, one has to consider which request form is better. Evaluation of the effects on the data quality will be discussed later. If the same results are obtained or the quality is not clearly better for the open requests, then the closed requests should be preferred because of the efficiency in information processing. It will be clear that in our opinion the conclusion of Krosnick and Fabrigar (1997) is still premature and we think that further research is required before a conclusion about the choice between open and closed requests can be stated with certainty. We speculate that the request choice will depend on the type of issue the request is aiming at as was the case with our examples.

5.2 CLOSED CATEGORICAL REQUESTS

The first of the requirements regarding closed response answer categories is that they should be *complete*. In practice, however, sometimes the answer alternatives are not complete, which can result in nonresponse. Such an example is given below:

5.6a What is the composition of your household?
1. *One single adult*
2. *Two adults*
3. *Two adults and one child*
4. *Two adults with two children*
5. *Two adults with three children*
6. *One adult with one child*

After scanning the answer options for 5.6a, it becomes clear that the answer categories are not exhaustive since there are several variations of adults and children possible and one for communes is missing. Hence 5.6b is a more complete version:

5.6b What is the composition of your household?
1. *Number of adults ...*
2. *Number of children ...*

The second requirement is that the answer categories are *exclusive*, or in other words they should not overlap. An example of overlapping answer categories is found in request 5.7a:

5.7a What is the most important reason why you are against nuclear energy?
1. *Too expensive*
2. *Too dangerous*
3. *Causes environmental problems*
4. *Other*

In request 5.7a the second and third categories are not exclusive because environmental problems can cause dangers and dangers, like radioactive waste, can cause environmental problems. Therefore, a respondent may be confused about which choice to make. The remedy is to reformulate these two categories in order to make them exclusive:

5.7b What is the most important reason why you are against nuclear energy?
1. *Too expensive*
2. *The probability of an accident is too high*
3. *Too much radioactive waste*
4. *Other ...*

Here the second category focuses on accidents and the third, on radioactive waste, which are now distinct and no longer overlap.

A third requirement is that answer categories *match* with the information provided in the request or statement asked (Lessler and Forsight 1996; Graesser et al. 2000a,b):

5.8a How far do you agree or disagree with the statement that governmental decisions are always carried out
1. *Completely agree*
2. *Agree*
3. *Neither agree nor disagree*
4. *Disagree*

In the example the statement refers to an objective concept (a behavior), while the answer categories relate to subjective concepts. The appropriate answer categories would be "true/false." The request could be reformulated in the following manner:

5.8b Do you think that the following statement is true or false? Governmental decisions are always carried out.
1. *True*
2. *Neither true nor false*
3. *False*

Finally, a requirement is that all the response categories represent *the same concept*. Sometimes a mismatch of answer categories occurs because they concern different concepts and then it is difficult for the respondent to choose a category. Example 5.9 illustrates a case where this is not correct:

5.9 What is your opinion about a ban on driving a car in downtown area?
1. *Inconvenient*
2. *Acceptable*

The first category refers to a feeling, while the second is a right. In order to be consistent, it is possible to provide either a feeling (unpleasant/pleasant) or a right (acceptable/unacceptable) as options of the uncertainty space. All requests for an answer with closed answer categories should satisfy the abovementioned requirements.

In the following sections we want to illustrate the different types of response categories that are available to the survey designer. The first type uses nominal categories without any ordering, while the second type provides ordinal response categories and the third consists of what is called vague quantifiers.

5.2.1 Nominal categories

Requests for an answer using unordered response categories are an alternative for the open requests asking for one option out of a set. An example is

5.10 What is the most important problem that our country faces at the moment?
1. *Terrorism*
2. *Unemployment*
3. *Racism*

4. *Criminality*
5. *Other, please specify ...*

Similar requests can be asked for the most important aspect of the work and many other topics. There is no ordering in the different response possibilities even though they can be numbered in the questionnaire and certainly in the database, but, the numbers cannot suggest an ordering on any dimension because that dimension does not exist. Response scales that are not ordered are called *nominal* scales.

A special nominal scale is a scale for *dichotomous* responses where only two answers are possible for example:

5.11 Did you vote in the last elections?
1. *No*
2. *Yes*

In this case the scale is officially nominal, indicating no ordering. However, it is possible to use the scale in the ordinal sense and apply analyses that at minimum require ordinal data and it is arbitrary if the coding by the researcher is completed as 0–1 or 1–2 for the dichotomous scale.

5.2.2 Ordinal scales

Ordinal response categories require that there is an ordering of the response categories. Such sets of response alternatives are very common in subjective judgments. For example

5.12 How good do you think Clinton was as president?
1. *Very bad*
2. *Bad*
3. *Neither good nor bad*
4. *Good*
5. *Very good*

In this case there is an ordering in the response categories, and one can say that the numbers in front of the categories suggest an ordered scale where 1 is the lowest and 5 is the highest category. Similar scales can be made with any predicate with "high" and "low," "friendly" and "unfriendly," "active" and "passive," to name only a few examples.

Although such an ordinal scale is called a 5-point scale – a scale with 5 possible answers – a person with a positive evaluation of Clinton has only two possibilities: good or very good. If it is desirable to have a more precise answer, it can be specified as a 7-point scale such as the one below:

1. *Very bad*
2. *Rather bad*
3. *Bad*
4. *Neither good nor bad*
5. *Good*

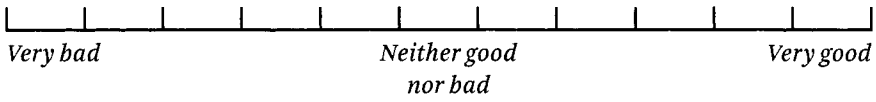
6. *Rather good*
7. *Very good*

Along the same line one can also construct a 9- or 11-point scale. Keep in mind that there is a limit to the possibilities of labels for the different categories, and that it is also possible to specify ordinal scales with labels for only a limited number of categories. Common examples are the following:

5.13a How good do you think Clinton was as president?
Express your opinion in a number between 0 and 10, where
0 = very bad and 10= very good

or

5.13b How good do you think Clinton was as president?
Express your opinion by placing an x at the point of the scale that expresses your opinion the best



Examples 5.13a and 5.13b are both 11-point scales, the distinction is that the former has only two labeled categories while the latter has three labeled categories; and that the first request uses numbers while the second is a typical example of what is called a *rating scale.*

Many alternative presentations can be developed with ordinal response scales. What is important is that the categories are ordered in some way from low to high. It can also be done by pictures of faces that are more or less happy or ladders where each step indicates a different level of satisfaction (Andrews and Withey 1974) or a thermometer where the increasing grades indicate the warmth of the feelings of respondents toward parties and party leaders. The United States' National Election Studies are exemplar for this type of creative ordinal response scale grading.

When developing ordinal scales a range of decisions is at the researchers' disposal. We will discuss some of these choices with their alternatives. First, we have seen that either all or some of the possible responses can be labeled. Therefore, the responses can be *completely labeled or partly labeled.*

In example 5.13a the numbers in front of the categories were ordered in the same way from low to high as the labels and they started with the lowest or most negative category. It can also happen that there is no *correspondence* between the category labels and the numbers or that the scale does not go from low or negative to positive but vice versa.

All the scales presented so far are *symmetric* around the middle of the scale, which means that there are as many categories at the positive as at the negative side.

In general it is advisable to use symmetric scales; the reason can be demonstrated by the example:

5.14
1. *Very unhappy*
2. *Unhappy*
3. *Neither unhappy nor happy*
4. *Happy*
5. *Rather happy*
6. *Very happy*

This example demonstrates that it appears awkward to be using an asymmetric scale in this case. However, if we know that all respondents' answers are on the happy side of the scale, it is not very efficient to use a 5-point scale from "very unhappy" to "very happy" because the distribution of happiness in the population is reduced to a 2-point scale. Therefore, an asymmetric 5-point scale is more appropriate and precise:

5.15
1. *Not happy*
2. *A bit happy*
3. *Happy*
4. *Rather happy*
5. *Very happy*

Example 5.15 has a 5-point scale that favors the positive side, while the "not happy" side of the scale, is represented by only one response category. Such a scale presupposes knowledge about the happiness of a survey population; otherwise, such an asymmetric scale is biased.

So far, except for in the last example, all sets of response scales were also *bipolar*, which means that there are two opposite sides of the scales: positive to negative or active to passive. The last scale of happiness was made one-sided or *unipolar*, but happiness itself is in principle a bipolar concept, going from unhappy to happy. Therefore we also said that the unipolar scale presupposed knowledge of the distribution of feelings within the population. There are, however, also concepts that are typically unipolar. For example, "attachment to a party" goes from "no attachment" to "strong attachment" because it is impossible to imagine a negative side of the scale of attachment.

The discussion above has served to demonstrate that both the provided scale for responses and the concept can be in *agreement* with each other (both bipolar or both unipolar) or in disagreement if the concept is bipolar, but the responses are only unipolar, as in example 5.15.

So far we have used a *neutral category or a middle category*, but it is not always necessary to do so. If it is necessary to force people to make a choice in a specific direction, then the middle category can be omitted. Schuman and Presser (1981) have shown that this has no effect on the distribution of the respondents over the positive and negative categories. However, it might have the effect that fewer people are willing to answer the request because, according to them, their response is not provided and consequently they choose for a "don't know" or "refusal" (Klingemann 1997).

The *"don't know"* category has been the subject of serious investigation. Research has centered around the question of whether it should be offered, and if so, in what form. One can ask, for instance, before the request itself is asked whether people have an opinion or not about the topic in question. This is the most explicit "don't know" check. The second possibility is to provide "don't know" explicitly as one of the response options. The third possibility is that "don't know" is not mentioned but that it is an admissible response alternative that can be found on the questionnaire of the interviewer but is not mentioned as a possibility to the respondent. Finally, there is the possibility of omitting it altogether.

Providing the "don't know" option explicitly creates several obstacles. The most important issue is that respondents can choose this option for several reasons which have nothing to do with their own opinion. Krosnick and Fabrigar (forthcoming) mention that this option is chosen because respondents don't want more requests or because they do not want to think about the request and therefore an acceptable option "don't know" is easily available. The authors call this "satisficing behavior of a respondent."

Schuman and Presser (1981) argue that people who normally would say that they "don't know" would make a difference in the relationships between variables under investigation. They report on a study where without respondents using the "don't know" category the correlation between two variables was close to zero while with them it went up to .6.

Another problem with people choosing "don't know" is that fewer representatives of the population are left for the analysis. If the option is available for several requests, the number of people with complete data on a larger set of variables can decrease and it becomes questionable whether the respondents who are left in the sample are on the whole representative for the population. These three arguments have led researchers to allow for the "don't know" option, but only if the respondent explicitly asks for it. However, whether this is the most scientific course of action, we will evaluate later.

So far the focus of our discussion has been specification of response categories for subjective variables. However, ordinal response categories are also used for objective variables such as the frequency of activities or categories of income and prices. An example could be

5.16a How often do you watch TV during a week?

1. *Very often*
2. *Often*
3. *Regularly*
4. *Seldom*
5. *Never*

If we had omitted the response alternatives, this could have been an open-ended request, but researchers often add response categories to such requests and the issue is that respondents can differ in their interpretation of the

different labels: what is "often" for one person means "seldom" for another. It all depends on the reference point of the respondent. Therefore these ordinal scales are called *vague quantifiers*. We could have also asked the following:

5.16b How often do you watch TV during a week?

1. *Every day*
2. *5 or 6 times*
3. *3 or 4 times a week*
4. *1 or 2 times a week*
5. *Never*

This request is more precise and less prone to different interpretations. Even so, 5.16b is an ordinal scale because it is not clear what numeric values the categories 2–4 represent.

Table 5.1: The results of Hippler and Schwartz with respect to TV watching

Categories at the low side		Categories at the high side	
Categories	**Percentage of respondents**	**Categories**	**Percentage of respondents**
< ½ hour	11.5		
½ – 1 ½ hours	53.8		
1 ½ – 2 ½ hours	34.7	< 2 ½ hours	70.6
> 2 ½ hours	0.0	> 2 ½ hours	29.4
Total	100		100

Similar scales can be used for income and prices with the option of using vague quantifiers or more precise category labels. Hippler and Schwarz (1987) made a remarkable observation when they varied the category labels in an experiment about the amount of time people watch TV. In it they did not use vague quantifiers like those of example 5.16a but two different and separate categorizations for the number of TV viewing hours. Their results are presented in Table 5.1. The table shows that the different categories had a considerable effect on the responses. Their explanation was that respondents do not have an answer readily available for this type of request. Instead, they use the response scale as their frame of reference. Respondents estimate their TV watching time on whether they view themselves as more or less TV watching than other persons. Therefore, if they consider that they watch more TV than others, they will choose the high end of the scale and vice versa. This experiment shows that even for objective variables the answers do not represent absolute judgments but relative judgments. It has been suggested that people always make relative judgments. If that is so, it is better to adjust the approach of asking requests to the human judgment factor. We will investigate this problem in the next section in more depth.

5.2.3 Continuous scales

Another form for response possibilities is to give respondents instructions to express their opinions in numbers or lines. This approach was developed with the idea that it would result in more precise information than would the other methods discussed in previous sections.

Originally such approaches were used in psychophysics. For an overview we refer to Stevens (1975). Presently these measurement devices have been introduced in the social sciences by Hamblin (1974), Saris et al. (1977), Wegener (1982) and Lodge (1981). When these approaches were introduced special procedures, called *magnitude estimation* and *line production*, were used. The basic idea is rather simple and will be illustrated by several examples of procedures used in practice. Originally a request for an answer was formulated as follows:

5.17a *Occupations differ with respect to status. We would like to ask you to estimate the status of a series of occupations. If we give the status of the occupation of a schoolteacher a score of 100, how would you evaluate the other occupations? If an occupation has a status that is twice as high as that of a schoolteacher, give a twofold larger number or 200. If the status of the occupation is half that of a schoolteacher, divide by 2, which gives 50.*

What is the status of a physician? ...

.

.

.

Of a carpenter? ...

And so on.

People are asked to match the ratios of status judgements with the ratios of numbers. This could also be done using "line production," as has been shown in the following instruction:

5.17b *Occupations differ with respect to status. We would like to ask you to estimate the status of a series of occupations. We express the status of the occupation of a schoolteacher by a standard line as follows:*

If an occupation has a status that is twice as high as that of a schoolteacher, draw a line which is twice as long. If the status of the occupation is half of that of a schoolteacher, draw a line that is half the size of the standard line.
What is the status of a physician?

.

.

Of a carpenter?

With these procedures a striking precision of responses was obtained (Hamblin 1974; Saris et al. 1977). However, in their embryonic stage these approaches were used only for evaluation of stimuli, as we have previously indicated. Currently other concepts are also measured in this way. For example, we could reformulate the frequently asked satisfaction request using continuous scales as follows:

5.18 a How satisfied are you with your house? Express your opinion with a number between 0 and 100, where 0 is completely dissatisfied and 100 is completely satisfied.

This request differs in several points from the original instruction. The first point is that the ratio estimation is no longer mentioned. The reason is that the results are not very different whether one gives this instruction explicitly, while at the same time, omitting this instruction makes the formulation much simpler. The second point is that two *reference points* have been mentioned instead of just one. This is due to research showing that people use different scales to answer these requests if only one reference point is provided, while using two reference points it is less of a concern (Saris 1988b). A condition for this conclusion is that *fixed reference points* are used. With fixed reference points, we mean that there is no doubt about the position of the reference point on the subjective scale in the mind of the respondent. For example, "completely dissatisfied" and "completely satisfied" must be the endpoints of the opinion scale of the respondent. If we would use "dissatisfied" and "satisfied" as reference points, then respondents may vary in their interpretation of these terms because some of them see them as endpoints of the scales while others do not.

The disadvantage of using numbers is that people tend to use numbers which can be divided by 5 (Tourangeau et al. 2000). This leads to rather peaked distributions of the results. This can be largely avoided by the use of line length instead of requesting a numerical evaluation. For request 5.18a the instruction, using line length as response mode, would be as follows:

5.18b How satisfied are you with your house? Express your opinion in length of lines, where completely dissatisfied is expressed by the following line:

–

and completely satisfied by the following line:

Now express your opinion by drawing a line representative of your opinion:

The disadvantage of the line production is, of course, that later the lines need to be measured. This is a challenge if paper-and-pencil procedures for data collection are used but with computer assisted interviewing (CAI) the programs can measure the length of lines routinely.

Although these methods gained some popularity around the 1980s, they are still not frequently employed. One reason is that researchers want to continue with existing measurement procedures and do not want to risk a change in their time series due to a method change. Another reason is that several researchers have argued that the lines do not increase precision a lot. The most outspoken author is Miethe (1985). Some other people (Alwin 1997; Andrews 1984; Költringer 1995) do not agree with Miethe's argument, and they have shown that better data are indead obtained if more categories are used. In the next section we will argue why we think that it is better to use more than 7- point category scales and why we prefer line drawing scales as a standard procedure.

Before moving to the next section we should clarify a point about the measurement level of continuous scales. So far we have discussed nominal scales and ordinal scales, however, it is interesting to know what kind of measurement level is obtained using the continuous scales discussed here. One may think that the scales discussed represent ratio scales given the ratio instructions originally requested. However, Saris (1988b) has found that the line and number responses are nearly perfectly linearly related (after correction for measurement error and logarithmic transformation) and he concludes that on the basis of these results the measurement level of these continuous scales is log-interval (Stevens, 1975). This means that the data obtained with the suggested response procedure, after logarithmic transformation, can be analyzed using interval-level statistics. From this it follows that continuous scales have a higher measurement level than do the previously discussed category scale procedures.

5.3 How many categories are optimal

Most researchers are in agreement that it is better to use more than two categories if it is possible and they are even inclined to accept that 7-point scales are even better. For example, Krosnick and Fabregar (1997) make this recommendation very explicitly and conclude not to use more categories. Several studies share this opinion and they have tried to indicate that people can not provide more information than suggested by a 7- point category scale.

However, we are of the opinion that respondents are capable of sharing more information. This can be shown by asking people the same judgment 3 times: once expressed on a category scale and once expressed in numbers and once expressed in lines. If people did not have more information than can be expressed in the number of categories of the scale, the correlation between line and number judgments of stimuli placed in the same category of the category scale would be zero. This is, however, not the case. The correlation between the line and number responses of stimuli that all received the same categorical scale score, can go as high as .8. This reveals that people have indeed more information than they can express in the verbal labels of the standard category scales (Saris 1998).

Why this extra information normally is not detected has to do with the problem that the respondents may use different scales in answering requests even from one occasion to the next. Saris (1988b) calls this type of phenomenon "Variation in the response function." He suggests that respondents answer very precisely, but in their own manner. Figure 5.1 illustrates this phenomenon.

In this figure respondent 1 expresses herself in rather extreme words compared to the others: if she has an opinion which is close to 0, she also gives responses close to zero, and if she has an opinion close to 100, she also gives responses close to 100. The other two respondents give much more moderate responses even though they have the same opinions. Of course, this is just a fictional illustration of the problem. For empirical illustrations we refer to Saris (1988a). In this illustration we have assumed that all respondents will give the response 50 if they have an opinion of 50 about the evaluated stimuli. In practice, this is only necessarily so if one reference stimulus is provided with a standard response of 50; otherwise this point will also vary across respondents.

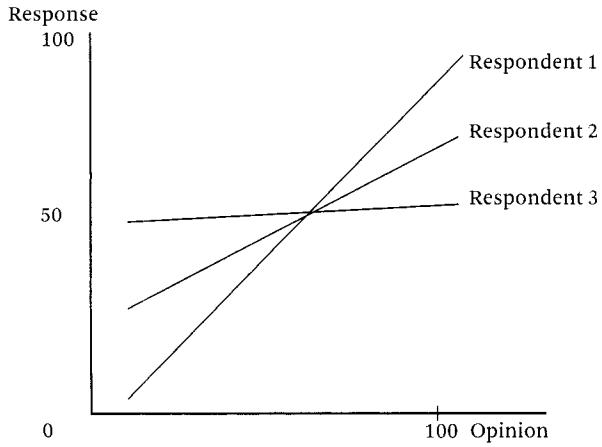


FIGURE 5.1: *Variations in the response function*

Let us now look at what happens when only one stimulus is provided for which all respondents have an opinion of 100. In accordance with Figure 5.1, we see that the respondents will give a different response even though they have the same opinion. This means that the varying responses cannot be explained by substantive variables. They are a consequence of the differences in response function and could be mistakenly interpreted as measurement error. This is a problem for researchers because this kind of variation will occur while the respondents may have very precise, reliable responses if you look at their individual data.

The variation in responses due to variation in response function is larger at the extreme ends of the scale than closer at the middle. This phenomenon can explain that extension of scales with more categories, for example, above 7, will increase what is seen as measurement error, and it is for this reason that many researchers believe that they do not gain more information by increasing the length of the scales.

On the basis of our research with respect to the amount of information that people can provide and the problem of variation in response functions we would like to suggest that people often have more information than they can express in the labels of the standard 7-point category scales, but increasing the number of categories also increases the problem that respondents will start to use their own scale. The latter problem can be reduced by the use of more than one fixed reference point. If two fixed reference points are given on the response scale, then the endpoints of the opinion and response scale are the same for all people and if a linear response function is used, the responses will be comparable. It has been shown that in that case the variation in response functions is indeed smaller. In this way it is possible to obtain more information from respondents than using response scales with 7-point scales (Saris and De Rooy 1988).

That such procedures are not so difficult to formulate has been illustrated above because the last examples of continuous scales (examples 5.18a and 5.18b) provided in the previous section satisfied the abovementioned criteria. It was also mentioned there that the line production is the better procedure because the respondents will not round off their answers, when using the line method. In Part III of this book, where we discuss the empirical evidence for the effects of the different choices that we discuss here, we will come back to this issue.

5.4 Summary

In this chapter we have discussed the different options that exist with respect to the specification of the uncertainty space or the set of possible responses. We have seen that some researchers do not recommend explicitly specifying response options. However, we are not of the same opinion. We would say that depending on the context, an open request for an answer may be preferable to a closed request. On the other hand, open requests are much less efficient because the answers have to be coded, but the advantage of open requests is that people are not forced into the frame of reference of the researcher.

One type of open request, the "WHY requests," was given special attention in this chapter because it is commonly used. However, we share Krosnick and Fabrigar's (forthcoming) view in not recommending this type of request because respondents may be led into rationalizations and may not give their true reasons for the answer. It was also shown through a research review that introspection is not a very scientifically valid procedure.

Furthermore, we have seen that there are some requests that are difficult to

translate into closed request form, such as open requests about sequences of events and open requests that are open with respect to the concept measured. In those specified cases open requests are probably the preferred method. Therefore, it depends on the topic, context, and researcher's intent, whether open or closed requests should be selected for a request for an answer.

With respect to closed requests a distinction was made between nominal and ordinal categorical response scales, and continuous response scales. There are many forms of categorical scales, especially ordinal scales. Several examples were discussed. In doing so, we introduced choices that are connected with the development of such scales such as:

- Correspondence between the labels and the numbers of the categories
- Symmetry of the labels
- Bipolar and unipolar scales and agreement between the concept and the scale
- The use of neutral or middle category
- The use of "don't know" options
- The use of vague quantifiers or numeric categories
- The use of reference points
- The use of fixed reference points
- The measurement level

Furthermore, we introduced the advantages and disadvantages of choosing the number of possible responses. Our logical argument is that more information can be obtained than is possible in the standard 7-point category scales if we allow respondents to provide more information. However, in order to obtain responses that are comparable across respondents at least two fixed reference points need to be specified in the response procedures that are connected to the same responses across all respondents. In this context we suggested that line production scales provide better results than magnitude estimation; since respondents have a tendency to prefer numbers that can be divided by 5, this leads to peaked response distributions and this does not happen with line production scales.

It should not be concluded that the line production scales should be used for all topics and at all times. If researchers don't need more information than "yes" or "no," it does not make sense to force the respondents to use a continuous scale. Also the continuity in survey research often requires the use of the standard category scales. The continuous scales may have a future when computer-assisted interviewing becomes more popular.

EXERCISES

1. Below is an example of a request for an answer:
   All in all, nowadays are you feeling very happy, quite happy, not so happy, or not at all happy?
   1. Very happy

2. *Quite happy*
3. *Not so happy*
4. *Not at all happy*

What can you say about this response scale with respect to

a. The correspondence between the labels and the numbers of the categories?
b. The symmetry of the labels?
c. The bipolar and unipolar scales and agreement between the concept and the scale?
d. The use of a neutral or middle category?
e. The "don't know" option?
f. The use of vague quantifiers or numeric categories?
g. The use of reference points?
h. The use of fixed reference points?
i. The measurement level?

2. Could you reformulate the request in order to improve the quality of the request in the light of the evaluation on the different characteristics mentioned in exercise 1?
3. Is it also possible to formulate this request in an open request form? If so, how?
4. Is it also possible to formulate this request using continuous scales? If so, how?
5. Which of the three scales would be the most attractive one and why?
6. One could also have asked: *How are you these days?*
   a. Do you see a problem with this request?
   b. Is it possible to reformulate this request in a closed form?
7. Now look at your proposal for a questionnaire. Do you think that you have chosen the best response categories? If not, make improvements and indicate why you have made these improvements.

CHAPTER 6

The structure of open-ended and closed survey items

So far we have discussed the basic form of requests for an answer, but often they are placed in a larger textual unit called a "survey item," which consists of an entire text that requires one answer from a respondent (Saris and de Pijper 1986). Andrews (1984) defined a survey item as consisting of three different parts of text or components, namely, an introduction, one or more requests for an answer, and a response scale. Molenaar (1986) uses quite similar components. In this chapter we propose distinguishing even more components of a survey item. First, we will describe the components and thereafter we will present different structures of survey items for open and closed requests. The structure of batteries of requests for an answer, such as those using stimuli or statements will be the topic of Chapter 7. We close this chapter with a discussion of the advantages and disadvantages of the different forms of open-ended and closed survey items.

6.1 Description of the components of survey items

Figure 6.1 shows the basic components of a survey item. The reader should notice that we make a distinction between parts embedded in the request for an answer as discussed before and parts that can be juxtaposed before or after the request for an answer.

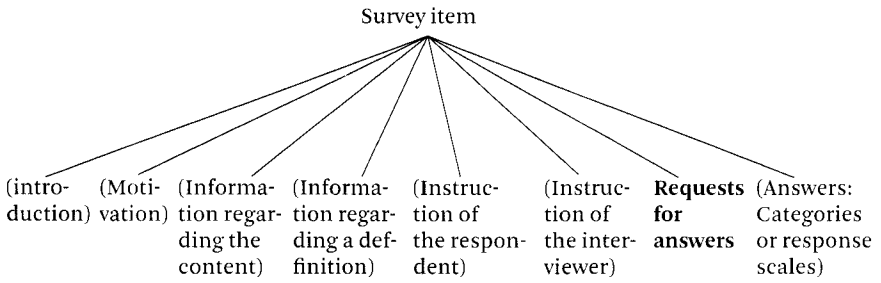


FIGURE 6.1: *Decomposition of a survey item into its components*

In our opinion the following parts can be added: an introduction, a motivation, an information regarding the content, information regarding a definition, an instruction of the respondent, an instruction for the interviewer, the request for an answer and response categories or scales, as shown in Figure 6.1. The components indicated within parenteses in Figure 6.1 are optional. This implies that the request for an answer is the core unit of a survey item, and it also means that the simplest form of a survey item is just an open request for an answer and nothing more. The figure also demonstrates that a survey item can consist of many more components. How many and which ones, will be discussed further. But, first, we begin with a description and illustration of the different components.

Introductions (INTRO) are meant mainly to indicate the topic of the request for an answer to the respondent. In general they consist of one or more sentences. Examples are as follows:

6.1 *Now, a couple of questions follow about your health.*

6.2 *The next question is on the subject of work.*

Sometimes two requests for an answer are formulated and the first request functions just as an introduction because no answer is asked for it. The second request for an answer is the one to be answered that is indicated by the answer categories. Examples 6.3 and 6.4 are an illustration:

6.3 *Would you mind telling me your race or ethnic origin (INTRO)? Are you white, black, Hispanic American, Alaskan native, Asian, or Pacific Islander?*
1. White but not Hispanic
2. Black but not Hispanic
3. Hispanic
4. American Indian or Alaskan native
5. Asian or Pacific Islander

6.4 *What is your opinion on each of the following proposals (INTRO)? Could you tell me if you are for or against it?*
There should not be death penalty anymore.
1. For it
2. Against it

The next component of a survey item we introduce is called *motivation (MOTIV)*. This part of text explains the broader purpose of the research to stimulate the respondent to answer the question(s). It consists of one or more sentences and contains keywords like "purpose," "research," "representative." Examples 6.5 and 6.6 demonstrate our point:

6.5 *We are doing research to find out the best way to ask questions.*

6.6 *For the statistical processing of a survey, it is important that the research be representative for the entire population. In order to obtain this, we need to know the general range of incomes of all people whom we interview.*

Information regarding the content (INFOC) clarifies or explains something about the content of the survey item. It is included in a survey item because many people do not have an opinion about many issues (Converse 1964). Linguistically it consists of one or more sentences. Examples 6.7–6.9 illustrate this concept:

6.7 *The European Political Union may include a common defense the arrangement involving the member states of the European Community. Successive Irish governments have accepted that moves toward the European Political Community could mean scraping Ireland's policy of military neutrality.*

6.8 *There are different ways by which people can show their disagreement with the measures employed by the government.*

Frequently the explanation or the clarification contains arguments for and/or against a point of view. Kay (1998) has used this approach to test the stability or strength of opinions. Example 6.9 is an illustration:

6.9 *Since the crime rate among young people has been drastically increasing over the last few years, some citizens and political parties think that the government has to take strong action against crime.*

However, example 6.9 provides only one-sided information. Using such information, one can get very different results depending on the information given. If arguments are given, they should include arguments for both points of view (Sniderman and Theriault 2004), as is done in example 6.10:

6.10 *As you probably know some politicians fear that within a few years they will have to deal with an energy shortage. They therefore propose building more nuclear reactors. Other politicians warn about the dangers of nuclear energy and therefore suggest that no new reactors should be built.*

Saris et al. (1984) have developed a choice questionnaire using this approach in order to solicit *well-considered opinions* from respondents. For an elaborate discussion of this approach and its evaluation, we refer the reader to Neijens (1987).

We also defined a component about *information regarding a definition (INFOD)*. This part of text defines some "concept" used in the survey item like "abortion" or "euthanasia" or some scales. It can consist of one or more sentences but frequently it is shorter than a sentence, implying that it is embedded in another component, which is often an instruction to the respondent or a request for an answer. Illustrations of this component type might look like examples 6.11–6.13:

6.11 *By abortion we understand the deliberate termination of a pregnancy by a physician.*
6.12 *The "net income" is the amount you receive after tax deduction.*
6.13 *You get a ladder with steps that goes from 0 at the bottom to 10 at the top, 5 is the middle step at the ladder. At the top of the ladder are the very best feelings you might expect to have, and at the bottom of the ladder are the worst feelings.*

The next two components relate to *instructions*. Researchers can give instructions to the respondents or the interviewers. Linguistically they are characterized by sentences in the imperative mood or polite variations of it. Here we discuss only instructions that are used outside the request for an answer. Instructions can also be used within the request for an answer that has already been discussed in Chapter 3. Examples 6.14 and 6.15 illustrate *instructions to the respondents* (INSTRR):

6.14 *What do you think of President Bush? Express your opinion in a number between 1 and 0, where 0 is very bad and 10 very good.*
6.15 *Look at the card and tell me which answer seems to fit your situation.*

In the first example the survey item begins with a request for an answer and continues with an instruction for the respondent (INSTRR). In the second example the positions are reversed. Examples of *instructions of interviewers* (INSTRI) are as follows:

6.16 *Hand over the show card. Only one answer is possible.*
6.17 *Read out the following text. If unclear, repeat instructions.*

The next component of a survey item indicated in Figure 6.1 is the *request for an answer (REQ)*. We will not repeat our discussion of this form because it has already been done detailed in Chapter 3.

The last component of a survey item presented in Figure 6.1 relates to *answer categories or response scales (ANSWERS)*. They are optional, as open requests for an answer do not require them and respondents have to give their own answers. Since Chapter 5 is entirely devoted to this topic, we only wish to alert you the presence of this component at this time.

6.2 DIFFERENT STRUCTURES OF SURVEY ITEMS

In this section we will discuss some different structures of survey items occurring in questionnaires as a consequence of researchers' choices. We will also indicate the position of the components in the item as far as possible. First we present the structures encountered in a number of selected questionnaires. For this purpose we used a sample of 518 Dutch survey items selected on the basis of a random procedure from a larger sample of 1527 survey items by Molenaar (1986: 34–44). Since this sample contains only requests for an answer with

closed answer categories, we added a sample of 103 open-ended Dutch requests for an answer from a database of the Steinmetz archive collected by Wouters (2001). A convenience sample of factual requests for an answer is studied on the basis of a collection of questionnaires from the Dutch Gallup institution NIPO, the Telepanel, and The Dutch Bureau of Statistics of the late 1980s and early 1990s. In order to compare these Dutch survey items with structures of English survey items, we collected 200 closed and open-ended requests for an answer that actually do not constitute a representative sample, from the Institute of Social Research (ISR, Ann Arbor, MI) questionnaires from the period 1979–1981, the Eurobarometer (1997) and survey items from Sniderman et al. (1991). Factual requests for an answer were also collected from Sudman and Bradburn (1983). A similar collection of 250 German survey items coming from surveys from the IFES Institute in Austria was also used for comparative purposes.

The abovementioned databases of survey items serve as an overview of different structures that occur in practice. At the end of the chapter we present a quantitative estimate of the frequency of occurrence of the structures for subjective variables on the basis of the random sample of survey items collected by Molenaar (1986).

In this chapter we will separately discuss two groups of survey items: open-ended requests for an answer and closed ones. This distinction is made because we expected a considerable difference between them.

6.2.1 Open-ended requests for an answer

First we illustrate the structure of an open-ended survey item that consists only of a request for an answer. There are no answer categories or rating scales mentioned since the request for an answer is open-ended. Examples 6.18–6.20 illustrate this type of structure:

6.18 *What is, in your opinion, the most important problem with which our country is confronted nowadays (REQ)?*

6.19 *Please, give me the reasons why you changed your job last year (REQ)?*

6.20 *How many hours a day do you watch television (REQ)?*

It will be obvious that the first two examples (6.18 and 6.19) are open-ended subjective requests for an answer where respondents are free to give their answers. Example 6.20 is a factual request for an answer, where the respondent provides the appropriate answer in terms of a number, which we also consider as open-ended since no answer categories are provided (Tourangeau et al. 2000). The following structures of open-ended requests for an answer contain two components. The first one that we illustrate consists of an introduction and a request for an answer.

6.21 *Now we would like to ask a question about your job (INTRO). What do you think of your present job (REQ)?*

In this example the first sentence is a typical introduction while the second sentence is an open request. Sometimes the introduction is also formulated as a question:

6.22 *What do you think about the political situation in Europe (INTRO)? Do you think that the developments go in the right direction (REQ)?*

The first request in this example must be seen as an introduction because no answer is expected from it. The second request in example 6.22 indicates that an evaluation is requested but as it is an open-ended request for an answer, answer categories are not provided.

Another structure of an open-ended survey item is used in combination with a closed request for an answer. This structure consists of three components, namely, a preceding closed request for an answer with answer categories either embedded or specified separately where an open-ended request for an answer follows. In this case an answer to both requests is expected. We mention this type of question because the closed question is normally used only as an introduction, while the actual open request is not complete as its content relates to the closed question. This can be shown by the following open-ended examples such as, "could you explain your answer?" or "could you tell me why?" which are not sufficient alone. An example of this combination is as follows:

6.23 *Do you think that in our country many or few people (ANSWER categories) make easy money (closed REQ)?*

1. *Many people*
2. *Few people (ANSWER categories)*

How did you get this idea (REQ open-ended)

It is also obvious that this open-ended request for an answer relates to the specific answer given and asks for more detailed information.

6.2.2 Closed survey items

We first will discuss the structure of a closed survey item that consists only of a request for an answer with explicitly mentioned answer categories. Examples could be as follows:

6.24 *Do you think that we should have a world government for all countries (REQ)?*

1. *Yes*
2. *No (ANSWER CATEGORIES)*

6.25 *Please tell me if your parents were members of a church (REQ)?*

1. *Yes*
2. *No (ANWSER CATEGORIES)*

These structures are rather normal in mail surveys or other self-completion surveys. In such surveys the response categories are presented after the

request for an answer and are not embedded in the request for an answer. In oral surveys presented by an interviewer such forms are rather unusual, except very simple requests for an answer with clear "yes" or "no" answer categories. In surveys presented by an interviewer the answer categories are also often presented on a card given to the respondent (showcard) or embedded in the request for an answer. This latter structure is the next case to be discussed.

In interviewer-administered surveys closed survey items consist of a request for an answer in which answer categories are embedded. They are mentioned in the interview form again after the request. The interviewer does not repeat them again. These second set of answer categories are presented for administrative purposes. Therefore in these cases we indicate their presence in the printed form of the questionnaire by enclosing them in parentheses. Examples 2.26 and 2.27 illustrate this structure:

6.26 Do you own or rent your home? (REQ + ANSWERS)
(1. I rent the home
2. I own my home) (ANSWER CATEGORIES)

6.27 Please tell me whether you consider our President a good or bad leader or neither good nor bad? (REQ + ANSWERS)
(1. Good leader
2. Neither good nor bad
3. Bad leader) (ANSWER CATEGORIES)

After some discussion of the simplest structures of closed survey items, we will show how structures of survey items can be expanded by adding components between the request for an answer and the answer categories.

The structures earlier mentioned can be expanded by inserting, for instance, an instruction for the respondent between the request for an answer and the answer categories:

6.28 Are you working in the household or do you have a paid employment (REQ+ ANSWER categories)?
Please, mark only one answer category. If you engage in both activities choose the one you consider most important (INSTRR).
(1. Works in household
2. Has paid employment) (ANSWER CATEGORIES)

The inserted component can also be an instruction to the interviewer that leads to the following example:

6.29 What is the main reason why you might not go and vote at the next European elections (REQ)?
Interviewer show card, one answer only (INSTRI).
1. I am not interested in politics.
2. I am not interested in the European elections.
3. I am against Europe.
4. I am not well enough informed to vote.

5. Other reasons.
(ANSWER categories on card)

The added component can also be information regarding a definition, for example:

6.30 *Are you a member of a public library (REQ)?*
By "public library" we understand a library other than at a school or university (INFOD)
1. *Yes*
2. *No (ANSWER categories)*

The examples just mentioned demonstrated that survey items can be expanded by inserting an instruction for the respondent, an instruction for the interviewer, or information regarding a definition after the request for an answer before the answer categories.

The next examples will present extensions of survey items by inserting components such as an introduction, information regarding the content, an instruction for the interviewer, or a motivation for the researcher before the request for an answer. Typical examples are as follows:

6.31 *The next question concerns the upcoming elections (INTRO).*
Please, tell me, if there were elections tomorrow, would you go to vote (REQ)?
1. *Yes*
2. *No (ANSWER CATEGORIES)*

6.32 *People look for different things in their jobs. Some people like to earn a lot of money. Others prefer an interesting work (INFOC).*
Which of the following 5 items do you most prefer at your job (REQ)?
1. *Work that pays well*
2. *Work that gives a feeling of accomplishment*
3. *Work where you make most decisions yourself*
4. *Work where other people are nice to work with*
5. *Work that is steady with little chance of being laid off*

(ANWSER CATEGORIES)

6.33 *Interviewer: Ask this question also when the respondent did not answer the previous question (INSTRI).*
Is the monthly income of your household higher than $ 10,000 (REQ)?
1. *Yes*
2. *No (ANSWER CATEGORIES)*

6.34 *We need the information about income of your household to be able to analyze the survey results for different types of households. An income group is enough.*
It would help us a lot if you would be able to state what income group your household belongs to (MOTIV).

Please, tell me, to which one of the following income groups your household belongs after tax and other deductions (REQ).
1. *$ 20,000 – $50,000*
2. *$ 50,000 – $ 100,000*
3. *Higher than $ 100,000*

(ANSWER CATEGORIES)

One also can find more complex structures with three components by inserting more components before the request for an answer. Typical examples are as follows:

6.35 *The following request for an answer deals with your work (INTRO).*
Some people think that work is necessary to support themselves and their families (INFOC).
Do you like your work or do you do it as a necessity to earn money? (REQ + answer categories)
1. *Likes his work*
2. *Work is a necessity to earn money (ANSWER CATEGORIES)*

6.36 *Now I would like to talk about abortion (INTRO).*
"Abortion" means the deliberate termination of a pregnancy by a physician (INFOD).
Are there in your opinion situations that justify an abortion (REQ)?
1. *Yes*
2. *No (ANSWER CATEGORIES)*

In the same way an instruction for the respondent or a motivation for an answer can be used.

Other structures with four components are also possible. For example, starting with an introduction followed by a request for an answer, an extra instruction or information and were the answer categories are at the end. An example could be

6.37 *In the following requests for an answer we would like to ask you about your leisure activities (INTRO).*
Please, tell me which of the following activities you prefer most in your spare time? (REQ)
Indicate only one activity (INSTRR).
1. *Sports*
2. *Watching TV*
3. *Reading*
4. *Going shopping*
5. *Talking with people*
6. *Something else*

(ANSWER categories in request for an answer)

Finally, even more complex structures can be found in the literature such as inserting several components before and after the request for an answer, for example:

6.38 *We want to ask you about your education (INTRO).*
It is very important for us to get a good picture of the education of our citizens (MOTIV).
By "education" we understand the schools you finished for a degree and we want to know the highest education you finished with a degree (INFOD).
What was the highest level of educational training that you finished with a degree (REQ)?

1. *Primary school*
2. *Lower vocational training*
3. *High school*
4. *Higher vocational training*
5. *University*
6. *Other (ANSWER CATEGORIES)*

Here follows another example:

6.39 *Now we would like to ask you about issues that are frequently discussed in the media (INTRO).*
When a physician assists a patient at his/her own request to die, we call this euthanasia (INFOD).
Some people and political parties think that euthanasia should be forbidden. Others are of the opinion that a physician always must comply with a patient's request to die. However, there are also people whose opinion lies in between (INFOC).
What is your opinion about euthanasia (REQ)?
Interviewer present the card (INSTRI).
People who favor euthanasia should choose the number 7, which means that "a physician has to comply with a patient's request."
People who are against euthanasia should choose the number 1, which means that "euthanasia should be forbidden. People who are neither for or against it should choose the number 4 (INSTRR)

Where would you place yourself on the scale (REQ)?

1		4		7
Euthanasia should be forbidden (ANSWER scale on card)		*Neither for nor against it*		*A physician has to comply with a patient's whish*

This type of request is not used often, because the text becomes very complex and it is unclear whether respondents can answer these questions at all.

6.2.3 The frequency of occurrence

After having introduced various structures of closed survey items on the basis of selected data, we can now investigate the frequency of occurrence of the

different structures of survey items, as in Table 6.1.

The frequency of occurrence of survey items relating to subjective requests for an answer is studied on the basis of Molenaars sample. Table 6.1 summarizes the structures of closed survey items that we encountered in this data set. This table shows clearly that structures where answer categories are embedded in the request for an answer are more frequent than structures without embedding them. This is because most interviews were still interviewer-administered in The Netherlands at that time. With the increase of the number of self-administered interviews like mail and WEB surveys this distribution might change quite rapidly.

The table also shows that researchers avoid highly complex survey items. Although complex items are possible as we have shown, they are seldom used in market and opinion research. Most frequently the items consist of two components. An inspection of the English and German survey items we had collected also confirmed that the structures mentioned in Table 6.1 were similar. Survey items consisting of more than three components were infrequent. Also, the most common extension of a basic structure of a survey item is to start with some information about the content of the following survey item. Another possibility is the use of an introduction. These two structures may be substitutes of one other as they seldom occur simultaneously.

Table 6.1: Overview of structures of closed survey items encountered in the sample of requests for answers for subjective variables with closed response categories

Structure of closed survey items relating to subjective requests	Number of components	Frequency	
		%	Absolute
REQ + answer categories	2	8	14
REQ +embedded answer categories	2	39	73
REQ + embedded answer categories + answer categories	2	3	5
INFOC + REQ + embedded answer categories	3	30	53
INFOC + REQ + answer categories	3	5	9
INTRO + REQ + answer categories	3	1	1
INTRO + REQ + embedded answer categories	3	2	3
INTRO + REQ + embedded answer categories + answer categories	3	10	19
INTRO + REQ + REQ + answer categories	4	1	2
INTRO +INFOD + REQ + embedded answer categories	4	1	2
Total		100	181

6.2.4 The complexity of survey items

From the respondents' perspective, survey items consisting of various components are more difficult to understand than requests with only one component.

In the literature (Graesser et al. 2000b; Tourangeau et al. 2000; Molenaar 1986) different measures for complexity are used that coincide partially with the ones we make use of. In our research (see Chapter 12) we register which components are present in a survey item. In addition, we determine the complexity of the introduction and the request separately. For both parts the complexity is studied with indices that will be discussed in more detail.

One of these indicators for complexity is the *number of interrogative sentences*. If there is more than one interrogative sentence in a request, the respondent has to decide which one should be answered which in turn complicates the comprehension of the whole request.

Another characteristic that increases the difficulty of comprehension relates to the *number of subordinated clauses*. If a component contains one or more subordinate clauses that are embedded in the main clause, it can be assumed that the respondent needs several mental operations before fully understanding the sentence. As an example, we mention the following three requests:

6.40 *Do you think, although there is no certainty, that your financial situation will improve in the future?*

6.41 *Do you think that your financial situation will improve in the future?*

6.42 *Will, in your opinion, your financial situation improve in the future?*

Example 6.40 contains two subordinate clauses, where the second contains the proper request. Example 6.41 consists of only one subordinate clause containing a request.

The third example (6.42) has no subordinate clauses, and the request is stated in the main clause. This example is the easiest to comprehend.

The number of words of a component also contributes to its complexity. The more words it contains, the more difficult it is to understand. This also can be studied by means of *the average number of words for each sentence.*

Still another characteristic that adds to the complexity of a sentence is the mean *number of syllables* in the words. It is assumed that the more syllables a sentence contains, the more difficult it is to understand.

The last characteristic relating to complexity is *the number of abstract nouns on the total number of nouns. Abstract nouns* indicate objects that in principle can not be touched, which means that they do not refer to living beings or physical objects, while concrete nouns refer to the latter categories. We assume that the comprehension becomes more difficult with the increase of abstract nouns in comparison to the number of concrete nouns.

This overview of complexity characteristics suggests reducing complexity by using only one interrogative sentence in the request, few subordinate clauses, and short sentences with a minimal amount of abstract nouns. Most survey researchers agree with these recommendations (see literature).

6.3 WHAT FORM OF SURVEY ITEM SHOULD BE RECOMMENDED?

Our knowledge about the optimal form of survey items is still rather limited. Some new results on this topic will be mentioned in Part III of this book. However, about the use of some components research has already been done. First of all, Belson (1981) studied different forms of media items. For example, after respondents answered the question *"How often did you watch TV during the last week?"* he asked them how they had interpreted the terms "watch TV," "you," "last week." He wanted to see how many people interpreted the question according to the wishes of the researcher. It turned out that many people interpreted the terms differently as expected. For example, "watch TV" for some people meant that they were in the room while the TV was on. "You" could mean the respondent or his/her family. "Last week" was by some people interpreted as in the evening only, ignoring daytime viewing. Also weekend viewing was occasionally ignored.

In order to improve the question, Belson (1981) tried to include definitions of what was meant. This led to the following formulation:

6.43 *How often did you watch TV during the last week?*
By TV watching we mean that you, yourself, are really watching the TV while we would like to ask you to include day viewing and weekend viewing.

Somewhat surprising was that the number of misinterpretations of the request for an answer in the new form was not much lower than for the former. Belson's explanation was that the question was too long and people had already made up their minds before the definitions were given.

This does not mean that the length of the survey item always has negative effects on the quality of the responses. Schuman and Presser (1981) and Sudman and Bradburn (1983) suggest that the length of the survey item can have a positive effect if the topic is announced early and no further substantial information is given. For example, the question

6.44 *Should abortion be legalized?*
0. No
1. Yes

can be extended without adding new information as follows:

6.45 *The next question concerns the legalization of abortion. People have different opinions about this issue. Therefore, we would like to know your personal opinion.*

Could you tell me what your opinion is, should abortion be legalized?
0. No
1. Yes

The important difference between the longer form of Belson and the last example is that no relevant information is added after the request for an answer has been made clear. In the last form the respondents have more time to think and thereby improve their answers.

The other side of the coin is that one has to give extra information if the object of the question is not known to many respondents. For example, the meaning of the terms: euthanasia, democracy, globalization, the WTO, and so on may be unknown to large portions of the population. In that case an explanation of the term is necessary if they are to be used in a survey.

The findings of Belson suggest that these definitions definitely should not be given after the question, but before the request. We suggest starting a request with a definition of the concept. For example, if we want to know the opinion about a policy of the WTO with respect to free trade, we could use the following survey item:

6.46 In order to regulate world trade, an organization of the UN, called WTO, develops rules for the world trade to reduce the protection of countries of their own products and therefore to promote free trade in the whole world. What do you think of these activities of the UN? Express your opinion in a number between 0 and 10 where 0 = completely against, 5 = neutral and 10 = completely in favor.

In this case a definition of the concept of interest is needed, and it should be given in relative simple words *before* the question is asked. This is to ensure that the respondents listen to the explanation before they decide their response.

6.4 SUMMARY AND CONCLUSIONS

In this chapter we introduced different components of survey items and described the combinations of components that occur in open-ended and closed survey items. Given the data in Table 6.1, we can conclude that closed survey items consist of two (50%) or three (49%) components. The most often encountered structure of a closed survey item with two components consisted of a request for an answer with an answer component. Since some requests require an introduction or more information or an instruction these components were sometimes added (introductions 15%; information regarding the content 35%). However, there is always a tradeoff between precision and complexity. We have mentioned the following aspects of a survey item which increase its complexity:

- The number of components in the survey item
- The presence of more than one interrogative sentence
- The number of subordinated clauses
- Te number of words in a sentence
- The mean number of words in a sentence
- The mean number of syllables per word
- The ratio of abstract and concrete nouns

Although it is in general suggested that complex sentences should be avoided, there is also research suggesting that increasing the length of the questions improves the quality of the answers. Some new results with respect to the effects of these features of survey items on the their quality will be discussed in Part III when we present the effects of these above mentioned choices on data quality.

Exercises

Below are several survey items from empirical research.

Decompose the different survey items into their components.

1. *Before we proceed to the main topic of the questionnaire, we would like to ask the following question:*
   How long have you lived in your present neighborhood?
   Number of years _____
   Don't know _____
   Not answered _____
   ENTER YEAR ROUNDED TO NEAREST YEAR.
   PROBE FOR BEST ESTIMATE
   IF LESS THAN 1 YEAR, CODE 0

2. *How far would you say you discuss politics and current affairs?*
   SHOW CARD
   1. *A few times a week* ☐
   2. *A few times a month* ☐
   3. *A few times a year* ☐
   4. *Never or almost never* ☐
   8. *Don't know* ☐

3. *Do you actively provide any support for ill people, elderly neighbors, acquaintances or other people without doing it through an organization or club?*
   REGISTER ONLY UNPAID, VOLUNTARY ACTIVITY. INCLUDE ANY FINANCIAL SUPPORT GIVEN BY THE RESPONDENT TO ILL, ELDERLY, ETC.
   1. *Weekly* ☐
   2. *Monthly* ☐
   3. *Yearly* ☐
   4. *Never, or almost never* ☐

4. *We all know that no political system is perfect but some may be better than others. Therefore we would like to ask you the following about the functioning of democracy in our country. How satisfied are you with the way democracy functions in our country?*
   1. *Very dissatisfied*
   2. *Quite dissatisfied*
   3. *Neither satisfied nor dissatisfied*
   4. *Quite satisfied*
   5. *Very satisfied*
5. In your questionnaire, did you also use components other than requests for answers and answer categories? If so, check whether they are in line with common practice or whether you did something unusual. According to your judgment, is that good or bad?

CHAPTER 7

Survey items in batteries

In the last chapter we discussed the forms that single survey items can take. However, in Chapter 4 we mentioned that researchers in the social sciences often bring items together in batteries. In that case the different survey items do not stand alone anymore but often are connected by one introduction, instruction, and one request for an answer and a set of answer categories. Since we treat each text unit that requires one response as a survey item, we have to give special attention to the definition of survey items of batteries. The problem is that the different survey items in a battery contain very different text components even though they are often assumed to be equal and treated the same.

What distinguishes batteries is the mode of data collection in which they have been placed. Therefore, we start this chapter with batteries that are used in oral interviews, followed by a discussion about batteries in mail surveys and finally batteries employed in computer-assisted-self-interview (CASI) are discussed. In each case we will discuss which components should be seen as belonging to each survey item. In the summary and discussion we also will give some recommendations.

We will discuss the different battery types, not because we think that batteries are a good tool for survey research, but because they are so popular. As we have indicated in Chapter 4, we think that the standard batteries of agree/disagree responses have done more harm than good for the social sciences. Having given our words of caution and advice, let us start with battery forms employed in oral interviews.

7.1 Batteries in oral interviews

Typical for batteries in oral interviews is that the interviewer reads the items for the respondent. Within this battery class there is a difference between the face-to-face interview with *show cards* containing information and the telephone interview, where show cards cannot be used. Let us start with an example of a battery without show cards. A typical example of an oral battery without show cards has been presented as oral battery 1.

In this example, information about the content is read first, and then a

request for an answer with implied "yes" or "no" answers is read. Next the interviewer has to read the first item, wait for the answer, present the next item, and so on. Hence, the introduction and the request are read only before the first survey item and then not anymore. As a consequence we assume that each survey item after the first one consists only of the statement (or stimulus) and the response categories.

Given the interview process we have suggested above, the information about the content and the request for an answer belong to the first item, while all other items consist only of a stimulus since the interviewer does not repeat the answer categories for each item. We think that this is formally correct even though it may not be in agreement with the intention of the original designer of the battery. Moreover, it may be that the introduction and the question retain a strong presence in the mind of the respondent when they are not repeated for each item.

Oral battery 1

There are different ways of attempting to bring about improvements or counteract deterioration within society.

During the last 12 months, have you done any of the following ?

First ..... READ OUT

	Yes	No
A. Contacted a politician		
B. Contacted an association or organization		
C. Contacted a national, regional, or local civil servant		
D. Worked for a political party		
E. Worked for a (political) action group		
F. Worked for another organization or association		
G. Worn or displayed a campaign badge/sticker		
H. Signed a petition		
I. Taken part in a public demonstration		
J. Taken part in a strike		
K. Boycotted certain products		

This kind of battery can be used only for very simple response categories as in this example. For more complex response categories the quality of the responses will improve if the respondent is provided with visual aids. Visual aids help the respondent in two ways: (1) to provide the response alternatives on a card so that they can answer each item consistently and (2) to provide the respondent with the statements. The latter method makes sense if the state-

ments are complex or for emphasis. We will give an example of both types from research practice.

In the following example presented in oral battery 2, the respondents are provided with information about the response alternatives on card D1. First we present the form provided to the interviewer and then card D1.

In this case the introduction and the question belong to the first item, and the next items all contain a stimulus and response categories because the respondents have these answer categories always in front of them.

Oral battery 2

CARD D1: Policies are decided at various different levels. Using this card, at which level do you think policies should be decided mainly about ...

READ OUT AND CODE ONE ON EACH LINE

	International level	European level	National level	Regional or local level	(DK)
D1 ... protecting the environment	1	2	3	4	
D2 ... fighting against organized crime	1	2	3	4	
D3 ... agriculture	1	2	3	4	
D4 ... defense	1	2	3	4	
D5 ... social welfare	1	2	3	4	
D6 ... aid to developing countries	1	2	3	4	
D7 ... immigration and refugees	1	2	3	4	
D8 ... interest rates	1	2	3	4	

Card D1

Card D 1
International level
European level
National level
Regional or local level

If the answer categories are simple but the statements are complex or important, the content of the card can be changed. The next example, oral battery 3, demonstrates this point. First we present the card for the respondents and after that the form provided to the interviewer.

In this case the response alternatives are rather simple but the researcher wants the respondents to carefully consider the different possible situations and therefore the show card presents the different conditions that have to be evaluated. All the information before the first item belongs to the first item, while the second to the last item contains the statement because the response alternatives are not repeated.

CARD for oral battery 3

The woman got pregnant because she was raped.

The woman got pregnant even though she used a contraceptive pill.

.

.

The woman got pregnant although there are already enough children in the family.

Oral battery 3

An issue often discussed nowadays is abortion. By abortion we understand the purposeful termination of a pregnancy .

On this card some circumstances are indicated under which an abortion might be carried out.
Could you tell me for each circumstance mentioned on the card whether you think that an abortion is permissible ?

READ OUT

	Permissible	Not permissible
The woman got pregnant because she was raped.		
The woman got pregnant even though she used a contraceptive pill.		
.		
.		
The woman got pregnant although there are already enough children in the family.		

It is even possible that both the stimuli and the answer categories are provided on a show card. An example is given (oral battery 4). The first form is the interviewer version of the battery. Card K2 contains the same information for the respondents. In this case the show card (K2) contains the stimuli as well as the response alternatives. For item A the introduction to the question, the request for an answer and the answer categories belong to the survey item. Items B–G consist of the stimulus and the response categories.

Oral battery 4

Looking at card K2, how important is each of these things in your life?

First ... READ OUT

	Not important at all										Very important	(Don't know)
A ...family?	00	01	02	03	04	05	06	07	08	09	10	88
B ...friends?	00	01	02	03	04	05	06	07	08	09	10	88
C ...leisure time?	00	01	02	03	04	05	06	07	08	09	10	88
D ...politics?	00	01	02	03	04	05	06	07	08	09	10	88
E ...work?	00	01	02	03	04	05	06	07	08	09	10	88
F ...religion?	00	01	02	03	04	05	06	07	08	09	10	88
G ...voluntary organizations	00	01	02	03	04	05	06	07	08	09	10	88

Card K2

	Not important at all										Very important
A ...family?	00	01	02	03	04	05	06	07	08	09	10
B ...friends?	00	01	02	03	04	05	06	07	08	09	10
C ...leisure time?	00	01	02	03	04	05	06	07	08	09	10
D ...politics?	00	01	02	03	04	05	06	07	08	09	10
E ...work?	00	01	02	03	04	05	06	07	08	09	10
F ...religion?	00	01	02	03	04	05	06	07	08	09	10
G ...voluntary organizations?	00	01	02	03	04	05	06	07	08	09	10

7.2 BATTERIES IN MAIL SURVEYS

Batteries are also used in mail surveys. We provide common structures of this type in the examples below starting with mail battery 1.

The difference between the mail battery and the oral battery is that the respondents have to do all the work themselves. They have to read the question, the first survey item, and the answer categories. Then they have to fill in an answer and read the next statement and look for a proper answer again. Hence, the question is read only before the first survey item and then not again. As a consequence we assume that each survey item after the first one consists of the statement (or stimulus) and the response categories.

Mail battery 1

For each statement, could you tell me, which answer seems to fit your situation ?					
	Agree strongly 1	Agree moderately 2	In the middle 3	Disagree moderately 4	Disagree strongly 5
My financial situation has improved over the past year.					
My carrier prospects have improved the past year.					
.					
.					
.					
My relational problems has worsened over the past year.					

A slightly more complex battery is presented in mail battery 2. The battery begins with an introduction or an information regarding the content, after which the request for an answer with answer categories is given, followed by the statements.

Mail battery 2

Here are some statements about our society. Do you agree or disagree with the following statements?		
	Agree	Disagree
People in our society still have great ideals.		
The government should improve gun control.		
.		
Most of our citizens are interested only in making money.		

In this case an introduction and the question are given before the first survey item. As we have suggested above, we assume that these two belong to the first item while the next item consists of only a statement and answer categories.

In the next example (mail battery 3) the complexity of the battery is increased by adding other components.

This battery starts with information about the content, then a request for an answer with embedded answers follows. Next a request for an answer without answer categories is provided. However this is not the real request for an

answer because the answer categories don't match with the answer categories suggested later. Furthermore, the next sentence is also a request for an answer. The former request should be seen as an introduction, and the latter question is the real request for an answer. The first item and the answer categories that follow are presented in a table. As we mentioned previously, we assume that the first item contains all the information including the item, while the second and subsequent items consist only of the statement plus their answer categories.

Mail battery 3

There are different ways people can express their disapproval about the measures of authorities.

I will mention some to you, and then indicate then whether you approve of them or not. How much do you approve or disapprove of this action ?

	Approve completely	Approve	Disapprove	Disapprove completely
At a school some teachers are in danger of loosing their jobs. They therefore organize a strike without the approval of the trade union.				
Suppose that people are against a new law and therefore they occupy the Parliament in order to hamper the work of the representatives .				
.				
.				
.				
Suppose that the government had decided to increase the number of pupils in the classes of elementary school. Some teachers don't accept this and threaten to go on strike.				

A thoroughly developed and tested mail battery is the *choice questionnaire*. This procedure has been developed to collect a *well-considered opinion* of the respondent (Saris et al. 1984; Neijens 1987; Bütchi 1997). The reason for this development was that it was realized that people may understand a question, like the one about the policy of the WTO, after an explanation is given. However, this does not mean that they have a well-informed opinion about it (Converse 1964). Saris et al. (1984) suggested the use of a procedure called the *choice questionnaire* to collect a well-considered public opinion. Typical for the choice questionnaire is that respondents are provided with arguments on

paper both in favor and against the choice they have to make. The procedure is rather elaborate and cannot be shown in detail here, for more information we refer the reader to Neijens (1987). The problem mentioned in Chapter 6, about the free trade policy of the WTO, was developed in line with the choice questionnaire approach as presented in Table 7.1.

Table 7.1: An illustration of a simple choice questionnaire

The possible consequences of the reduction of the protection of national products and the promotion of free trade in the whole world are presented below.

Please, evaluate the following consequences of this policy by first determining whether a consequence is an advantage or a disadvantage and consequently evaluate the size of the advantage or disadvantage with a number, where a neither large nor small advantage or disadvantage is 400.

How large is the advantage or disadvantage of the following:	Advantage neither large nor small=400	Disadvantage neither large nor small=400
The bankruptcy of some local companies in some underdeveloped countries		
The investments of international companies		
More efficiency in the companies so that they can compete internationally, etc.		

In order to regulate world trade, an organization of the UN, called WTO, develops rules for world trade.
To reduce the protection of countries of their own products and to promote free trade in the whole world.
Are you, in favor or against free trade in the world?
0 against 1 in favor

The table shows that the respondents are provided with information about the choice they have to make. This information is provided in the form of questions concerning evaluations of consequences for the possible options. In this example, only one option has been discussed; however, the option that no free trade policy is introduced could be treated in the same way. People are asked to give numeric evaluations of the size of the possible advantages and disadvantages because in this way it is possible to get a total of the advantages and disadvantages for each option and to make a choice on the basis of these total evaluations of the options.

Saris et al. (1984) and Neijens (1987) have demonstrated that with this approach the final choice of the respondents was consistent with their judgments for approximately 70% of the cases. On the other hand, the consistency was around 30% if the judgments were asked after the choice was made without

the information provided by this approach. We conclude that the information aids in creating a well-informed choice. The choice questionnaire is discussed here because it is a very elaborate procedure to provide the respondents with information before they are asked to make their choice. This approach differs from the previously discussed batteries of survey questions because all battery items were prepared for the final choice. However, most of the time the items are supposed to measure different aspects of an object and are not aimed at preparation for a later judgment.

The final result of choice questionnaires can be very different from that of the naive opinion of the respondent without the given information. Therefore, the choice questionnaire should be used only to measure well-informed opinion and not to measure naive opinions that are based mainly on the first ideas that come to the mind of the respondent (Zaller 1992).

7.3 BATTERIES IN CASI

In the early development of computer assisted data collection, the computer-assisted-self-interviewing (CASI) mode often contained a series of identical requests for an answer, and answer categories for a series of stimuli or statements. Typical for such series of survey items is that the formulation is exactly the same for each item and that only one introduction with other possible components is given before the first survey item is mentioned. The items are treated equally because the interview programs use substitution procedures. An example of such an instruction to an interview program could look as follows:

#Casibattery 10 1
 # item 1
healthcare
#item 2
social services
#item 3
.
.
 # item 10
social security
#
#Question with 5 answer categories
What is your opinion about our "S"?
1. *Very satisfactory*
2. *Satisfactory*
3. *Neither satisfactory nor unsatisfactory*
4. *Unsatisfactory*
5. *Very unsatisfactory.*

The first line indicates that this battery consists of 10 stimuli and only 1 question; then follow the 10 stimuli and after that the request follows with the answer categories. In the request "S" is mentioned, which is substituted by the different stimuli or statements in the presentation of the questions to the respondents. Using this interview program the following computer screens will be presented:

Screen 1 of CASI battery 1

What is your opinion about our health care?

1. Very satisfactory
2. Satisfactory
3. Neither satisfactory nor unsatisfactory
4. Unsatisfactory
5. Very unsatisfactory.

Screen 2 of CASI battery 1

What is your opinion about our social services?

1. Very satisfactory
2. Satisfactory
3. Neither satisfactory nor unsatisfactory
4. Unsatisfactory
5. Very unsatisfactory.

Screen 10 of CASI battery 1

What is your opinion about our social security?

1. Very satisfactory
2. Satisfactory
3. Neither satisfactory nor unsatisfactory
4. Unsatisfactory
5. Very unsatisfactory.

In contrast to the previous batteries, all survey items contain exactly the same information and therefore have the same complexity.

This kind of battery has been used not only for stimuli but also for statements as the next example shows. The screens of CASI battery 2 look as follows:

Screen 1 of CASI battery 2

Which of the following statements is true or false?

Screen 2 of CASI battery 2

The European Monetary Union will be governed by a government composed of all participant nations.

Is this true or false ?

1. True 2. False

Screen 3 of CASI battery 2

The language of the European government will be French.

Is this true or false ?

1. True 2. False

There is one difference with battery 1, namely, that before the first item an introduction (in question form) is presented (screen 1 of CASI battery 2). Although the introduction is aimed at all items, it will be read only once before the first items. Therefore, we have decided that in this case only the first item has an introduction and the other items have no introduction.

Both examples are very simple. Far more complex examples can be found in the research literature. In CASI battery 3 we provide an example of a rather complex case.

Screen 1 of CASI battery 3

We would like to ask you how serious you find some illnesses.

You can indicate the seriousness of the illnesses with a number between 0 and 100, where 100 means very serious and 0 means not at all serious.
Thus, the more serious the illness, the larger the number.

Now comes the illness:

Screen 2 of CASI battery 3

Cancer

How serious do you consider this illness?
0=not at all serious; 100=very serious

Which number indicates the seriousness of this illness?

Screen *n* of CASI battery 3

Aids

How serious do you consider this illness?
0=not at all serious, 100=very serious

Which number indicates the seriousness of this illness?

This example is more complex because an introduction is presented, then an instruction including a definition, followed by another instruction, and finally a second introduction. The first item comes after all this information and all items are treated equally. But the first item is very different because of the large amount of information provided before it. It is doubtful how much of the information provided will be available in the mind of the respondent when the second and following items are answered. It is for this reason that we suggest connecting all information to the first item and not to the other items. This would lead to the second and the following survey items consisting of information that is placed on the screen for that item such as a stimulus, a request with answer categories and a second rephrasing of the basic question. Repetition of the main request for an answer after each stimulus is more typical for computerized questionnaires than in other modes of data collection.

In the later days of computer-assisted data collection and in the present Web surveys, many survey items are presented on one computer screen. An example is given in CASI battery 4. It is the measurement of "political action" in Web survey format, which was also presented in oral battery 1. We think that the respondents will read the introduction, the instruction, and the request first. Then they will proceed to read the first statement and click on the box. Next they will read the subsequent statement and decide whether to select the box. Having finished one, they will proceed to the next statement and complete it in the same manner, until whole list is completed. Here we assume that the first item contains much more information than the second to the last item, which consists only of the statement itself. A new element employed in this battery form is that the respondents can click on the boxes to indicate their answers.

Screen of CASI battery 4

There are different ways of trying to improve things in this country or help prevent things from going wrong. During the last 12 months, have you done any of the following? Tick all that apply:

Contacted a politician, government or local government official	☐
Worked in a political party or action group	☐
Worked in another organisation or association	☐
Worn or displayed a campaign badge/sticker	☐
Signed a petition	☐
Taken part in a lawful public demonstration	☐
Boycotted certain products	☐
Donated money to a political organization or group	☐
Participated in illegal protest activities	☐

CASI battery 5 illustrates that an 11-point scale can be used in this manner with many items being displayed on the same computer screen. The information belonging to the different items can also be determined in the same way as was described earlier.

Screen CASI battery 5

Please indicate how much you personally trust each of the institutions below. Selecting the leftmost button means you do not trust an institution at all, and the rightmost button means you have complete trust.

	No trust at all									Complete trust	(Don't know)
The Parliamant	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
The legal system	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
The police	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
Politicians	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
Political partners	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
The European Parliament	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍
The United Nations	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

7.4 SUMMARY AND DISCUSSIONS

In this chapter we have presented the most common ways in which batteries are used in survey research. One issue we have emphasized here and that has not been discussed in the literature thus far is the difference between the batteries as they are operationalized in their respective modes of data collection. We have also shown that this has distinct implications for the description of the survey items in the battery. We suggested including those components in the different survey items that are explicitly provided to the respondent when the survey item is introduced. That means that the first item in the battery has

a form very different than the other items in the battery, because the information given before the first item belongs to the first item and not to the others. This is particularly relevant because it affects the estimation of complexity (as discussed in the codebook) of the survey items within a battery, especially between the first and the other items, and across batteries between items in different modes of data collection. There are no studies showing that the use of stimuli has negative effects on the quality of the data collected. Batteries with statements have been more frequently criticized. These issues were covered in Chapter 4 and will not be repeated here. We will only note again that there is sufficient empirical evidence pointing in the direction that in the majority of cases trait-specific questions are of better quality than are battery items.

The battery method is an obvious choice asking for a reaction to many different statements. Often information has to be added to the battery about concepts or the procedure of response; however, there are limits to these possibilities. These limits depend on the topic discussed and the responses that are asked, which points to a difference between batteries with stimuli and batteries with statements.

Although we have discussed batteries with stimuli and statements together, they have their differences. On the basis of a random sample of survey items mentioned in the last chapter (Molenaar 1986), we have found that both types of batteries have in more than 90% of the cases an introduction before the first item is presented. But it was also found that they differ in that batteries with statements need extra instructions (78% of the batteries), versus those with stimuli (less than 2% of the cases). This may be a consequence of the length of the statements used, since otherwise there is no difference between these two types of batteries.

Let us give a final example (complex battery with 2 requests) to emphasize that there are limits to the use of batteries and that these limits vary for the different modes of data collection.

In this example two questions are presented in one table. Instructions are given to the interviewer above the table about how the questioning should be conducted.

Complex battery with 2 requests

CARD K1
For each of the voluntary organizations I will now mention, please use this card to tell me whether any of these things apply to you now or in the last 12 months, and, if so, which.

READ OUT EACH ORGANIZATION IN TURN. PROMPT "Which others?"
ASK K2 FOR EACH ORGANIZATION CODED 1–4 AT K1.

K2 Do you have personal friends within this organization?
READ OUT EACH ORGANIZATION CODED 1–4

	K1 CODE ALL THAT APPLY FOR EACH ORGANIZATION					K2	
	None	Member	Participated	Donated money	Voluntary work	Personal friends? Yes	No
A. Firstly, a sports club or club for outdoor activities; do any of the descriptions on the card apply to you?	0	1	2	3	4	1	2
B. An organization or cultural or hobby activities?	0	1	2	3	4	1	2
C. A trade union?	0	1	2	3	4	1	2
D. A business, professional, or farmers' organization?	0	1	2	3	4	1	2

NOW ASK K2 ABOVE FOR EACH ORGANIZATION CODED 1–4

In this case a show card is provided with the response alternatives to the first question:

Show card for the complex battery:

- A member of such an organization
- Participated in an activity arranged by such an organization
- Donated money to such an organization
- Have done voluntary (unpaid) work for such an organization

It should be clear that this combination of two batteries is rather complex. This format is recommended only in the case where the interviewers have been trained well and it is not for a mail questionnaire or a computer-assisted-self-interview.

With this illustration we finish the discussion about batteries of survey items; much further research is needed in order to determine what the effect of the different forms on the quality of the responses are. In Part III results of such research will be presented.

EXERCISES

1. Two interviewer forms of batteries are presented below.

Example 1:

How important is each of the following in your life? Here I have a card with a scale of 0–10 where 10 means "very important" and 0 means "not important at all." Where would you place yourself on this scale? SHOW CARD.

	Not important at all										Very important	Don't know
A. Family and friends	0	1	2	3	4	5	6	7	8	9	10	88
B. Leisure time	0	1	2	3	4	5	6	7	8	9	10	88
C. Politics	0	1	2	3	4	5	6	7	8	9	10	88
D. Work	0	1	2	3	4	5	6	7	8	9	10	88
E. Clubs	0	1	2	3	4	5	6	7	8	9	10	88
F. Community organization	0	1	2	3	4	5	6	7	8	9	10	88

Example 2:

As you know, there are different opinions as to what it takes to be a good citizen. I would like to ask you to examine the characteristics listed on the card. Looking at what you personally think, how important is it:

SHOW CARD

	Not at all important									Very important	Don't know
	0	1	2	3	4	5	6	7	8	9	10	88
A. To show solidarity with people who are worse off than yourself												
B. To vote in public elections												
C. Never to try to evade taxes												
D. To form your own opinion, independently of others												
E. Always to obey laws and regulations												
F. To be active in organizations												
G. To think of others more than yourself												

Answer the following questions:

a. What would you put on the card for these two examples?
b. Given the choice in 1a., indicate for both batteries what text belongs to which survey item.
c. What kind of components are presented before the first item in each scale?

2. Is there a reason to use batteries in your own questionnaire?
   a. If so, why?
   b. What is your proposal?

This Page Intentionally Left Blank

CHAPTER 8

Mode of data collection and other choices

The first part of this chapter is dedicated to discussion of some choices over which the researcher has little control, but which can have considerable influence over the survey results. The first choice to be discussed is the mode of data collection. In principle the researcher is free to choose any mode, however, in reality the options are restricted by the budget that is available for the study. Anyway, the mode of data collection affects the quality of the survey and also its costs. We will spend considerable attention to the mode of data collection since it is a very important decision and because its possibilities increase rapidly.

The second choice we will discuss concerns the position of the question in the questionnaire. Not all questions can be placed in the best place of the questionnaire. Therefore, a policy should be developed that deals with how to place the questions in the questionnaire.

A third issue to discuss is the layout of the questionnaires. Unfortunately there is still very little known about the effects of layout; however, we will give references where relevant information about this issue can be found.

Finally there is the choice of the language for the questionnaire. This is, of course, not a real choice. Nevertheless, a limited choice exists with respect to the language related to minority groups in a country. For example, should the Moroccan people in France be interviewed in French or in Moroccan Arabic? It is also important for comparative research to know whether it makes a difference on the substantive conclusions if the questions are asked in a given language.

8.1 The choice of the mode of data collection

Data collection is developing rapidly. In the 1960s and 1970s there were only three procedures for data collection: paper-and-pencil interviewing (PAPI) by an interviewer in the home of the respondent; traditional telephone interviewing, where the interview was done by telephone; and, finally, mail questionnaires, which were done without the presence of an interviewer and where respondents had to fill in the forms themselves.

In the mid-1980s the computer made its entry into survey research. First, the telephone interview was computerized. The computer-assisted telephone interview (CATI) is now a very common form of data collection for survey research. Also in the beginning of the 1990s the first experiments with computer-assisted personal interviewing (CAPI), as a substitute for PAPI, were done even though at that time the portable computer was not yet available (Danielson and Maarstad 1982; Saris et al. 1981). In 1985 the first experiments with computer-assisted self-administrated interviews (CASI) as a substitute for mail questionnaires (Saris and de Pijper 1986; Kiesler and Sproull 1986) were conducted.

The first two forms of computer-assisted interviewing systems did not change much for the respondents; they caused mainly a change for the interviewer. The last form really made an impact on the respondents because they had to answer questions on a computer and not on paper. There were two forms to this approach. The form that resembled mail questionnaires was the disk-by-mail (DBM) approach. In this case a diskette with the interview and interview program were sent to the respondents, who would answer the questions on their own computers and send the disks back to the research agency. The second approach was the *telepanel* (Saris and De Pijper 1986; Saris 1991, 1998). In this approach a random sample of the population was provided with computers and modems (if necessary). With the equipment interviews could be sent to the households via the telephone and the answers could be returned without intervention of an interviewer. These two approaches required the respondents to have a computer. Using the DBM system one could only study populations with computers like businesses or physicians, etc. The telepanel system required a large investment in computers in order to allow for studies of representative samples of the population. The telepanel approach also required the use of the same households for many studies because of the large startup investment. As a consequence, this research design became automatically a panel survey design.

The next developmental phase was the further automatization of the telephone interview. Two new forms have been developed for large scale surveys: touchtone data entry (TDE) and voice recognition entry (VRE). In both cases the questions are presented to the respondent by a device that can present a question in natural language via a recorder or a computer. The respondent is asked to choose an answer by pressing keys on the handset of the telephone (TDE) or to mention the answer loudly; these answers are interpreted by a computer and coded (VRE). These two approaches were developed for very large surveys in the United States (Phipps and Tupek 1991; Harrel and Clayton 1991).

A new phase in the data collection has been the development of the World Wide Web with its possibilities to reach very many people with relatively low costs. This approach can become the alternative for mail questionnaires, DBM, and the telepanel if the Web facilities are sufficiently widespread to allow research with representative samples of the population. As long as this is not the case the telepanel procedure should be used by providing a representative

sample of respondents with the necessary equipment to participate in research. This is done by Knowledge Net (Couper 2000) in Palo Alto (United States) and by Centerdata in Tilburg (The Netherlands). The development of the Web survey is not a fundamental change compared with DBM or the telepanel, but is an improvement in efficiency and in cost reduction.

The most recent development is that experiments are done with audio self-administered questionnaires (ASAQ) and audio computer-assisted self-interviewing (ACASI). The purpose of these approaches is to make it possible for illiterate people to fill in self-administered questionnaires because the questions are read to them via a recorder (ASAQ) or a computer (ACASI). On the other hand, it also provides the respondents with an environment where they can answer questions without being influenced by people in their surrounding (an interviewer or people in the household). This is especially important for sensitive questions where socially desirable answers can be expected. In order to provide such an environment, the questions are presented through a headphone and the respondents can mark their answers on a response form (ASAQ) or by pressing keys on a computer (ACASI). This approach leads to results for sensitive issues that are very different from results obtained by the standard data collection methods (Jobe et al. 1997; Tourangeau and Smith 1998; Turner et al. 1998). A more detailed overview of the historical development of data collection can be found in Couper et al.(1998)[1].

8.1.1 Relevant characteristics of the different modes

More important than the historical sequence of events are the differences in characteristics of the different modes of data collection for

1. The presence of an interviewer
2. The mode of presentation (oral or visual)
3. The role of the computer

Tourangeau et al. (2000) make a fourth distinction between oral responses: written and keyed responses. Evidently more cognitive skills are required for writing than for keying answers and oral responses, although this has not led to different data collection methods. Most methods use oral responses if an interviewer is present but keyed or written answers if no interviewer is available. Therefore this distinction is completely confounded with the role of the interviewer, although this is not absolutely necessary.

1 Our overview deviates on only one point of the report of Couper et al. (1998), which is the early development of CAPI. Although they were informed that experiments with CAPI were done as early as 1980, they did not mention this in their overview. They thought that it was not possible at that time because the PC was not yet available. But these experiments were not done with PCs but with the first Apple computer, which appeared on the market in 1979. This computer was placed in a box with a screen. With this box interviewers went to the houses of respondents for interviews. These experiments have been described in Saris et al. (1981).

Table 8.1: Different methods for data collection distinguished on the basis of the role of the interviewer, the mode of presentation and the role of the computer

		Presented		
		ORAL	ORAL/VISUAL	VISUAL
Presence of interviewer	Role of computer			
Present	CAI	CAPI	CAPI+	CASI(-IP)
	NO	PAPI	PAPI+	-
Distant	CAI	CATI	-	WEB(-IP)
	NO	Tel.In	-	-
Absent	CAI	ACASI	ACASI+	DBM
	NO		ASAQ	ASAQ+
			TDE/VRE	MAIL

CAI=computer-assisted interviewing; CAPI=computer-assisted personal interviewing, CAPI+=CAPI plus show cards; PAPI=paper- and- pencil interviewing, PAPI+=PAPI plus show cards; CASI-IP=computer-assisted self Interviewing with an interviewer present; CATI=computer-assisted telephone interviewing; Tel. In=telephone interviewing; ACASI=audio computer-assisted self Interviewing; ACASI+=ACASI plus possibility to read the questions; Web=Web survey; Web-IP=Web survey plus interviewer present at a distance; DBM=Disk by Mail; Tel. in=telephone interview; T.P.=Telepanel; ASAQ=audio self-administered questionnaire; ASAQ+=ASAQ plus possibility of reading the questions; TDE=touchtone data entry; VRE=voice recognition entry. .

The other three distinctions mentioned previously have led to different procedures, displayed in Table 8.1. The majority of methods use oral presentation, but the newer methods increasingly use a visual presentation of the questions, with the exception of different experiments with audio systems. The procedures employing oral and visual presentations started with show cards in oral procedures or by reading facilities next to audio facilities. Below we will briefly discuss the possible consequences of the different choices.

8.1.2 The presence of the interviewer

A distinction that has always existed in survey research concerns the role of the interviewer. In personal interviewing the interviewer is present during the interview and normally has to ask the questions and record the answers. In telephone interviewing the task of the interviewer is the same, but the interviewer is not physically present at the interview and can ask the questions from a long distance. Finally, in all self-administered interviews there is no interviewer present at all.

The advantage of the presence of the interviewer during the interview, either in person or at a distance, is that the respondents do not have to have reading and writing abilities. Normally the interviewer will read the requests for an answer to the respondent. Another advantage is that the interviewer can help

with difficult questions. Some people suggest that the interviewer in a personal interview can also see the nonverbal reaction of the respondent, which is not possible in telephone interviews. Another distinction between the two approaches with an interviewer is that a limited number of people do not have a telephone and therefore cannot be reached by telephone. Another point is that people are more inclined to participate if an interviewer appears at the door than when the interviewer asks for cooperation through the telephone. Finally, the item nonresponse may be lower in a personal interview if the interviewer can build up a good relationship with the respondent during the interview.

However, the personal interview with an interviewer is costly because of the travel expenses and that the interview takes relatively more time. The presence of an interviewer might lead to more socially desirable answers to sensitive questions in a personal interview than over those administered over the telephone because of the "closeness" of the interviewer. However, although several studies showed this effect, others did not, as was summarized by Tourangeau et al. (2000).

A major difference between using interviewers or the approaches without an interviewer is that in the latter it is not possible for an interviewer to reformulate the questions as often happens in personal and telephone interviews (Van der Zouwen and Dijkstra 1996). This might be useful in helping the respondent but it also makes the answers incomparable if the question asked is not the same.

Another major difference is that the interviewer is not present, which in turn affects the answer results for sensitive issue questions. In general there is a lesser social desirability effect in self-administration procedures. A series of studies by Aquilino and LoSciuto (1990), Aquilino (1994), Gfrörer and Hughes (1992), Turner et al. (1998), and Schober and Conrad (1997) have demonstrated these effects. Tourangeau and Smith (1996) and Tourangeau et al. (1997) show that these effects are even larger if ACASI is used instead of simple CASI or SAQ.

The last findings are certainly of major importance and suggest that for sensitive issues it is much better to opt for self-administrated methods instead of interviewer-administered questionnaires. Otherwise, the differences are not significant. Even the coverage error in telephone interviewing does not have a dramatic effect because the number of people without a telephone is rather small in most countries and therefore the effect of this deviant group on the total result is in general rather insignificant (Lass et al. (1997).

That does not exclude the possibility that in certain cases one mode can be more effectively used than another. For example, Kalfs (1993) demonstrated that the time spent on transport is more suited for telephone than for self-administered interviewing because the interviewer can be instructed to check the sequence of events better than respondents can. On the other hand, she also found that self-administered interviewing worked better for recording TV watching than telephone interviewing because higher-educated people

were not willing to report their total TV viewing time in the CATI mode. They reported approximately half an hour more TV viewing time in the CASI mode.

So we have to conclude that the quality of the measures depends very heavily on the topic studied. If the topic is rather complex an interviewer can be helpful. If the topic is simple but sensitive, self-completion is advisable. However, if the topic is simple and not sensitive, the presence or absence of the interviewer will not make much difference.

8.1.3 The mode of presentation

The second characteristic to be discussed is the presentation mode of the requests for answers. In personal and telephone interviewing it is natural to make an oral presentation. For self-administered methods the most natural procedure of presenting is visual.

More recently mixed procedures have also been developed. For example, a very important tool in personal interviewing nowadays is the show card. While the interviewer reads the question, the respondent receives a card with the response alternatives. In case of a battery of statements, the respondents can also be provided with a card representing the different statements about which they have to give a judgment.

Another new possibility in self-administered surveys with a computer is that the text on the screen is presented and read to the respondent. This is typically the case for TDE, VRE ASAQ, and ACASI.

Tourangeau et al. (2000) point out that the cognitive requirement is higher for a visual presentation than for an oral because people have to be able to read. The problem of illiteracy received more attention in the United States than in Europe because of that the ASAQ and ACASI methods are more popular there than in Europe. Much of the research in Europe is still administered by personal or telephone interviews, and so the illiteracy problem does not play such an important role.

However, the visual or mixed presentation is very helpful because it appears that people in general have a limited capability to remember the information with which they are provided. For example, for the oral interview the respondent is unable to go back and check information. However, if visual information is provided in the form of response categories or statements, then the respondent can quickly go back and recapture the information to give a better response. That is why show cards have gained in popularity in personal interviews, which is not possible for telephone interviews.

There is also the unexpected result of a completely visual presentation that is the effect on the actual formulation of the questions. Normally survey researchers try to include the response categories in the requests for an answer. For example:

8.1 *Are you very much in favor, much in favor, in favor, against, much against or very much against the extension of the EU with central European countries?*

This request for an answer can be much simpler and more naturally formulated in a completely visual presentation:

8.2 How much are you in favor or against the extension of the EU with central European countries?

1. *Very much in favor*
2. *Much in favor*
3. *In favor*
4. *Against*
5. *Much against*
6. *Very much against*

This difference does not have to exist, of course, because the second form of the question could also be used in oral presentations if the interviewer is instructed to read all the response categories or if a show card with the response categories is used. However, looking at the survey literature, questions like the first one are rather common even though their formulation is not very attractive.

A major disadvantage of the completely visual paper representation is the lack of process control about the way the respondent answers the questions. A respondent can skip questions that should not be skipped or answer questions in a standard way without giving due diligence for the specific characteristic of each question. Dillman (2000), who has focused his attention on addressing this problem, suggests designing the layout of the pages in order to guide the respondent in the desired sequence to answer the questions. We will come back to this issue when we talk about layout design. On the other hand, lack of individual attention for each separate question also occurs during oral presentations in the case of the interviewer who would like to advance the interview as quickly as possible. In both cases this can lead to the problem of response set (Krosnick 1991). However it should be mentioned that visual representations used in self-completion surveys often take more time than do personal or telephone interviews and are often done at a time chosen by the respondent. From this it can be inferred that there is less time pressure in visual presentations than in oral presentations, and this could improve the overall response quality.

8.1.4 The role of the computer

Although the use of computers in survey research has revolutionized data collection, the difference between procedures with and without a computer is not always as great as expected. For example, from the respondents' perspective, personal interviewing and telephone interviewing has not fundamentally changed because of the introduction of the computer (Bemelmans-Spork and Sikkel 1986). The procedural difference has occurred mainly for the interviewers. In fact, their task is simplified because in a good CAPI and CATI interview the routing in the questionnaire is controlled by the computer and the interviewer no longer has a need to plan the next question anymore and can focus on the interview process.

Moreover, the computer could perform more of the interviewer's tasks. For example, the computer can

- Check whether the answers are appropriate
- Provide the respondents with more information, if necessary
- Code open answers into predetermined categories
- Store the given answers in a systematic way

The computer has the potential to reduce the interviewer's tasks considerably. However, the interviewer is still better at

- Obtaining cooperation
- Motivating the respondent to answer
- Building up sufficient confidence for honest answers

But the computer can do certain tasks quicker than the interviewer, such as

- Calculations for routing
- Substitution of previous answers
- Randomization of questions and answer categories
- Validation of responses
- Complex branching
- Complex coding
- Exact formulation of questions and information
- Provision of help

For a detailed discussion of how these tasks can be performed by computer programs, we refer the reader to Saris (1991). However, these extra facilities have their price. In order to obtain all of them, a considerable amount of time to develop computer-assisted interviews has to be invested. In addition, more time is needed to check whether all these tasks have been formulated properly. Programs are available to check automatically that a question can be reached and does not lead to a dead end, but all sequences need to be checked on substantive correctness. This cannot be done by computer programs. Wrong routings in a paper questionnaire can be corrected rather easily by the interviewer; however, this is not the case in computer-assisted interviews. Therefore checking the questionnaire is an essential task of development.

Nevertheless, checking the structure of the questionnaire is not enough. The respondent or interviewer can also make mistakes, and as a consequence the routing might be wrong. For example, a respondent who does not have a job can be coded as having one. In that case the respondent may get all kinds of questions about labor activities that do not apply. To prevent such confusions, Saris (1991) has suggested summary and correction (SC) screens, which summarize prior provided information to be checked by the respondent before the program makes the next branching decision. This SC screen turns out to be a very efficient tool in computer-assisted data collection. An interviewer can be taught complex computer tasks that are too much for respondents. Therefore, for self-administered interviews, without an interviewer present, the SC screen

is even more important than for CAPI and CATI. How this tool and others can be developed can be found in Saris (1991).

The role of the computer in self-administrated questionnaire applications can be most helpful. First, the interview program:

- Can perform all the tasks we have previously mentioned, which can be used to substitute the interviewer except for obtaining cooperation
- Can ensure that the respondent will not skip a question by mistake
- The program provides automatic visual questions and can also show pictures or short movies
- The program can read the text of the questions as has been developed in ACASI
- Can use line production scales which have turned out to be very effective to obtain judgments on a continuous scale
- Can be used to present complex instructions which are not possible in oral presentations

In the case of a panel of respondents for the data collection it is possible to use previous answers to reduce the effort of the respondent and to verify the answers on unlikely responses. An example is the *dynamic SC screen*" where respondents can check the correctness of previous answers without filling in the forms again. If a change has occurred, they can restrict their work to the necessary corrections on the screen. Another example is the *dynamic range* and *consistency check*, where previous answers with respect to income, prices, or quantities are used in later interviews, that take into account the variability in these quantities by using a dynamic confidence interval around the point estimator (Hartman and Saris 1991).

This brief overview of the potential of computer-assisted interviewing indicates that a considerable increase in data quality can be obtained by efficient use of the computer. Tortora (1985) conducted an experiment that demonstrated that checks available in CATI could prevent nearly 80% of all the corrections that were normally needed after the data collection. As we have seen, even more advantages can be obtained in self-administered questionnaires improving mail questionnaires, including the commonly used diaries (Kalfs 1993). So far these data quality improvements have been obtained mainly with respect to factual information. For subjective variables the results are more difficult to improve because it is more difficult to specify rules for inconsistent answers. An interesting possibility is the development of procedures to detect unlikely answers elaborated by Martini (2001).

We have to mention that this quality improvement by CAI also has its costs, not only in terms of hardware but also in the time one has to spend for developing computer-assisted interviews, followed by checking the correctness of questionnaire routing. The mere fact that a computer is used is not enough to improve data quality. Considerable time has to be spent on the development of computer-assisted questionnaires which makes sense only for very expensive or often repeated studies. Therefore, improved quality is not guaranteed by

using computer-assisted data collection. Only if special attention is given to the development of the questionnaire, these advantages are obtained.

This point is important to emphasize because a mistaken assumption with the development of the Web survey is that it does do survey research at very low costs. However, there are three problems with this approach: (1) it is impossible to get a representative sample of a national population through the Web at this moment – there are still too few people connected to the Web, and this group is rather deviant from the population as a whole; (2) the cooperation with Web surveys is not such that one can expect to get a representative sample from the population; and (3) few researchers incorporate the possibilities of computer-assisted interviewing that were mentioned previously. As a consequence, formulation of the questionnaires is most usually on the level of the standard mail survey. In this case a step backward is taken in terms of survey quality, in comparison with what is possible, and even though computer-assisted data collection is used. Better procedures are used by Centerdata and Knowledge Network (Couper 2000), which follow the telepanel procedure, providing a random sample of the population with the necessary equipment to use real computer-assisted data collection with consistency checks and other possible tools for data collection.

8.1.5 Procedures without asking questions

So far we have discussed procedures that present requests for an answer to the respondents. In marketing research procedures have been developed where no questions are asked at all anymore. Two popular approaches are the "people meter" and the "barcode scanner." The former is used for recording the amount of time that people spend watching different TV programs. It is a box placed on the TV that can record any program that is viewed on it. In order to know who is watching the program the viewers have to indicate on a remote control whether they are watching. In this way their TV watching behavior is registered without asking any question.

The barcode scanner is used for recording the purchases of consumers. The most efficient procedure uses the barcode reader at the cash register. In order to be able to connect the purchases to a person, the consumer is first asked to present an identity card with a barcode. The code is registered, followed by all goods bought to account for all purchases without asking any question in person.

These systems also have their shortcomings, but they are much more efficient than the traditional procedures using diaries or asking questions. For a discussion around issues concerning these topics, we can refer to Belson (1981) and Kalfs (1993), for TV viewing to Silberstein and Scott (1991), and to Kaper (1999) for consumer behavior. We will not elaborate on further details here, because we will concentrate on approaches using requests for an answer. We will conclude by stating that automatic registration is more efficient but also much more expensive.

8.1.6 Mixed-mode data collection

The most commonly used procedures have different response rates and different costs associated with them. A general trend for all modes of data collection is that the response rate decreases. In order to improve the response rate, mixed-mode data collection has been suggested which employs several data collection methods for one study. The mixed-mode design is developed on the basis of the knowledge that different methods lead to different levels of nonresponse for different groups. These unequal response probabilities make mixed-mode data collection attractive, even more so when the (financial) costs of the different data collection methods are taken into account. Mail or web questionnaires are by far the most economical of the three traditional data collection modes, followed by telephone interviewing, while face-to-face interviewing is the most expensive one (Dillman 1991; Pruchno and Hayden 2000). Because the response rates of these three data collection methods are inversely related to financial expenses, it seems to pay off to start with the cheapest data collection method (mail, web), following up the nonrespondents by telephone and approach those who still are not reached or not willing to participate, by a personal interview. In this way, the highest response level at the lowest cost can be achieved and the differences of the selection process between the three data collection modes are turned to the advantage of the survey researcher.

The main disadvantage of mixed-mode data collection is the possibility of mode effects. The results with respect to mode effects are not so clear yet. It is most likely that it depends on the topic being studied. This means that pilot studies are needed before the mixed-mode data collection can be used in large-scale survey research. An example of such a study is done by Saris and Kaase (1997) with respect to a comparison of telephone and face-to-face research for the Eurobarometer. They found significant differences between the methods but indicated as well procedures for how to overcome these problems. A study has been done by Voogt and Saris (2003) dealing with election studies. They did not find mode effects comparing mail and telephone interviewing, but the personal interviewing gave results different from the other two mentioned earlier. A lot of useful suggestions to minimize the effects of the mode can be found in Dillman's work (2000).

8.2 THE POSITION IN THE QUESTIONNAIRE

A next issue that requires attention is the construction of the questionnaire. So far we have discussed only single requests for an answer or batteries but not the ordering of the different requests or batteries in a questionnaire. In this context there are four principles that require attention. The first is that a prior request for an answer can have an effect on a later request for an answer (Schuman and Presser 1981). The second principle is that one should not mix all requests randomly with each other as is often done in omnibus surveys. Dillman (2000) is a strong supporter of ordering the question by topic. However, this increases the risk of ordering effects. He also suggests a third principle to start question-

naires with the topic that has been mentioned to the respondents to get their cooperation. These questions should be relatively simple, apply to all respondents, and also be interesting in order to increase the cooperation to respond. Contrary to this rule is a fourth principle that suggests that the answers to the first requests are probably not as good as later responses because the respondents have to learn how to answer and to gain confidence with the interviewer and the interview process. According to this principle, it is advisable not to ask about the major topic of the survey immediately at the beginning of the interview.

The first two rules concern the choice to order the survey items by topic or not. In psychological tests the items are normally randomly ordered going from one topic to another without any connecting statement between the different questions. This is done to avoid sequence effects. Dillman (2000) argues that the respondents get the impression that the interviewer does not listen to their answers because the next question has nothing to do with their prior answer. Another problem is that it also puts a heavy cognitive burden on the respondents because they have to search for answers in a different part of their memory and it is questionable whether most respondents are willing to do so. This might lead to satisficing behavior as suggested by Krosnick (1991), which means that the respondent does not look for the optimal answer anymore, but only for an acceptable answer with minimal effort, as for example, a "don't know" answer or the same response.

The other side of the coin is that grouping the questions by topic can lead to order effects. An example of an order effect is called the "evenhandedness effect" (Hyman and Sheatsley 1950). The number of "yes" answers to "Should a communist reporter be allowed to report on a visit to America as he/she saw it?" increases when the respondents were first asked "Should an American reporter be allowed to report on a visit to the Soviet Union as he/she saw it?" Dillman (2000) mentions several other similar examples of effects of one question on the next for distinct reasons. One is called the "anchoring effect" and suggests that a distant object is evaluated differently if asked first than if it is evaluated after a more familiar local object. Another is called the "carryover effect," which means that, for example, happiness judgments are more positive after a question about marriage than before the question. It has also been shown that overall evaluations become lower after the evaluations of some specific aspects. For more examples we refer the reader to Schuman and Presser (1981); their examples suggest watching out for the possibility of order effects if requests for answers have an obvious relationship. Nevertheless, we think that it is better to order the questions by topic instead of using a random order of the request for answers because it improves the cooperation of the people and reduces their cognitive burden, which is in general quite high.

The second pair of contradictory rules suggests (1) starting with the main topic of the questionnaire using simple questions that apply to all and are interesting for the respondents and (2) not starting with the main theme of the

study until the people are more familiar with the procedure and the interview process and have confidence in the interviewer. According to rule 1, starting with questions about the background of the respondent or household would be a mistake, while according to rule 2 it could be acceptable.

Although we prefer the rule to start as soon as possible with the topic announced in order to get the cooperation of the respondent and to start with simple questions that apply to all people and are interesting, we don't suggest that the main topic should be in the first question. In the second part of this book we will demonstrate that the respondents continuously learn how to answer and as a consequence their answers improve. Therefore, it is preferable to delay the most important questions in order to elicit better responses. In general, respondents will not be surprised to answer some more general questions about their background before the main topic questions. Since these questions are simple and general, they serve to familiarize the respondent with the procedure and the interviewer

After a general introduction topics that are clearly related to the main topic should be introduced. The order of the different topics should be partially determined by the complexity of the questions, because the later that complex questions are asked, the better the respondents will be able to answer the questions. However, these complex questions should not be asked at a moment that the respondents are starting to get bored with the interview. Therefore, the best ordering of the topics should be based on the logical sequence of the topics and the quality of the responses, which will be discussed in Part III of this book.

8.3 THE LAYOUT OF THE QUESTIONNAIRE

Although the layout of the questionnaire is important for any questionnaire, the quality of the layout is more important in self-administrated questionnaires than in interviewer-administered questionnaires. Interviewers can be taught the meaning of the different signs and procedures in the layout. This is normally not possible in self-administered questionnaires; therefore, the layout of such a questionnaire should be completely self-evident.

This is even more important for paper questionnaires than for computer-assisted self- administrated questionnaires. For paper questionnaires the respondent has to find the routing of the questionnaire that skips the fewest requests for an answer by mistake and that answers the requests in the proper order. For computer administered questionnaires the computer can take over this task. However, in the new Web surveys often this routing task is not taken over by the program, therefore producing rather incomplete response files.

Unfortunately very little is known about the optimum rules for designing questionnaires. The best source for this information is Dillman (2000), who formulated a number of general rules mainly for mail questionnaires and in his recent work also for questionnaires presented on computer screens. For more detailed information, we refer the reader to his work. A general rule of thumb is to choose one system for the layout and to be consistent throughout the whole questionnaire so that the respondent and interviewer will know what to expect.

8.4 Differences due to use of different languages

The researcher does not have much choice over the language to be used in the questionnaire. Normally a country has one national language, and that is the language to be used to formulate the questions; for large minority groups a questionnaire in the language of the minority group can also be considered. Such translation tasks are rather complex because a functionally equivalent version of the questionnaire needs to be generated (Harkness et al. 2003) wherein the different translations of the requests for answers have an equivalent meaning.

In the European Social Survey, which is conducted in 23 countries in Europe, a translation into more than 20 languages was necessary. This translation process was organized in the following way. First a source questionnaire was developed in several steps in English (ESS 2002). This questionnaire, with annotations with respect to the concepts used was sent to the different countries. In each country two translators were asked to make a translation independently of each other. The resulting questionnaires were provided to a reviewer who looked at the differences between the two translations. Finally, the national coordinator of the study together with the reviewer made the final decisions with respect to the definite formulation of the questionnaire in each respective national language (Harkness et al. 2003). After this complex and laborious process, a set of questionnaires with as similar as possible requests for answers and response categories were developed.

Although all due diligence is taken in cross-national research there is no guarantee that the results from the different countries or cultural groups are comparable. Even if all the texts produced are completely functionally equivalent it can not be excluded that the same request for an answer generates more random errors and/or systematic errors in one country than in an other. Illustrations have been provided by Saris (1997) for the Eurobarometer and in Saris (2003) for the ESS. How such tests can be conducted will be discussed in further detail in Parts III and IV. Given these findings, their argument is to compare results of different countries after correcting for measurement error; otherwise it is unknown whether the differences between the countries are due to measurement error or represent real differences.

8.5 Summary and discussion

In this last chapter of Part II, an overview of the developments in data collection methods was provided. In that context we have discussed three choices that in combination determine the data collection method to select

1. The presence or absence of an interviewer
2. The mode of presentation (oral/visual)
3. The role of the computer

With respect to the presence or absence of an interviewer, we have to conclude that the quality of the measures depends heavily on the topic studied.

Depending on whether the topic is rather complex, an interviewer can be helpful. However, if the topic is simple but sensitive, self-completion is advisable. When the topic is simple and not sensitive, the presence or absence of the interviewer will not make much difference.

The mode of presentation can make quite a difference for sensitive topics. We have also mentioned that the mode of data collection will change the way the questions are formulated; for this reason it is difficult to study pure mode effects.

The role of the computer in survey research is very interesting as it can take over many tasks of the interviewer. The computer can also do certain tasks that an interviewer cannot do, leading to considerable improvement of a questionnaire. On the other hand designing computer-assisted questionnaires is a very time-intensive process and therefore also rather expensive. Because of the costs involved we recommend this method for large-scale studies and longitudinal studies. However, computer-assisted data collection does not always improve the quality of the surveys. To date the improvements that have been noted are mainly for objective variables. For subjective variables it is more difficult to specify efficient consistency checks.

A second issue discussed was the ordering of the questions within the questionnaire; we advise the reader to be aware of the possibility of order effects in case of requests for answers that have an obvious relationship. Nevertheless, we think that it is better to order the questions by topic instead of using a random order of the requests for an answer. This will improve the cooperation of the respondents and reduce their cognitive burden.

The third issue discussed concerned the layout of questionnaires. In general it is recommended that the researcher chooses only one system for the layout and adhere to it throughout the whole questionnaire. This is important not only for self-administered questionnaires but also for interviewer-administered ones, because then the respondent and interviewer both know what to expect.

The last topic was the effect of the language used. Normally, the official language cannot be chosen, but often there is a choice in using a different language for large minority groups. In such cases, the questionnaires are not just translations but have to be made as functionally equivalent as possible. However, this does not guarantee that the results for different groups or countries can be compared. It is possible that in the different groups the reactions on the optimally equivalent requests create different types of errors which lead to different results. Therefore it was suggested that one can never compare results directly across groups without correction for measurement error, which will be further discussed in Part IV.

Exercises

Imagine that we want to design a questionnaire to determine the satisfaction of the people with the government and the reasons for this satisfaction. This could be done in different ways, and the next questions deal with these different options.

1. The most important question, of course, is about satisfaction with the government with a 7-point completely labeled response scale. Formulate such a request for the following data collection modes:
   a. For a mail survey
   b. For a personal interview in two ways: one with a show card and one without
   c. For telephone interviewing, also in two ways: one in two steps, one direct.
   d. Which one would you prefer?
2. Next we would like to know which policies influence the satisfaction with the government the most. We want to ask about the economy, health service, education system, security, and so on.
   a. Formulate such a block of requests in different ways (and in detail)
   b. Should these requests be asked before or after the general satisfaction request mentioned in question 1.
3. Routing in paper questionnaires requires special attention. Take as an example the two step procedure asked as request 1c.
   a. Indicate in detail how you would make the layout of the page to avoid problems
   b. How would this routing be done in CAI?
4. Other questions concern the sources of income (salary, pension, unemployment money, savings, alimentation, etc.), especially the amount of money they get from each source.
   a. Formulate a question block for a face-to-face study to ask these questions
   b. How can a computer-assisted data collection simplify the task taking into account the position of the people on the labor market?
   c. In computer-assisted panel research this can be made even simpler using prior information. How?
5. As a last part in this questionnaire we want to formulate questions about satisfaction with income.
   a. Formulate a question asking for an absolute judgment with a labeled 11-point scale for a face-to-face study
   b. Formulate also a question asking a relative judgment with a labeled 11-point scale for a face-to-face study.
   c. Can we ask the same questions by telephone? If not, how should the procedure be adjusted?
   d. Do you think that the order of these questions will make a difference?
   e. Can we use the answers to these two questions to check for consistency? If so, how can that be done?
6. For the study you are designing choose the most appropriate mode of data collection. Determine the order of the questions and specify the layout of the questionnaire.

Part III

The effects of survey characteristics on data quality

Until now we have been discussing the different choices that have to be made in order to design survey items, a questionnaire, and a data collection instrument. The design of survey questionnaires can be a scientific activity, if the effect of the different choices on the data quality can be estimated.

In order to study these effects, we need to consider the following steps:

1. Establishing criteria for the quality of survey questions (Chapter 9)
2. Estimation of reliability, validity and method effects (Chapter 10)
3. Estimation of the effects of the measurement characteristics on the quality of the survey questions (Chapter 12)

Chapter 11 is rather specific and can be skipped without loosing sight of the argument in this book. It was added for professionals who would like to do data quality experiments themselves.

This Page Intentionally Left Blank

CHAPTER 9

Criteria for the quality of survey measures

In Part I and II we have seen that the development of a survey item demands making choices concerning the structure of the item and the data collection procedure. Some of these choices follow directly from the aim of the study, such as the choice of the topic of the survey item(s) (church attendance, neighborhood, etc.) and the concept measured by the request for an answer (evaluations, norms, etc.). But there are also many choices that are not fixed and these choices will influence the quality of the survey items. They have to do with the formulation of the request, the response scales, and any additional components such as introduction, motivation, position in the questionnaire, and the mode of data collection. So, it is highly desirable to have some information about the quality of a survey item before it is used in the field.

Several procedures have been developed to evaluate survey items before they are used in the final survey. The oldest and most commonly used approach is, of course, the use of pretests and debriefing of the interviewers regarding any problems that may arise in the questionnaire. Another approach, suggested by Belson (1981), is to ask people during a pretest, after they have answered a request for an answer, how they interpreted the different concepts in the survey item while they were answering the requests. A third approach is the use of "think aloud" protocols during interviews. A fourth approach is to assess the cognitive difficulty of a request for an answer. This can be done by an expert panel (Presser and Blair 1994) or on the basis of a coding scheme (Forsyth et al. 1992; Van der Zouwen 2000) or by using a computer program (Graesser et al. 2000a,b). The latter authors developed a computer program to evaluate survey items in relation to their linguistic and cognitive difficulty. A fifth approach, which is now rather popular, is to present respondents with different formulations of a survey item in a laboratory setting in order to see what the effect of these wording changes is (Esposito et al. 1991; Esposito and Rothgeb 1997; Snijkers 2002). For an overview of the different possible cognitive approaches to the evaluation of requests we recommend Sudman et al. (1996). A rather different approach is interaction or behavioral coding. This approach checks to see whether the interaction between the interviewer and the respondent

follows a standard pattern or whether deviant interactions occur (Dijkstra and Van der Zouwen 1982). Such deviant interactions can indicate problems in the questionnaire related to specific concepts or the sequence of the survey items.

All these approaches are directed at detecting response problems. The hypothesis is that problems in the formulation of the survey item will reduce the quality of the responses of the respondents. However, the standard criteria for data quality, such as validity, reliability, method effect, and item nonresponse, are not directly evaluated. Campbell and Fiske (1959) suggested that validity, reliability, and method effects can be evaluated only if more than one method is used to measure the same trait which is in our research a concept by intuition. Their design is called the multitrait-multimethod (MTMM) design and is widely used in psychology and psychometrics (Wothke 1996). This approach has also attracted attention in marketing research (Bagozzi and Yi 1991). In survey research, it has been elaborated and applied by Andrews (1984), whose method has been used for different topics and request forms in several languages: English (Andrews 1984), German (Költringer 1995), and Dutch (Scherpenzeel and Saris 1997). Andrews (1984) also suggested using a meta-analysis of the available MTMM studies to determine the effect of different choices made in the design of survey requests on the reliability, validity, and method effects. Following his suggestion, Saris and Gallhofer (2007) conducted a meta-analysis of the available 87 MTMM studies to summarize the effects that different request characteristics have on reliability and validity. Chapter 12 describes their results, and Chapter 13 indicates how these results can be used to predict both the quality of survey items before they are used in practice and how formulations of survey items can be improved where the quality of the original formulation is insufficient. In this chapter we will illustrate some effects of these choices, followed by indicating what criteria should be used to evaluate the quality of survey requests. In Chapters 10 and 11 we discuss procedures to evaluate questions with respect to the selected quality criteria.

9.1 DIFFERENT METHODS, DIFFERENT RESULTS

Normally all variables are measured using a single method. Thus, one cannot see how much of the variance of the variables is random measurement error and how much is systematic method variance. Campbell and Fiske (1959) suggested that multiple methods for multiple traits should be employed in order to detect error components. The standard MTMM approach nowadays uses at least 3 traits that are measured by at least 3 different methods, leading to 9 different observed variables. In this way a correlation matrix of 9×9 is obtained. In order to illustrate this type of procedure, Table 9.1 presents a brief summary of a MTMM experiment conducted in a British pilot study for the first round of the European Social Survey (2002). Three different traits and three methods were used in the study.

Table 9.1: A MTMM study in the ESS pilot study (2002)

For the three traits the following three requests were employed:

1. On the whole, how satisfied are you with the present state of the economy in Britain?
2. Now think about the national government. How satisfied are you with the way it is doing its job ?
3. And on the whole, how satisfied are you with the way democracy works in Britain?

In this experiment the following response scales were used to generate the three different methods:

Method 1: (1) Very satisfied; (2) fairly satisfied; (3) fairly dissatisfied: (4) very dissatisfied

Method 2:

Very dissatisfied *Very satisfied*

0 1 2 3 4 5 6 7 8 9 10

Method 3: (1) Not at all satisfied; (2) satisfied; (3) rather satisfied; (4) very satisfied

In this study, the topic of the survey items (national politics/economy) remained the same across all methods. Also the concept measured (a feeling of satisfaction) is held constant. Only the way in which the respondents are asked to express their feelings varies. The first and third methods use a 4-point scale, while the second method uses an 11-point scale. This also means that the second method provides a midpoint on the scale, while the other two do not. Furthermore, the first and second methods use a bipolar scale while the third method uses a unipolar scale. In addition, the direction of the response categories changes in the first method compared with the second and the third methods. The first and third method have completely labeled categories, while the second method has labels only at the endpoints of the scale.

There are other aspects in which the requests are similar, although they could have been different. For example, in Table 9.1 direct requests have been selected for the study. It is, however, very common in survey research to specify a general request such as "How satisfied are you with the following aspects in society?" followed by the provision of stimuli such as the present economic situation, the national government, and the way the democracy functions. Furthermore, all three requests are unbalanced, asking "how satisfied" people are without mentioning the possibility of dissatisfaction. They have no explicit "don't know" option, and all three have no introduction and subordinate clauses, making the survey items relatively short. There is no need to discuss here other relevant characteristics of requests because they have already been covered in Part I and II of this book.

Identical characteristics of the three requests cannot generate differences, except for random errors, but those aspects that do differ, can generate differ-

ences in the responses. Many studies have looked at the differences in response distributions (Schuman and Presser 1981). Table 9.2 presents a summary of the responses. We have made the responses as comparable as possible by summing up categories that can be clustered.

Table 9.2: The response distribution for the 9 requests specified in Table 9.1.

	Satisfaction with the								
	Economy			Government			Democracy		
	method			method			method		
Responses	1	2	3	1	2	3	1	2	3
Dissatisfied	167	134	99	268	208	169	187	152	169
Neutral	-	102	-	-	100	-	-	100	-
Satisfied	273	193	320	191	128	258	223	176	258
Very satisfied	26	7	12	11	2	4	43	7	4
Missing	19	49	54	15	47	54	32	50	54

Table 9.2 shows that quite different results are obtained depending on what method for the formulation of the answer categories is used. If we did not know that these answers come from the same 485 people for the same requests, we could conclude that these responses come from different populations or that the requests measure different opinions.

One obvious effect is the effect of the neutral category of the second method, which changes the overall distribution of the answers. Another phenomenon that seems to be systematic is that the third method generates much more "satisfied" responses than do the other two. This has to do with the unipolar character of the scale and the extreme label for the negative category, which is "not at all satisfied." This label seems to move some respondents to the category "satisfied" where they otherwise are "neutral" or "dissatisfied." It also appears that the number of people saying that they are "very satisfied" decreases if more response categories are available to express satisfaction: method 1 has only 2, method 2 has 5 and method 3 has 3 possibilities.

Finally, a very clear effect can be observed for the number of missing values. This might have to do with the positioning of the request for an answer in the questionnaire and other characteristics of the questionnaire.

Because the same people were asked for all 9 requests, it is possible to look at the cross-tables of the requests for the same topic, which demonstrates what the link between the different responses is. In Tables 9.3 and 9.4 we present the results for two of the three possible combinations for satisfaction with the economy.

Table 9.3: The cross-table of satisfaction with the economy measured with methods 1 and 3*

Measured with	Method 1				
Method 3	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied	Total
Satisfied	10	173	44	7	234
Rather satisfied	6	54	13	4	77
Very satisfied	8	4	0	0	12
Total	24	250	99	47	420

* These data were taken from the British pilot study of the ESS.

Table 9.3 has the following striking results:

- 68 respondents claim that that they are dissatisfied with method 1 and satisfied with method 3.
- 19 respondents claim that that they are satisfied with method 1 and dissatisfied with method 3.
- 16 respondents claim that that they are very satisfied with method 1 and are only satisfied with method 3.
- 4 respondents claim that they are only satisfied with method 1 and are very satisfied with method 3.

Table 9.4: The cross-table of satisfaction with the economy measured with methods 1 and 2

Measured with	Method 1				
Method 2	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied	Total
Very dissatisfied					
0	0	1	5	12	18
1	0	3	2	8	13
2	0	4	8	8	20
3	0	12	22	7	41
4	0	19	18	1	38
5	3	61	24	9	97
6	0	41	12	1	54
7	2	60	6	1	69
8	7	43	4	1	55
9	5	9	0	0	14
10	7	0	0	0	7
Very satisfied					
Total	24	250	99	47	420

A similar analysis can be made for Table 9.4. In this table we see that:

- 25 respondents claim that they are dissatisfied with method 1 and satisfied with method 2.
- 39 respondents claim that they are satisfied with method 1 and dissatisfied with method 2.
- 9 respondents claim that they are very satisfied with method 1 and only satisfied with method 2.
- 9 respondents claim to be only satisfied with method 1 and very satisfied with method 2.

In summary, Tables 9.3 and 9.4 had 107 inconsistent answers out of approximately 420 respondents who answered both requests. Some of these inconsistent answers may be mistakes, but it is also possible that there is a systematic pattern in these inconsistencies due to the 3 methods. To clarify whether that is the case, we also looked at the same two tables for the topic "satisfaction with the government." The results are presented in Tables 9.5 and 9.6.

Table 9.5: The cross-table of satisfaction with the government measured with methods 1 and 3

Measured with	Method 1				
Method 3	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied	Total
Not at all satisfied	0	12	89	67	168
Satisfied	4	124	67	2	197
Rather satisfied	4	35	15	1	55
Very satisfied	2	2	0	0	4
Total	10	173	171	70	424

Table 9.5 shows the following errors:

- 85 respondents claim to be dissatisfied with method 1 and satisfied with method 3.
- 12 respondents claim to be satisfied with method 1 and dissatisfied with method 3.
- 8 respondents claim to be very satisfied with method 1 and satisfied with method 3.
- 2 respondents claim to be only satisfied with method 1 and very satisfied with method 3.

In Table 9.6 it can be seen that:

- 26 respondents claim to be dissatisfied with method 1 and satisfied with method 2.
- 31 respondents claim to be satisfied with method 1 and dissatisfied with method 2.

Table 9.6: The cross-table of satisfaction with the government measured with methods 1 and 2

Measured with	Method 1				
Method 2	Very satisfied	Fairly satisfied	Fairly dissatisfied	Very dissatisfied	Total
very dissatisfied					
0	0	1	5	27	33
1	0	5	8	13	26
2	0	3	27	12	42
3	0	7	37	9	53
4	1	14	35	2	52
5	1	49	38	7	95
6	1	28	13	0	42
7	0	29	9	0	38
8	2	32	3	1	38
9	4	5	0	0	9
10	1	1	0	0	2
very satisfied					
Total	10	174	175	71	430

- 7 respondents claim to be very satisfied with method 1 and only satisfied with method 2.
- 1 claims to be satisfied with method 1 and very satisfied with method 2.
- 36 respondents claim to be very dissatisfied with method 1 and only dissatisfied with method 2,
- 5 respondents claim to be only dissatisfied with method 1 and very dissatisfied with method 2.

In Tables 9.5 and 9.6 we discovered 107 inconsistent answers out of the approximately 425 respondents who answered both requests. As mentioned previously, some inconsistency may be *random errors*, but there are also some evident patterns in these errors. For example, there are many more people claiming to be dissatisfied with method 1 and satisfied with method 3 than vice versa. This is also true for Tables 9.3 and 9.5. This phenomenon cannot be found in Tables 9.4 and 9.6, where the effect is rather reversed. But there we see that there are many more extreme responses for method 1 than for method 2. These results seem to suggest that there are random errors due to mistakes but also *systematic effects* connected with the differences between the methods. This issue will be discussed in a later section.

Given the results, the general conclusion should be that the correspondence between the different measures for the same people is very low. A measure

commonly used to determine the correspondence of responses is the correlation coefficient. It does not come as a surprise that the correlations in these tables are not very high, even though the requests are supposed to measure the same traits. The correlations in these tables are as follows: Table 9.3, –.502; Table 9.4, –.626; Table 9.5: –.608; Table 9.6: –.663. We have also calculated the same correlations for the concept "satisfaction with the democracy." The respective correlations are –.566 and –.669. We see that these relationships are rather weak since the proportion of similarity is equal to the correlation squared, which peaks around .40.

Let us now look at what happens to the correlations between the measures of the different traits. Table 9.7 presents the correlations between the 9 measures.

Table 9.7: Correlations between the 9 variables of the MTMM experiment with respect to satisfaction with political outcomes

	Method 1			Method 2			Method 3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Method 1									
Q1	1.00								
Q2	.481	1.00							
Q3	.373	.552	1.00						
Method 2									
Q1	-.626	-.422	-.410	1.00					
Q2	-.429	-.663	-.532	.642	1.00				
Q3	-.453	-.495	-.669	.612	.693	1.00			
Method 3									
Q1	-.502	-.347	-.332	.584	.436	.438	1.00		
Q2	-.370	-.608	-.399	.429	.653	.466	.556	1.00	
Q3	-.336	-.406	-.566	.406	.471	.638	.514	.558	1.00

These results clearly indicate the need for further investigation of the quality of the different measures, since the correlations between the three requests Q1 to Q3 are very different for the different methods. For the first method the correlations vary between .373 and .552; for the second method between .612 and .693; and for the third method between .514 and .558.

All these results raise questions such as:

- How can such differences be explained?
- What are the correct values?
- What is the best method?

To answer these requests quality criteria for survey measures are required which is the topic of one of the next sections.

9.2 HOW THESE DIFFERENCES CAN BE EXPLAINED

In order to explain the differences discussed earlier, something has to be said about relations between variables in general; namely, how such relationships can be formulated and how these relationships and correlations are linked. After this general introduction we will apply our knowledge on the measurement situation we were discussing above.

9.2.1 Specifications of relationships between variables in general

In the literature a distinction is made between direct effects, indirect effects, spurious relationships, and joint effects. The different relations are illustrated in Figure 9.1.

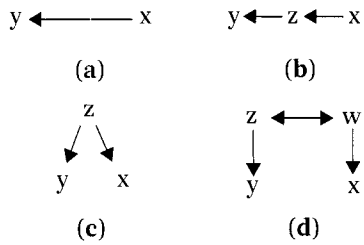


FIGURE 9.1: *Different relationships between the variables x and y: (a) a direct effect; (b) an indirect effect; (c) a spurious relation ; (d) a joint effect.*

The arrow in Figure 9.1a going from x to y indicates a *direct effect* of a variable x on a variable y. In Figure 9.1b there is no arrow from x to y directly, so there is no direct effect, but there is an *indirect effect* because x influences z and z in turn influences y, and so x has an influence on y but only indirectly. This is not the case in Figure 9.1c. There the relationship between y and x is called *spurious* because the two variables have neither a direct nor an indirect effect on each other, but there is a third variable z that influences both. Therefore, we can expect a relationship between x and y but this relationship is not due to any effect of x on y. Finally, in Figure 9.1d there is also no effect of x on y, but it is unclear where the relationship between x and y comes from, as the direction of the effect between z and w is unknown (indicated by a double-headed arrow). In this case z could produce a spurious relation and w could do so as well, making the relationship unclear. This type of relationship is called a *joint effect* due to z and w.

In order to make the discussion less abstract, let us look at the example in Figure 9.2. Figure 9.2 represents a causal model explaining "political interest." It is assumed that this variable is directly influenced by education (b_3) and "SES" (socio economic status) (b_4), and that "SES" is directly influenced by "income" (b_1) and "education" (b_2). We could continue with the explanation of the relationship between "income" and "education," but here we are not interested in these details, so we are leaving this relationship unspecified (ρ_{ie}).

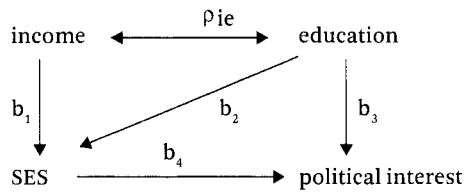


FIGURE 9.2: *An example of causal relationships.*

The different b coefficients indicate the size of the effects (for an explanation, we refer the reader to the appendix of this chapter) while ρ_{ie} represents the size of the correlation between the variables "income" and "education." Note that we only have to specify the direct effects because from these follow the indirect effects, spurious relations and joint effects.

In order to be more precise about the size of the indirect effects, spurious relationships, and joint effects, we specify that:

Theorem 1. *The indirect effect, spurious relations, and joint effects are equal to the product of the coefficients in the direction of the arrow from one variable to the other, without passing the same variable twice and going against the direction of the arrow.*

From such causal models predictions can be made concerning the size of the correlations between each pair of the variables, assuming that no other variables play a role in this process. These predictions are as follows:

Theorem 2. *The correlation between two variables is equal to the sum of the direct effect, indirect effects, spurious relationships, and joint effects between these variables.*

In the literature on structural equation modeling these two theorems have been proven. For a simple introduction, we refer the reader to Saris and Stronkhorst (1984); for a more complete discussion, Bollen (1989) is recommended.

Let us now apply these theorems to derive the size of the different correlations between the variables mentioned in the causal model of Figure 9.2. The results are presented below, where the different correlations are denoted by ρ(i,j).

ρ(income, education)	=	joint effect
	=	ρ_{ie}
ρ(income, SES)	=	direct effect+ joint effect
	=	$b_1 + \rho_{ie}b_2$
ρ(income, political interest)	=	indirect effect + joint effects
	=	$b_1b_4 + \rho_{ie}b_3 + \rho_{ie}b_2b_4$

$$
\begin{aligned}
\rho(\text{education, SES}) &= \text{direct effect} + \text{joint effect} \\
&= b_2 + b_1\rho_{ie} \\
\rho(\text{education, political interest}) &= \text{direct effect} + \text{indirect effect} + \text{joint effect} \\
&= b_3 + b_2b_4 + \rho_{ie}b_1b_4 \\
\rho(\text{SES, political interest}) &= \text{direct effect} + \text{spurious} + \text{joint effect} \\
&= b_4 + b_2b_3 + b_1\rho_{ie}b_3
\end{aligned}
$$

On the basis of the estimates of the different effects, predictions can be made about the size of the correlations of the variables. Note that in all cases the correlation between any two variables is not equal to the direct effect between the two variables. Sometimes there is even no direct effect at all while there will be a correlation due to other effects or relationships, for example, the correlation between "income" and "political interest."

In the next chapter we will show that these relationships between the correlations and the effects (also called *parameters*) can be used to estimate the values of these parameters if the sizes of the correlations are known from research. But in this chapter we concentrate on the formulation of models. So in the next section we will use this approach to specify measurement models.

9.2.2 Specification of measurement models

In the first part of this chapter we have shown that we can expect two types of errors: random and systematic. This means that the response variables we use in survey research will not be the same as the variables we want to measure. Looking only at the random errors, psychometricians (Lord and Novick 1968) have suggested the model of Figure 9.3.

FIGURE 9.3: *The classical model for random errors.*

This model suggests that the response variable (y) is determined directly by two other variables: the so-called *true score* (t) and the random errors represented by the random variable (e). The true score variable is the observed response variable corrected for random measurement error. For example; imagine that we measure "political interest" by a direct request "How interested are you in politics?" using an 11-point scale and we assume only random mistakes in

the answers. For example, by mistake some people pick a score 6 while their true score is 5, while others choose a 3 by mistake and their true score is 4, and so on. Then y represents the observed answer to the request, e represents the random errors (here in both cases 1), and t is equal to the observed answer corrected for random errors.

However, as we have seen before this model is too simple because systematic errors also occur. In that case the variable we want to measure is not the same as the variable measured. This can be modeled by making a distinction between the variable we want to measure (f), the true score (t), and the response variable (y). This idea is presented in Figure 9.4.

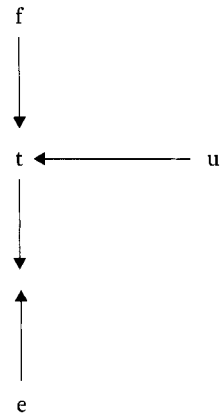


FIGURE 9.4: *The measurement model with random (e) and systematic (u) errors.*

To continue our example, if "political interest" is not measured by a direct request but by "the amount of time spent looking at the TV for political programs," then there is a difference between the variable we would like to measure and the variable really measured. We certainly expect that "political interest" affects "the time people spend watching TV for political programs," however, the relationship will not be perfect, especially because "the amount of leisure time" a person has will have an effect on "the time this person will watch TV." This could be modeled as is done in Figure 9.5.

If the relationship between f_2 and t is perfect so that $f_2 = t$, then this model will be reduced to the model presented in Figure 9.4. However, in Figure 9.5 there is a systematic error because the variable to be measured is measured indirectly using an indicator that also contains another component (leisure time).

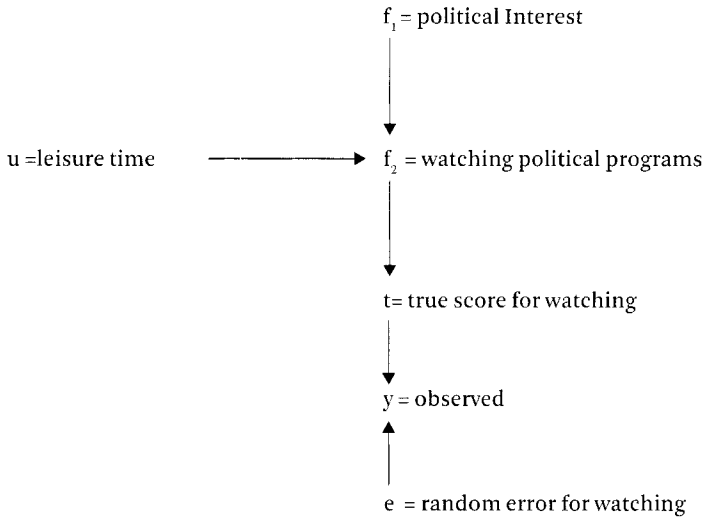


FIGURE 9.5: *A measurement for political interest with a systematic error (u) and a random error (e).*

There can also be another reason why f is not necessarily equal to the true score. We have seen that just varying the answer categories for measuring the opinion of respondents can also produce distinct and systematically different results. Therefore it appears that the method also has an effect on the response. For example, the direct request "How interested are you in politics?" So we can assume[1] that $f_1 = f_2$ in Figure 9.6. Even when we ignore this possible difference, we observe systematically different responses to the direct request depending on formulation choices of the direct request. We demonstrated that the reaction of respondents toward an 11-point scale can vary between respondents (with respect to the range of responses) and their individual reactions can be different for different response scales. The reaction of respondents to a specific method is called the *method factor*, which is modeled in Figure 9.6

[1] However it can be argued that the answers to the direct request are indeed not influenced by any variables other than "political interest," although the possibility that "a tendency to social desirable answers" might also play a role

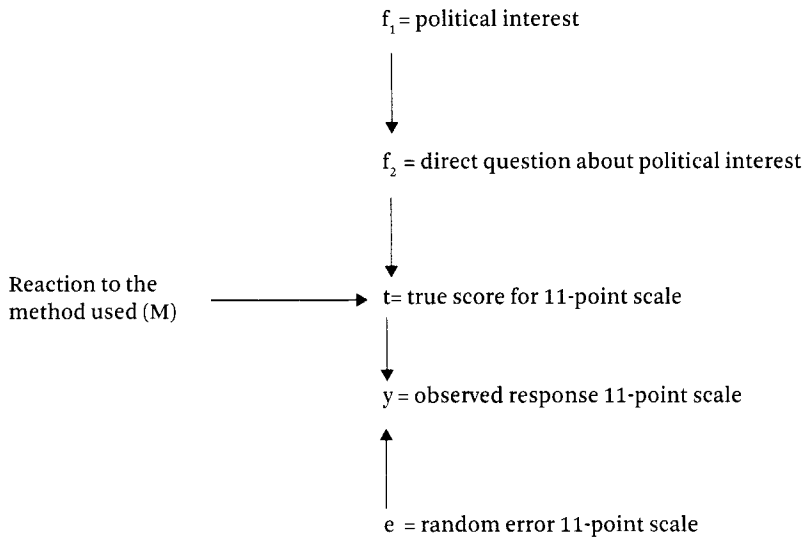


FIGURE 9.6: *A measurement for political interest with a systematic method factor (M) and a random error (e).*

If we assume again that the link between f_2 and f_1 is perfect ($f_2 = f_1$), then this model reduces to the model of Figure 9.4 with a specific component affecting the true score "the reaction to the method used."

Although a combination of the two reasons for differences between f_1 and t is rather common, in the next sections we will concentrate on the second kind of possible systematic error, because the first example represents a shift of concept that is discussed later in Part IV. Until then we will always assume that the different indicators do not contain systematic errors due to a conceptual shift and only contain errors due to the reaction of the method used. This modeling of measurement processes can help us specify quality criteria for measures and explain the differences in the data.

9.3 QUALITY CRITERIA FOR SURVEY MEASURES AND THEIR CONSEQUENCES

The first quality criterion for survey items is to have as little *item nonresponse* as possible. This is an obvious criterion because missing values have a disrupting effect on the analysis, which can lead to results that are not representative of the population of interest.

A second criterion is *bias*, which is defined as a systematic difference between real values of the variable of interest and the observed scores corrected for random measurement errors.[2] For objective variables real values can be obtained and thus the method that provides responses, corrected for random

[2] This simple definition serves for the purpose of this text. However, a precise definition is found in Groves (1989).

errors (true scores) closest to the real values is preferable. A typical example comes from voting research. After the elections the participation in the elections is known. This result can be compared with the result that is obtained by survey research performed using different methods. It is a well-known fact that using standard procedures, participation is overestimated. Therefore, a new method that does not overestimate the participation or produces a smaller bias is preferable to the standard procedures.

In the case of subjective variables, where the real values are not available, it is only possible to study different methods that generate different distributions of responses as we have done previously. If differences between two methods are observed, at least one method is biased; however, both can also be biased.

These two criteria have been given a lot of attention in split-ballot experiments [see Schuman and Presser (1981), for a summary]. Molenaar (1986) has studied the same criteria focusing on nonexperimental research (1986). In summary, these criteria give a description of the observed differences by nonresponse and differences by response distributions for differents methods.

There are also quality criteria that provide an explanation for the weak correlation between indicators that should measure the same variable and the differences in correlations between variables for different methods as we have seen in Table 9.7. To explain these observations, the concepts *reliability*, *validity*, and *method effect* need to be studied.

In order to do so, we extend the model of Figure 9.4 to two variables of interest, for example "satisfaction with the government" and the "satisfaction with the economy." The measurement model for two variables is presented in Figure 9.7.

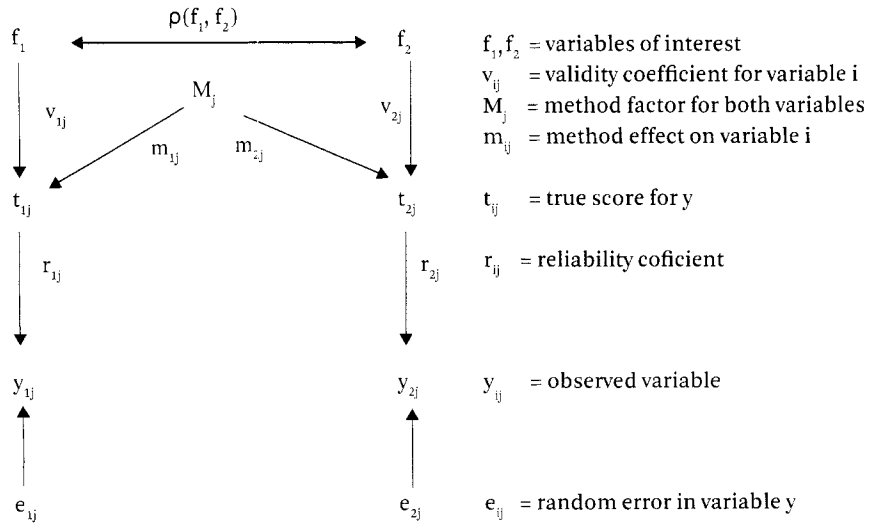


FIGURE 9.7: *The measurement model for two traits measured with the same method.*

In this model it is assumed that

- f_i is the trait factor i of interest measured by a direct question.
- y_{ij} is the observed variable (variable or trait i measured by method j).
- t_{ij} is the "true score" of the response variable y_{ij}.
- M_j is the method factor that represents a specific reaction of respondents to a method and therefore generates a systematic error.
- e_{ij} is the random measurement error term for y_{ij}.

The r_{ij} coefficients represent the standardized effects of the true scores on the observed scores. This effect is smaller if the random errors are larger. This coefficient is called the *reliability coefficient.*

The v_{ij} coefficients represent the standardized effects of the variables of interest on the true scores for the variables that are really measured. Therefore this coefficient is called the *validity coefficient.*

The m_{ij} coefficients represent the standardized effects of the method factor on the true scores, called the *method effect.* An increase in the method effect results in a decrease in validity and vice versa. It can be shown that for this model $m_{ij}^2 = 1 - v_{ij}^2$, and therefore the method effect is equal to the invalidity due to the method used.

Reliability is defined as the strength of the relationship between the observed response (y_{ij}) and the true score (t_{ij}), that is r_{ij}^2.

Validity is defined as the strength of the relationship between the variable of interest (f_1) and the true score (t_{ij}), that is v_{ij}^2.

The *systematic method effect* is the strength of the relationship between the method factor (M_j) and the true score (t_{ij}) resulting in m_{ij}^2.

The *total quality of a measure* is defined as the strength of the relationship between the observed variable and the variable on interest, that is $(r_{ij}v_{ij})^2$.

The *effect of the method on the correlations* is equal to $r_{1j}m_{1j}m_{2j}r_{2j}$.

The reason for employing these definitions and their criteria becomes evident after examining the effect of the characteristics of the measurement model on the correlations between observed variables.

Using the two theorems we have provided previously, it can be shown that the correlation between the observed variables $\rho(y_{1j},y_{2j})$ is equal to the joint effect of the variables that we want to measure (f_1 and f_2) plus the spurious correlation due to the method factor as demonstrated in formula (9.1):

$$\rho(y_{1j},y_{2j}) = r_{1j}v_{1j}\rho(f_1,f_2)v_{2j}r_{2j} + r_{1j}m_{1j}m_{2j}r_{2j} \qquad (9.1)$$

Note that r_{ij} and v_{ij}, which are always smaller than 1, will decrease the correlation (see first term) while the method effects, if they are not zero, can generate an increase in the correlation (see second term). This result suggests that it is possible that the low correlations for methods 1 and 3 in Table 9.7 are due to a lower reliability of method 1 and 3 compared to method 2. However, it is also

possible that the correlations of method 2 are higher because of greater systematic method effects of this method.

Using the specification of the model mentioned in Figure 9.7, the results presented in Table 9.8 are obtained for the variables of Table 9.7. How this result is obtained will be the topic of the next chapter. Now we will concentrate on the meaning of the results. Table 9.8 shows that method 2 has higher reliability coefficients than the other methods and that its method effects are intermediate.

Table 9.8: The quality criteria estimated from the ESS data of Table 9.7

	Validity coefficients			Method effects			Reliability coefficients
	F_1	F_2	F_3	M_1	M_2	M_3	
t_{11}	.93			.36			.79
t_{21}		.94		.35			.85
t_{32}			.95	.33			.81
t_{12}	.91				.41		.91
t_{22}		.92			.39		.94
t_{32}			.93		.38		.93
t_{13}	.85					.52	.82
t_{23}		.87				.50	.87
t_{33}			.88			.48	.84

If we know that the correlation between the first two traits was estimated at .69, it can be verified by substituting the values of the reliability, validity and method coefficients in equation (9.1), that such different observed correlations (Table 9.7) as .481 for method 1 and .642 method 2 can be obtained.

Equation (9.1) for method 1 gives

$$\rho_{y11,y12} = .79 \times .93 \times .69 \times .94 \times .85 + .79 \times .36 \times .35 \times .85 = .405 + .085 = .49$$

Equation (9.1) for method 2 gives

$$\rho_{y21,y22} = .91 \times .91 \times .69 \times .91 \times .91 + .91 \times .41 \times .39 \times .94 = .473 + .137 = .61$$

This result shows that the observed correlation between the same two variables was .12 higher for method 2 than for method 1 because its reliability was higher while the method effect was higher for method 2. So, with these quality estimates we can quite well explain the difference in correlations for the different

methods. However, both correlations were not very good estimates of the correlation between the two variables because of the random and systematic errors in the data. Our best estimate of the correlation between these two variables corrected for measurement error is .69. So, both correlations were incorrect. How we obtained the estimate of the relationship corrected for measurement errors will be discussed in the next chapter.

Our results show that the differences in the correlations obtained can almost entirely be explained by differences in data quality between the different measurement procedures. It also illustrates how important, for social science research, reliability and validity are as defined. Therefore, it is also important to know how these quality criteria can be estimated. However, let us now turn to some other commonly used criteria for data quality.

9.4 ALTERNATIVE CRITERIA FOR DATA QUALITY[3]

Out of the many possible criteria for data quality, we will discuss only the most common ones and indicate some problems associated with them.

9.4.1 Test-retest reliability

A very popular idea is that reliability can be determined by repeating the same observation twice as in the model of Figure 9.8

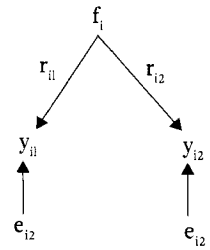


FIGURE 9.8: *The standard test-retest model.*

Here, f_i is the variable to be measured and y_{i_1} and y_{i_2} are the responses to the request used to measure this variable. This approach requires that the same method be used on two occasions. If the model holds true then the correlation between the two variables can be due only to the product of the two reliability coefficients of the two measures:

$$\rho_{yi1,yi2} = r_{i1} \cdot r_{i2}$$

[3] This section gives a wider perspective on the way reliability and validity can be defined; however, it is not essential for understanding the approach discussed in this book

But since the same measure is used twice, we can assume that $r_{i1} = r_{i2}$ and then it follows that the reliability = $r_{i1}^2 = r_{i2}^2 = \rho_{yi1,yi2}$. In this case the reliability of the measure is equal to the test-retest correlation.

However, the above representation is too simple and it is better to start with the model shown in Figure 9.9.

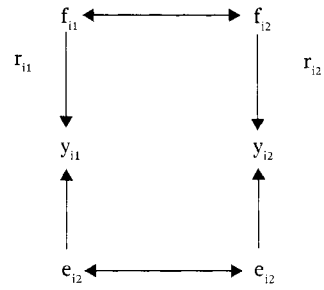


FIGURE 9.9: *A more realistic test-retest model.*

The difference with the previous model is that a distinction is made between the latent variable for the first and second measures, accounting for a change that might have occurred between conducting the two observations. Additionally, the possibility is left open that respondents remember their first answers, indicated by a correlation between the error terms. In order to arrive from this model to the earlier model, the following assumptions were made:

1. No change in opinion between the first and the second measurements
2. No memory effects
3. No method effects
4. Equal reliability for the different measures of the same trait

This approach is unrealistic because it assumes that the measurement procedure can be repeated in exactly the same way (assumption 4).

Furthermore, if the time between the repetitions is too short, we can expect a memory effect (assumption 2) and if the time is too long, the opinion may be changed (assumption 1). Finally, possible method effects cannot be detected, while they may play an important role (assumption 3).

Therefore, this approach is not an accurate representation of reality. Although many people think that it is a robust procedure, it is based on a number of unattainable assumptions and a less restricted approach is needed.

9.4.2 The quasi-simplex approach

The above specified approach can be made more manageable by using three observations instead of two. This approach has been suggested by Heise (1969),

improved by Wiley and Wiley (1970) and used by Alwin and Krosnick (1991) to evaluate measurement instruments for survey research. Its advantage of it is that it is no longer necessary to assume that no change has occurred, and it is suggested that the memory effect can be avoided by making the time gap between the observations so long that a memory effect can no longer be expected. Figure 9.10 displays the suggested model

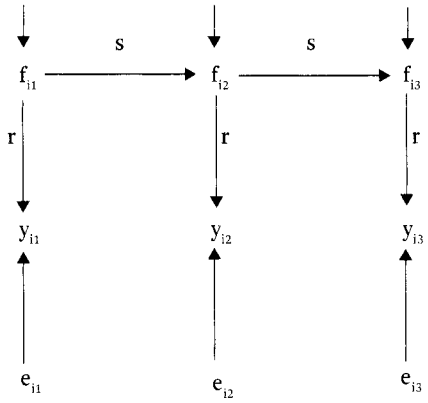


FIGURE 9.10: *The quasi simplex model for three repeated observations.*

In Figure 9.10 "s" is the stability coefficient and "r" is the reliability coefficient. This approach has two major problems. First, it assumes that it is not possible that considerations that are associated with the variable of interest are forgotten at time 2 but return at time 3. This would suppose that there is an effect of f_{i1} on f_{i3} that is not possible for technical reasons. However, because of these effects, wrong estimates of the quality of the measures will be obtained as discussed by Coenders et al. (1999).

The second problem is that any temporary component in the variables that is not present at the next occasion will be treated as error, while it might be a substantive part of the latent variable at a given point in time. For example, if we ask about "life satisfaction" and the respondent is in a bad mood, that person's score will be lower than if the same respondent is in a good mood on a different occasion. The mood component is a real part of the satisfaction variable, but because the mood changes rapidly, this component will end up in the error term. Therefore, the error term increases and the reliability decreases: not because of lack of data quality but because of the instability of a component within the variable of interest. For further discussion of this point, we refer the reader to Van der Veld (2006). However, this would not occur if the measures were conducted quickly in the same survey, but then memory effect might emerge again. For these reasons this approach is not preferable for defining the reliability coefficient.

9.4.3 Correlations with other variables

In order to evaluate the validity of different measures for the same variable, it has been suggested to use the correlation with other variables, that are known to correlate with the variable of interest. The measure with the highest correlation is then the best estimate. Following this line of reasoning this approach is modeled in Figure 9.11.

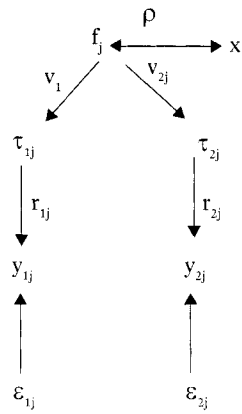


FIGURE 9.11: *A standard model to evaluate validity.*

In this Figure ρ is the correlation between the variable of interest and the external criterion variable (x). The other coefficients have their previously discussed meanings.

From this model it follows that:

$$\rho_{y1j,xi} = r_{1j}\, v_{1j}\, \rho \quad \text{and} \quad \rho_{y2j,xi} = r_{2j}\, v_{2j} \rho$$

This demonstrates that correlations can be different because of differences in validity, differences in reliability, or both. It also suggests that these correlations are not the proper criteria to evaluate the validity of measures. The validity of a measure should be evaluated by comparing the validity coefficients that we have presented in Section 9.3, in order to avoid confusion between reliability and validity, as is the case when using the correlation with a criterion variable.

9.5 SUMMARY AND DISCUSSION

In this chapter we have demonstrated that the influence of different choices in the development of a survey item can have a significant effect on the item nonresponse, the distribution of the variables and the correlation between different traits measured by different methods. These results led us to the conclusion that the first two quality criteria are:

1. Item nonresponse
2. Bias in the response

Furthermore, we have shown that the differences in the correlations between the variables for the different methods can be explained by the size of the random errors and systematic errors or the reliability and the validity and method effects, which were defined as follows: reliability, as the strength of the relationship between the observed variable and the true score; and validity, as the strength of the relationship between the true score and the latent trait of interest.

Other measures that are often used and their critiques were also elaborated. Given that we have clearly defined the quality criteria in the next two chapters we will discuss how these quality criteria can be and have been estimated in practice.

EXERCISES

1. Asking people for information about their salary is a problem because many people refuse to answer the request. Therefore, alternative procedures are employed and compared with factual information collected from the employers.
   Below we provide two requests used in practice and the results that were obtained are compared with the factual information.
   Q1: *What is your net monthly income?*
   Q2: *Could you tell me to what category your net monthly income belongs?*

In euros	Factual info	Q1	Q2
< 1000	5%	9%	5%
1000–1500	10%	12%	11%
1500–2000	30%	35%	32%
2000–2500	30%	32%	33%
2500–3000	10%	8%	11%
3000–3500	5%	2%	4%
3500–4000	4%	1%	2%
4000–4500	3%	1%	1%
> 4500	3%	0%	1%
100%=	1000	700	850

   a. What, do you think, is the best measure for income?
   b. What was the criterion for answering "a"?
2. To see the effect of reliability on the observed correlation, we evaluate the following cases of the model of Figure 9.7:
   If the correlation between the variables corrected for measurement error $\rho(f1,f2) = .9$, the validity = 1, and the method effect = 0 what is the correlation between the observed variables in the following cases?

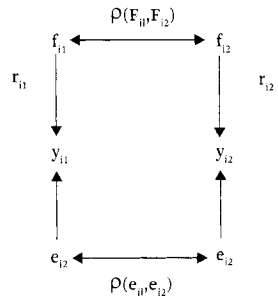
Reliability	Reliability	Observed correlation
y_{11}	y_{12}	$r_{y11,y12}$
1.0	1.0	
.9	.9	
.8	.8	
.7	.7	
.6	.6	

[To answer these questions, use equation (9.1)]

3. The effect of the validity and method effect on the correlation between the observed variables is studied while the correlation between the variables corrected for measurement error = .4 and the reliability = .8 for both measures.
   What is the correlation between the observed variables in the following cases?

Validity	Validity	Method effect	Method effect	Correlation
y_{11}	y_{12}	y_{11}	y_{12}	$r_{y11,y12}$
1.0	1.0	0.0	0.0	
.9	.9	.43	.43	
.8	.8	.6	.6	
.7	.7	.71	.71	

4. Is there any reason to assume that one or more of the requests of your own questionnaire are without random and/or systematic measurement error? If so why?
5. Can you think of a measure that is without measurement error?
6. For the test-retest approach we have specified the following model:[4]



   a. Express the correlation between the observed variables in the parameters of the model.
   b. Show that the specified restrictions actually lead to the simple test-retest model.

[4] This question requires reading of Section 9.4

APPENDIX 9.1: THE SPECIFICATION OF STRUCTURAL EQUATION MODELS

Structural equation models (SEMs) have been discussed extensively in the literature. For a simple introduction we refer the reader to Saris and Stronkhorst (1984), and for a more elaborate text we suggest the text of Bollen (1989). Therefore our introduction to this topic will be very brief and we mention only those aspects that are relevant for the models discussed in this and the next chapter.

The simplest case is a model with only two variables: a cause X and an effect variable Y. Assuming a linear relationship, we can formulate

$$Y = a + bX + u \tag{9A.1}$$

where "u" represent all variables affecting Y that are not explicitly mentioned. If we assume that the mean of u is zero, then the mean value of Y = a, if X=0. This coefficient is called the *intercept* of the equation. The mean value of Y will increase by "b" for any increase of X of 1 unit on its scale. This effect "b" of X, which is called the *slope* or the *unstandardized regression coefficient*, is always the same. Therefore the relationship between the two variables is linear, as presented in Figure 9A.1

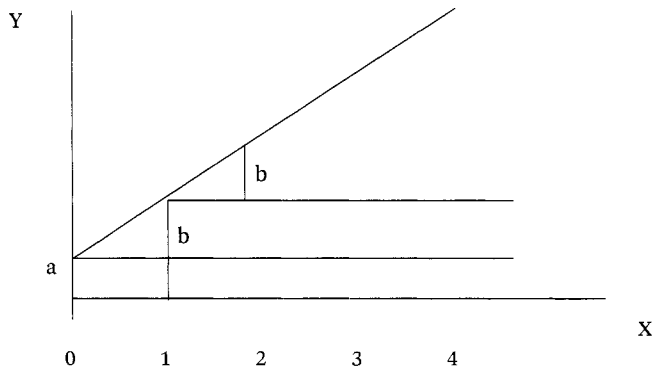


FIGURE 9A.1: *The linear regression model.*

From equation (9A.1) it follows for the mean of Y indicated by $\mu(Y)$ that

$$\mu(Y) = \frac{1}{n}\Sigma(Y) = a + b\frac{1}{n}\Sigma(X) + \frac{1}{n}\Sigma(u) \tag{9A.2}$$

where the summation is done over the scores of all people. Since the mean of "u" is zero, we derive that:

$$\mu(Y) = a + b \cdot \mu(X) \tag{9A.3}$$

From (9A.1) and (9A.3) it follows that:

$$(Y - \mu(Y)) = b\,(X-\mu(X)) + u \tag{9A.4}$$

In this case the scores of the variables are transformed to *deviation scores* because they are a numerical expression of the deviation from the means. It follows that the intercept of the equation is zero. This means that the line goes through the zero point or that Y= 0 if X=0. We also see that the regression coefficient remains the same.

It is very common and useful in our application to standardize the variables. This can be done by dividing each of the variables, expressed in deviation scores, by their standard deviation (σ(i)) and from (9A.4) we get formula (9A.5)

$$\frac{Y - \mu(Y)}{\sigma(Y)} = b\frac{\sigma(x)}{\sigma(y)}\quad\frac{X-\mu(X)}{\sigma(X)} + \frac{u}{\sigma(Y)} \tag{9A.5}$$

It can easily be verified that if (9A.4) is true, then (9A.5) is also true. The last equation can be rewritten as

$$y = \beta x + u \tag{9A.6}$$

Now y and x are standardized variables, while the effect of x on y is β, and the relationship with the previous coefficient is,

$$\beta = b\frac{\sigma(X)}{\sigma(Y)} \tag{9A.7}$$

This effect should be interpreted as the effect on y of an increase of x with 1 standard deviation and is therefore called the *standardized regression coefficient*. In order to indicate the *strength of the effect* of x on y, it is squared:

$$\beta^2 = \text{strength of the relationship} \tag{9A.8}$$

Note that β=0 if and only if b=0 because the standard deviations are always larger than zero. The standardized coefficient as well as the unstandardized coefficient indicate whether a variable has an effect on another variable.

This model can be extended by introducing more causal variables. In case we assume that the effects of the different variables are additive, the model (9A.9) becomes a model of multiple regression with standardized variables:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + u \tag{9A.9}$$

In this case the strength of the effects of the different variables on the effect variables can be compared by comparing the β_i coefficients.

Models can be further extended by introducing more effect variables. In that case for each effect variable a separate equation like (9A.9) should be formulated, which includes in the equation only those variables that supposedly have a direct causal effect on the effect variable. Such a set of equations forms a causal model, and these models are compatible with the graphical models used in this text. Theorems 1 and 2 discussed above can be proved using the causal models in algebraic form that were introduced in this appendix. For more detail, we refer the reader to Saris and Stronkhorst (1984) and Bollen (1989).

CHAPTER 10

Estimation of reliability, validity, and method effects

In the last chapter we discussed several criteria to consider while thinking about quality of survey measures. Two criteria, item nonresponse and bias, do not require much further discussion. If different methods have been used for the same trait, item nonresponse can be observed directly from collected data. The same holds true for bias if factual information is available. However, this is not the case if an estimate of the relative bias of measures has to be derived by comparing distributions of responses with each other. Molenaar (1986) has made useful suggestions for measures of relative bias.

More complicated is the estimation of the quality criteria of reliability, validity and method effect. Therefore this chapter and the next will concentrate on their estimation. In order to discuss the estimation of reliability and validity, we have to introduce the basic idea behind estimation of coefficients of measurement models. So far it is not so evident that the effect of an unmeasured variable on a measured variable, let alone the effect of an unmeasured variable on another unmeasured variable, can be estimated.

Here we start by addressing the problem of identifying the parameters of models with unmeasured variables. Next we will discuss the estimation of the parameters. Only after we have introduced these basic principles, will we concentrate on the estimation of the reliability and validity, demonstrating the kind of designs that have been used to estimate them, as was defined in the last chapter.

10.1 Identification of the parameters of a measurement model

In the last chapter we introduced the formulation of causal models and two theorems to derive the relationship between the correlations of pairs of variables and the parameters of a model. Figure 10.1 represents a causal model for "Satisfaction with the economy" where f_1 is a general measure of "satisfaction with the economy" while y_{11} is a measure of "satisfaction" using a bipolar 4-point scale. Measure y_{12} is a bipolar measure of satisfaction on an 11-point scale and y_{13} is a measure of satisfaction on a unipolar 4-point scale. The formulation of these measures has been presented in Table 9.1.

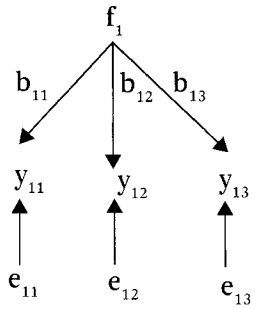


FIGURE 10.1: *A simple measurement model assuming only random measurement errors.*

In this model f_1 is the variable that we would like to measure but that cannot be directly observed, except by asking questions. The variables y_{11} through y_{13} represent answers to these different questions. For sake of simplicity, we assume that there are only random measurement errors (or $t_1=f_1$). Using the two theorems mentioned in the last chapter, the following relationships of the correlations between the variables and the parameters of the model can be derived:

$$\rho(y_{11},f_1) = b_{11} \tag{10.1}$$

$$\rho(y_{12},f_1) = b_{12} \tag{10.2}$$

$$\rho(y_{13},f_1) = b_{13} \tag{10.3}$$

$$\rho(y_{12},y_{11}) = b_{12}b_{11} \tag{10.4}$$

$$\rho(y_{13},y_{11}) = b_{13}b_{11} \tag{10.5}$$

$$\rho(y_{13},y_{12}) = b_{12}b_{13} \tag{10.6}$$

If the correlations between these variables were known, the effects of the general judgment (f_1) on the specific observed answers would be easily determined on the basis of the first three relationships. However, the problem with measurement models is that the general variable (f_1) is not a measured variable and by definition its correlations with the observed variables are also unknown. So, the effects have to be estimated from the relationships between the observed variables from y_{11} to y_{13} [equations (10.4) – (10.6)]. Formally it can be shown that this is possible. From (10.4) and (10.5) we can derive

$$\frac{\rho(y_{12},y_{11})}{\rho(y_{13},y_{11})} = \frac{b_{12}b_{11}}{b_{13}b_{11}} = \frac{b_{12}}{b_{13}}$$

and therefore

$$b_{12} = b_{13}\frac{\rho(y_{12},y_{11})}{\rho(y_{13},y_{11})}$$

If we substitute this result in (10.6) we get equation

$$\rho(y_{13},y_{12}) = b_{13}\frac{\rho(y_{12},y_{11})}{\rho(y_{13},y_{11})}b_{13} \qquad (10.7)$$

$$b_{13}^2 = \frac{\rho(y_{13},y_{11})\,\rho(y_{13},y_{12})}{\rho(y_{12},y_{11})}$$

From this it can be seen that if these three correlations are known, the effect b_{13} is also known. Also, if b_{13} is known, the other effects can be obtained from equations (10.5) and (10.6).

The proof shows that the estimate of an effect of an unmeasured variable on measured variables can be obtained from the correlations between the observed variables. Let us check this conclusion for the given example. In Table 9.7 the correlations for the three observed variables can be found. Assuming that the values mentioned are the correct values for the population correlations, we can derive the following:

$$\rho(y_{12},y_{11}) = -.626 = b_{12}b_{11}$$
$$\rho(y_{13},y_{11}) = -.502 = b_{13}b_{11}$$
$$\rho(y_{13},y_{12}) = .584 = b_{12}b_{13}$$

Applying the above indicated procedure, we first get that $b_{13}^2 = -.502 \times .584/-.626 = .468$, making $b_{13} = .684$ or[1] $-.684$, it follows from equation (10.6) that $b_{12} = .584/.684 = .853$ and from (10.5) that $b_{11} = -.502/.689 = -.734$.

The results show that "satisfaction with the economy" (f_1) has the strongest relationship with the observed scores of method 2 (y_{12}). Therefore this measure seems preferable, because it contains the smallest amount of random measurement errors. We also see that the effect for method 1 is negative. This is because the scale of the observed variable goes from positive to negative, while for the other two variables the scale goes in the opposite direction.

---

[1] That a positive or negative value is possible is not just a numeric result but also a logical one, given that the scale of the latent variable is not fixed and can go from low to high or from high to low. Therefore, the sign of the effect of this variable can be positive or negative. After the sign for one effect has been chosen, all others are also determined.

This example demonstrates that it is possible to estimate the size of the effect of an unmeasured variable on a measured variable. The question is, however, whether this is always the case and if not, when it is not the case. It can easily be verified that the necessary condition is that there should be at least as many correlations as unknown parameters. If there are only two observed variables, there is only one correlation; however, two coefficients need to be estimated, which is impossible.

Even if the necessary condition is fulfilled, the values of the parameters cannot always be obtained. The sufficient condition for "*identification*" is difficult to formulate. For a more complete analysis of sufficient conditions, we refer the reader to the literature on structural equation modeling (Bollen 1989). A practical approach is that one uses programs for the estimation of structural equation models to determine whether the parameters of a model can actually be estimated. The programs will indicate whether the model is not identified and even indicate which parameter cannot be uniquely estimated.[2]

A requirement for the quality of the estimates is that the model be correctly specified, because the relationships that are derived for the correlations and their parameters are based on the assumption that it is so. While the estimation of the values of the parameters is based on these relationships, the correctness of the specification of a model is determined by a test. Such a test for structural equations models requires that there be more correlations than parameters to be estimated. The difference between the number of correlations and the number of parameters is called the *degree of freedom* (df). So the necessary condition for any testing is that df > 0.

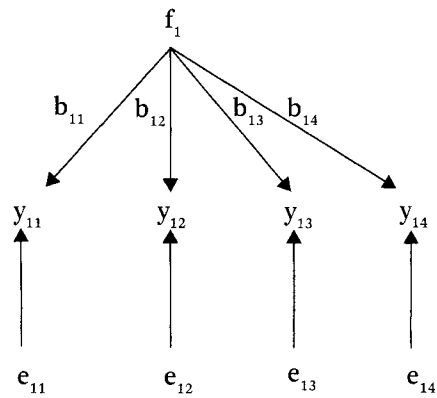


FIGURE 10.2: *A simple measurement model with one unobserved and four observed variables assuming only random measurement errors.*

[2] The programs mention the first parameter that cannot be estimated, but there are often more possible candidates. Therefore one should use such a suggestion only as one of the possible suggestions.

In the previous example (Figure 10.1) a test is not possible because there are only three correlations for three parameters that have to be estimated, having df = 0. There is a perfect solution presented, but there is no information for a test because all information has been used to determine the values of the parameters. Adding one more measure for "satisfaction with the economy" we have the model presented in Figure 10.2.

In Figure 10.2 there are three more correlations between the observed variables and only one extra effect parameter that needs to be estimated. In this case df=2 and a test is possible. Let us illustrate it again by extending the above given example: Suppose for y_{14} that we have found the following correlations with the other variables

$\rho(y_{14},y_{11}) = -.474$

$\rho(y_{14},y_{12}) = .551$

$\rho(y_{14},y_{13}) = .442$

Now we have six correlations of observed variables and only four effects to estimate as can be seen from the following sets of equations. The first three are the same as we have used before, but we can add three more for the three extra correlations:

$$\rho(y_{12},y_{11}) = -.626 = b_{12}b_{11} \tag{10.8}$$

$$\rho(y_{13},y_{11}) = -.502 = b_{13}b_{11} \tag{10.9}$$

$$\rho(y_{13},y_{12}) = .584 = b_{12}b_{13} \tag{10.10}$$

$$\rho(y_{14},y_{11}) = -.474 = b_{14}b_{11} \tag{10.11}$$

$$\rho(y_{14},y_{12}) = .551 = b_{14}b_{12} \tag{10.12}$$

$$\rho(y_{14},y_{13}) = .442 = b_{14}b_{13} \tag{10.13}$$

From the first three equations (10.8)–(10.10) b_{11}, b_{12} and b_{13} can be estimated as we did before, while from the last three equations (10.11)–(10.13) only one is needed to estimate b_{14}. So there are indeed 2 degrees of freedom. Keeping in mind that b_{11}= -.734, it follows from equation (10.11) that b_{14} = -.474/-.734 = .646. Now all parameters are determined but we have two equations left that have not been used for determining the values of the parameters.

Equations (10.12) and (10.13) can be used for a test because now all coefficients are known and we can control whether the correlations in these two equations can be reproduced by the values obtained for the effect parameters. For these two equations we expect that:

$$\rho(y_{14},y_{12}) = b_{14}b_{12} = .853 \times .646 = .551$$

$$\rho(y_{14},y_{13}) = b_{14}b_{13} = .684 \times .646 = .442$$

In this constructed example the test would indicate that the model is correct because the effect parameters produce exactly the values of the two correlations that were not used for estimation. If, however, the correlation between the variables y_{14} and y_{13} had been .560, then one would think that there may be something wrong with the model, because we would expect, on the basis of the estimated values of the coefficients, the correlation to be .442 while in reality it is much larger (.560).

The differences between the correlations of the observed variables and their predicted values are called the *residuals*. It will be clear that the model has to be rejected if the residuals are too large. In the next section we will discuss this. However, keep in mind that such a test is possible only if df > 0. If df=0 the residuals are normally[3] zero by definition, because there is only one perfect solution.

10.2 ESTIMATION OF PARAMETERS OF MODELS WITH UNMEASURED VARIABLES

So far we have discussed only under what conditions the parameters of measurement models are identifiable and the fact that the models can be tested. We have not yet discussed how the parameters can be estimated. The procedure we have used so far to show that the parameters can be identified is not satisfactory for estimation for two reasons. The first reason is that the selection of the equations (10.8) – (10.11) to determine the values of the parameters was arbitrary. We could have used four other equations and in general, different estimates of the coefficients will be derived for each of the chosen sets of four equations, making this procedure unacceptable.

A second reason is that the correlations in the population are not known. Only estimates of these correlations can be obtained using a sample from the population, which will be denoted with "r_{ij}" in order to indicate very clearly that they represent *sample correlations*. The relationship between the correlations and the effect parameters holds perfectly only for the population correlations and not for the sample correlations. Here we simply cannot use the previously mentioned equations (10.8) – (10.13) to estimate coefficients.

There are several general principles to derive estimators for structural equations models. We will discuss the *unweighted least squares* (ULS) procedure and the *weighted least squares* (WLS) procedure. Both procedures are based on the residuals between the sample correlations and the expected values of these correlations that are a function of the parameters $f_{ij}(\mathbf{p})$. In this formulation $\mathbf{p}$

[3] There is the possibility that the necessary condition for identification is satisfied, but that the model is not identified anyway and in that case the residuals can be larger than zero.

represents the vector of parameters of the model and f_{ij}, the specific function that gives the link between the population correlations and the parameters for the variables i and j.

The ULS procedure suggests looking for the parameter values that minimize the unweighted sum of squared residuals[4]

$$F_{ULS} = \Sigma(r_{ij} - f_{ij}(\mathbf{p}))^2 \tag{10.14}$$

where r_{ij} is the correlation in the sample between variables y_i and y_j. The summation is computed on all unique elements of the correlation matrix.

The WLS procedure suggests looking for the parameter values that minimize the weighted sum of squared residuals

$$F_{WLS} = \Sigma w_{ij}(r_{ij} - f_{ij}(\mathbf{p}))^2 \tag{10.15}$$

where w_{ij} is the weight for the term with the residual for the correlation r_{ij}. The summation is also computed on all unique elements of the correlation matrix.

The difference between the two methods is only that the ULS procedure gives all correlations an equal weight while the WLS procedure varies the weights for the different correlations. These weights can be chosen in different ways. The most commonly used procedure is the maximum likelihood (ML) estimator, which can be seen as a WLS estimator with specific values for the weights (w_{ij}). The ML estimator provides standard errors for the parameters and a test statistic for the fit of the model. However, it was developed under the assumption that the observed variables have a multivariate normal distribution. More recently it was demonstrated that the ML estimator is robust under very general conditions (Satorra 1990, 1992). There are also other estimators using different weighting techniques; for further information about this topic we recommend the books on SEM by Bollen (1989) and Kaplan (2000).

Also, how the values of the parameters are determined is an issue that is beyond the scope of this text. However, the ULS procedure can easily be illustrated by the example with the four observed "satisfaction" variables discussed earlier. In this case the minimum of the ULS criterion has to be found for the following function:

$$\begin{aligned} F_{ULS} = {} & (-.626 - b_{12}b_{11})^2 + (-.502 - b_{13}b_{11})^2 + (.584 - b_{12}b_{13})^2 + (-.474 - b_{14}b_{11})^2 + \\ & (.551 - b_{14}b_{12})^2 + (.442 - b_{14}b_{13})^2 + (1 - (b_{11}^2 + \text{variance}(e_{11})))^2 + \\ & (1 - (b_{12}^2 + \text{variance}(e_{12})))^2 + (1 - (b_{13}^2 + \text{variance}(e_{13})))^2 + \\ & (1 - (b_{14}^2 + \text{variance}(e_{14})))^2 \end{aligned} \tag{10.16}$$

[4] For simplicity sake we specified the functions to be minimized for the correlation coefficients; however, it is recommended to use the covariance matrix as data. For details on this issue we recommend the work of Bollen (1989).

Each term on the right hand side of (10.16) is a residual. The first six terms come directly from the equations (10.8)–(10.13). The first value is always the observed sample correlation, and the second term of the function of the parameters is according to the model equal to the population correlation. The last four terms are introduced to obtain an estimate of the variance of the measurement errors. They are based on the diagonal elements of the correlation matrix representing the variances of the standardized variables, which are by definition equal to 1. The variances should be equal to the explained variance (b_{ij}^2) plus the unexplained variance or the variance of e_{ij}.

To obtain the optimal values of the parameters, an iterative procedure is used that looks for those values that minimize the function as a whole. In the case of a perfect fit we get exactly the same solution as we have found earlier. This is, however, not the case anymore if the sample correlation between y_{13} and y_{14} or $r(y_{13},y_{14}) = .560$ as we suggested before. Using the program LISREL to estimate the parameters and applying the ULS estimation method, we find the result presented in Table 10.1.

The ULS method tries to find an optimal solution for all cells of the correlation matrix, while employing the hand calculation, only the first four equations have been used. When using ULS the *root mean squared residuals* (RMSR) are considerably smaller and can be used as a measure for the fit of the model to the data. However, it is difficult to determine when the residuals are too large. Some people suggest to reject the model if the RMSR is larger than .1. If that criterion would have been used, this model would not be rejected. An alternative is to look for possible misspecifications in the model. The programs that can estimate these models provide estimates of the *expected parameter change* or EPC and provide also a test statistic for these estimates, the *modification index*. For this issue we refer the reader to the literature (Saris et al. 1987).

Table 10.1: The estimated values of the parameters obtained by a hand calculation compared to the ULS procedure of the LISREL program

Parameter	values of the parameters obtained	
	By hand equations (10.8-10.11)	By LISREL/ ULS method
b_{11}	-.734	-.713
b_{12}.	.853	.820
b_{13}	.684	.736
b_{14}	.646	.698
Fit of the model : RMSR	.037	.024

We could discuss about estimation and testing procedures of SEM much more, but we refer the reader to the literature mentioned in this section.

10.3 ESTIMATING RELIABILITY, VALIDITY, AND METHOD EFFECTS

Now that the general principles of identification, estimation, and testing of a measurement model have been introduced, we will discuss the estimation of

the three criteria: reliability, validity and method effects. Figure 10.3 presents the same model for two traits from the last chapter.

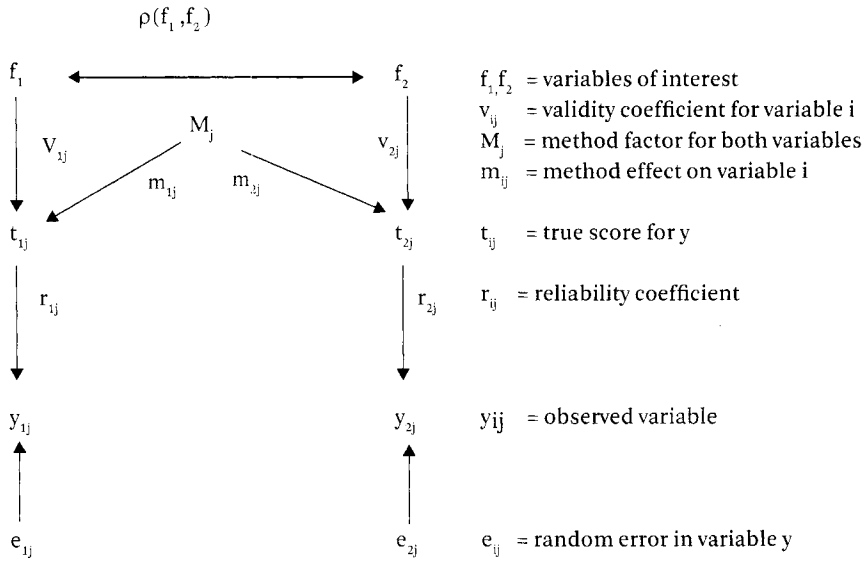


FIGURE 10.3: *The measurement model for two traits measured with the same method.*

This model differs from the models presented in Figures 10.1 and 10.2 in that method- specific systematic errors are also introduced. This makes the model more realistic, while not changing the general approach.

Using the two theorems presented in Chapter 9, it was demonstrated that the correlation between the observed variables, $\rho(y_{1j},y_{2j})$, is equal to the joint effect of the variables that we want to measure (f_1 and f_2) plus the spurious correlation due to the method effects, as follows:

$$\rho(y_{1j},y_{2j}) = r_{1j}v_{1j}\rho(f_1,f_2)v_{2j}r_{2j} + r_{1j}m_{1j}m_{2j}r_{2j} \qquad (10.17)$$

We have shown above that the reliability, validity, and method effects are the parameters of this model. The issue within this model is that there are two reliability coefficients, two validity coefficients, two method effects and one correlation between the two latent traits, leaving us with seven unknown parameters, while only one correlation can be obtained from the data. It is impossible to estimate these seven parameters from just one correlation. Therefore, in the following section we will discuss more complex designs to estimate the parameters.

Campbell and Fiske (1959) suggested using multiple traits and multiple methods (MTMM). The classical MTMM approach recommends the use of a minimum of three traits that are measured with three different methods

leading to nine different observed variables. The example of Table 10.2 was discussed in Chapter 9 Table 9.1.

Table 10.2: The classic MTMM design used in the ESS pilot study

The three traits were presented by the following three requests:

- *On the whole, how satisfied are you with the present state of the economy in Britain?*
- *Now think about the national government. How satisfied are you with the way it is doing its job ?*
- *On the whole, how satisfied are you with the way democracy works in Britain?*

The three methods are specified by the following response scales:
(1) Not at all satisfied; (2) Satisfied; (3) Rather satisfied; (4) Very satisfied

Very dissatisfied *Very satisfied*

0 1 2 3 4 5 6 7 8 9 10

1 not at all satisfied 2 satisfied 3 rather satisfied 4 very satisfied

Collecting data using this MTMM design, data for nine variables are obtained and from that data a correlation matrix of 9×9 is obtained. The model formulated to estimate the reliability, validity, and method effects is an extension of the model presented in Figure 10.3. Figure 10.4 illustrates the relationships between the true scores and their general factors of interest. Figure 10.4 shows that each trait (f_i) is measured in three ways. It is assumed that the traits are correlated but that the method factors (M_1, M_2, M_3) are not correlated. To reduce the complexity of the figure, it is not indicated that for each true score there is an observed response variable that is affected by the true score and a random error as was previously introduced in the model in Figure 10.3. However, these relationships, although not made explicit, are implied.

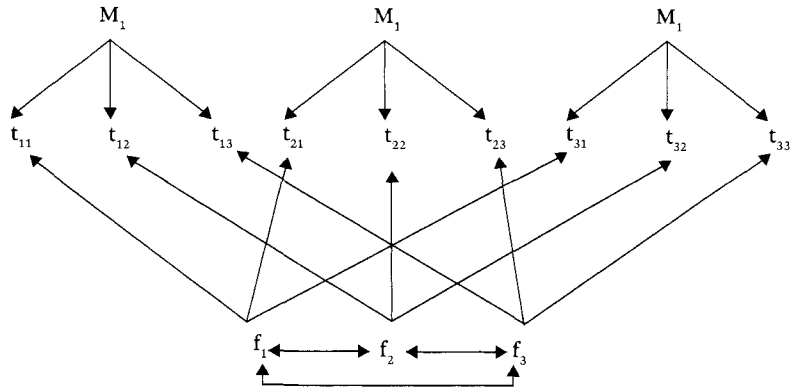


FIGURE 10.4: *MTMM model illustrating the relationships between the true scores and the factors of interest.*

It is normally assumed that the correlations between the factors and the error terms are zero, but there is debate about the actual specification of the correlations between the different factors. Some researchers allow for all possible correlations between the factors, while mentioning estimation problems[5] (Kenny and Kashy 1992; Marsh and Bailey 1991; Eid 2000). Andrews (1984) and Saris (1990) suggest that the trait factors can be allowed to correlate, but should be uncorrelated with the method factors, while the method factors themselves are uncorrelated. Using this latter specification, combined with the assumption of equal method effects for each method, almost no estimation problems occur in the analysis. This was demonstrated by Corten et al. (2002) in a study in which 79 MTMM experiments were reanalyzed.

The MTMM design of 3 traits and 3 methods generates 45 correlations and variances. In turn, these 45 pieces of information provide sufficient information to estimate 9 reliability and 9 validity coefficients, 3 method effect coefficients and 3 correlations between the traits. In total there are 24 parameters to be estimated. This leaves 45 - 24 = 21 degrees of freedom, meaning that the necessary condition for identification is fulfilled. It also can be shown that the sufficient condition for identification is satisfied and given that df=21 a test of the model is possible.

Table 10.3 presents again the correlations that we derived between the 9 measures obtained from a sample of 481 people in the British population. Using the specifications of the model indicated above and the ML estimator to estimate the quality indicators, the results presented in Table 10.4 are obtained[6]. (The input for the LISREL program that estimates the parameters of the model is presented in Appendix 10.1.)

The results in Table 10.4 indicate that the fit of the model is rather good. Therefore the model does not have to be rejected and the estimated values of the parameters are probably a good approximation of the true values of the parameters. The parameter values point to method 2 having the highest reliability for these traits. With respect to validity, the first two methods have the highest scores and are approximately equal. When considering all estimates method 2 is preferable to the other methods.

---

[5] This approach lends itself to non-convergence in the iterative estimation procedure or improper solutions such as negative variances.

[6] In this case the ML estimator is used. The estimation is done using the covariance matrix as the input matrix and not the correlation matrix (see Appendix 10.1). Thereafter, the estimates are standardized to obtain the requested coefficients. A result of this is that the standardized method effects are not exactly equal to each other.

Table 10.3: The correlations between the 9 variables of the MTMM experiment with respect to satisfaction with political outcomes

	Method 1			Method 2			Method 3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
Method 1									
Q1.	1.00								
Q2	.481	1.00							
Q3	.373	.552	1.00						
Method 2									
Q1	-.626	-.422	-.410	1.00					
Q2	-.429	-.663	-.532	.642	1.00				
Q3	-.453	-.495	-.669	.612	.693	1.00			
Method 3									
Q1	-.502	-.374	-.332	.584	.436	.438	1.00		
Q2	-.370	-.608	-.399	.429	.653	.466	.556	1.00	
Q3	-.336	-.406	-.566	.406	471	.638	.514	.558	1.00
Means	2.42	2.71	2.45	5.26	4.37	5.13	2.01	1.75	2.01
Standard deviation	.77	.76	.84	2.29	2.37	2.44	.72	.71	.77

Table 10.4: Standardized estimates of the MTMM model specified for the FSS data of Table 10.3

	Validity coefficients			Method effects			Reliability coefficients
	F_1	F_2	F_3	M_1	M_2	M_3	
t_{11}	.93			.36			.79
t_{21}		.94		.35			.85
t_{31}			.95	.33			.81
t_{12}	.91				.41		.91
t_{22}		.92			.39		.94
t_{32}			.93		.38		.93
t_{13}	.85					.52	.82
t_{23}		.87				.50	.87
t_{33}			.88			.48	.84

Note that the validity and the method effects do not have to be evaluated separately because they complement each other, as was mentioned previously: $v_{ij}^2 = 1 - m_{ij}^2$. With this example we have shown how the MTMM approach can be used to evaluate the quality of several survey items with respect to validity and reliability.

10.4 Conclusion and Discussion

The reliability, validity coefficients, and the method effects are defined as parameters of a measurement model and indicate the effects of unobserved variables on observed variables or even on unobserved variables. This chapter showed that these coefficients can be estimated from the data that can be obtained through research. After an introduction to the identification problem, general procedures for the estimation of the parameters and testing of the models were discussed.

Furthermore, it was demonstrated that the classic MTMM design suggested by Campbell and Fiske (1959) can be used to estimate the data quality criteria of reliability, validity, and method effects. This proved that the design can evaluate specific forms of requests for an answer with respect to the specified quality criteria.

There are many alternative models suggested for MTMM data. A review of some of the older models can be found in Wothke (1996). Among them is the *confirmatory factor analysis* model for MTMM data (Althauser et al. 1971; Alwin 1974; Werts and Linn 1970). An alternative parameterization of this model was proposed as the *true score* (TS) model by Saris and Andrews (1991), while the *correlated uniqueness* model has been suggested by Kenny (1976), Marsh (1989), and Marsh and Bailey (1991). Saris and Aalberts (2003) compared models presenting different explanations for the correlated uniqueness. Models with *multiplicative method effects* have been suggested by Campbell and O'Connell (1967), Browne (1984), and Cudeck (1988). Coenders and Saris (1998, 2000) showed that the multiplicative model can be formulated as a special case of the correlated uniqueness model of Marsh (1989). We suggest the use of the *true score (TS) MTMM model* specified by Saris and Andrews (1991) because Corten et al. (2002) and Saris and Aalberts (2003) have shown that this model has the best fit for large series of data sets for MTMM experiments. The classic MTMM model is locally equivalent with the TS model, meaning that the difference is only in its parameterization. For more details on why we prefer this model, see Appendix 10.2.

The MTMM approach also has its disadvantages. If each researcher performed MTMM experiments for all the variables of his/her model, it would be very inefficient and expensive, because he/she would have to ask six more requests to evaluate three original measures. In other words, the respondents would have to answer the requests about the same topic on three different occasions and in three different ways. This raises the questions of whether this type of research can be avoided; if this research is really necessary, and whether or

not the task of the respondents can be reduced.

So far all MTMM experiments have employed the classical MTMM design or a panel design with two waves where each wave had only two observations for the same trait while at the same time the order of the requests was random for the different respondents (Scherpenzeel and Saris 1997). The advantage within the latter method is that the response burden of each wave is reduced and the strength of opinion can be estimated (Scherpenzeel and Saris 2006). The disadvantages are that the total response burden is increased by one extra measure and that a frequently observed panel is needed to apply this design. Although this MTMM design has been used in a large number of studies because of the presence of a frequently observed panel (Scherpenzeel 1995), we think that this is not a solution that can be recommended in general. Therefore, given the limited possibilities of this particular design other types of designs have been elaborated, such as the split-ballot MTMM design (Saris et al. 2004b), which will be discussed in the next chapter. We recommend this chapter only if you are interested in going into the details of this design; otherwise please skip Chapter 11 and move directly to Chapter 12, where a solution of how to avoid MTMM research in applied research is presented.

Exercises

1. A study evaluating the quality of requests measuring "political efficacy" was conducted using following requests for an answer:
   How far do you agree or disagree with the following statements?
   1 *Sometimes politics and government seem so complicated that I can't really understand what is going on.*
   2 *I think I can take an active role in a group that is focused on political issues.*
   3 *I understand and judge important political questions very well.*

   The response categories were:
   1 *Strongly disagree*
   2 *Disagree*
   3 *Neither disagree nor agree*
   4 *Agree*
   5 *Strongly agree*

   The 5-point category scale was used twice: at the very beginning of the questionnaire and once at the end. Therefore the only difference between the two sets of requests was the positioning in the questionnaire. We call these requests "agree/disagree" requests or A/D requests. One other method was used to measure "political efficacy." Instead of the agree/disagree format, a "trait specific method" or TSM request format, was employed. The requests were:

1. *How often do politics and government seem so complicated that you can't really understand what is going on?*
   1 Never
   2 Seldom
   3 Occasionally
   4 Regularly
   5 Frequently

2. *Do you think that you could take an active role in a group that is focused on political issues?*
   1 Definitely not
   2 Probably not
   3 Not sure either way
   4 Probably
   5 Definitely

3. *How good are you at understanding and judging political questions?*
   1 Very bad
   2 Bad
   3 Neither good nor bad
   4 Good
   5 Very good

A MTMM study evaluating these requests led to the following results: First we represent a response distribution for the different requests presenting the means, standard deviations (sd), and the missing values of the distributions of the responses

	First A/D		**Second A/D**		**TSM**	
	Mean	**sd**	**Mean**	**sd**	**Mean**	**sd**
Item 1	2.91	1.21	2.87	1.12	2.90	1.10
Item 2	2.28	1.24	2.38	1.21	2.17	1.21
Item 3	2.94	1.12	3.06	1.08	3.23	.99
Missing	34		82		55	

Below we provide results of the estimation of the reliability, validity and method effects:

	Request 1	Request 2	Request 3
Reliability coefficient			
A/D core	.69	.76	.76
A/D dropoff	.82	.91	.79
TSM dropoff	.88	.92	.87
Validity coefficient			
A/D core	.84	.88	.87
A/D dropoff	1	1	1
TSMdropoff	1	1	1
Method effect[7]			
A/D core	.55	.48	.49
A/D dropoff	0	0	0
TSM dropoff	0	0	0

Please answer the following questions on the basis of the findings of the MTMM study:

a. What are, according to you, the best measures for the different traits?
b. Why are there differences between the measures?
c. Can these hypotheses be generalized to other requests?

2. In Figure 10.4 a MTMM model specifies the relationships between the true scores and their factors of interest:
   a. Express the correlations between the true scores in the parameters of the model. Do this only for those correlations that generate a different expression.
   b. Assuming that each true score has an observed variable that is not affected by any other variable except random measurement error, what do the correlations between the observed variables look like?
   c. Do you have any suggestion about whether the parameters can be estimated from the correlations between the observed variables? (Solving the equations is too complicated.)

APPENDIX 10.1: INPUT OF LISREL FOR DATA ANALYSIS OF A CLASSIC MTMM STUDY

```
Analysis of the British satisfaction data for ESS:
Data ng=1 ni=9 no=428 ma=cm
km
*
1.00
.481 1.00
.373 .552 1.00
-.626 -.422 -.410 1.00
-.429 -.663 -.532 .642 1.00
-.453 -.495 -.669 .612 .693 1.00
-.502 -.374 -.332 .584 .436 .438 1.00
-.370 -.608 -.399 .429 .653 .466 .556 1.00
-.336 -.406 -.566 .406 .471 .638 .514 .558 1.00
mean
*
2.42 2.71 2.45 5.26 4.37 5.13 2.01 1.75 2.01
sd
*
.77 .76 .84 2.29 2.37 2.44 .72 .71 .77

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fr ps=di,fi be=fu,fi ga=fu,fi ph=sy,fi
value -1 ly 1 1 ly 2 2 ly 3 3
value 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9

free ga 1 1 ga 4 1 ga 7 1 ga 2 2 ga 5 2 ga 8 2 ga 3 3 ga 6 3 ga 9 3

value 1 ga 1 4 ga 2 4 ga 3 4
value 1 ga 4 5 ga 5 5 ga 6 5 ga 7 6 ga 8 6 ga 9 6
free ph 2 1 ph 3 1 ph 3 2 ph 6 6 ph 5 5 ph 4 4
value 1 ph 1 1 ph 2 2 ph 3 3
start .5 all
out rs adm=off sc
```

APPENDIX 10.2: RELATIONSHIP BETWEEN THE TS AND THE CLASSIC MTMM MODEL

The structure of the classical MTMM model follows directly from the basic characteristics of the TS model that can be specified in equations (10.2A1) and (10.2A2).

$$y_{ij} = r_{ij}\, t + e_{ij} \tag{10.2A.1}$$

$$t_{ij} = v_{ij} f_i + m_{ij} m_j \tag{10.2A.2}$$

From this model one can derive the most commonly used MTMM model by substitution of equation (10.A.2) into equation (10.2A1). It results in the models (10.2A.3) or (10.2A.4):

$$y_{ij} = r_{ij} v_{ij} f_i + r_{ij} m_{ij} m_j + e_{ij} \tag{10.2A.3}$$

or

$$y_{ij} = q_{ij} f_i + s_{ij} m_j + e_{ij} \tag{10.2A.4}$$

where $q_{ij} = r_{ij} v_{ij}$ and $s_{ij} = r_{ij} m_{ij}$

One advantage of this formulation is that q_{ij}^2 represents the strength of the relationship between the variable of interest and the observed variable and is an important indicator of the total quality of an instrument. Besides, s_{ij} represents the systematic effect of method j on response y_{ij}. Another advantage is that it simplifies equation (9.1) to (10.2A5):

$$r(y_{1j}, y_{2j}) = q_{jr}(f_1, f_2) q_{2j} + s_{1j} s_{2j} \tag{10.2A.5}$$

Although this model is quite instrumental, some limitations are connected with it. One of these is that the parameters themselves are products of more fundamental parameters. This creates problems because the estimates for the data quality of any model are derived only after the MTMM experiment is completed and the data analyzed. Therefore, in order to apply this approach for each item in the survey two more requests have to be asked to estimate the item quality. The cost of doing this makes this approach unrealistic for standard survey research.

An alternative is to study the effects in terms of how different questionnaire design choices affect the quality criteria and to use the results for predicting the data quality before and after the data are collected. By making a meta-analysis to determine the effects of the question design choices on the quality criteria we would be eliminating the additional survey items needed in substantive surveys. It is an approach that has been suggested by Andrews (1984) and has

been applied in several other studies (Költringer 1995; Scherpenzeel and Saris 1997; Corten et al 2002; Saris and Gallhofer 2007).

In such a meta-analysis, it is desirable that the parameters to be estimated represent only one criterion and not mixtures of different criteria, in order to keep the explanation clear. It is for this particular reason that Saris and Andrews (1991) have suggested an alternative parameterization of the classical model: the true score model, presented in equations (10.2A.1) and (10.2A.2), where the reliability and validity coefficients are separated and hence can be estimated independently from each other. Both coefficients can also vary between 0 and 1 which does not occur if one employs the reliability and the validity coefficient as Andrews (1984) did, starting with the classical model (10.2A5). In agreement with Saris and Andrews (1991), we suggested that for the meta-analysis the true score MTMM model has major advantages and therefore we have presented the true score model in this chapter.

This Page Intentionally Left Blank

CHAPTER 11

Split-ballot multitrait-multimethod designs[1]

Although the classical MTMM approach is effective, it is not efficient because respondents have to answer similar requests three times. This may lead to decreased precision because of annoyance or to increased precision since there is more time to think or to correlated errors due to memory effects. Keeping in mind the correlation matrix of Table 10.3, we see that the correlations between the three variables are higher for the second and the third methods than for the first method. Can this be due to a difference in the method, as was argued, or because respondents had more time to think about the issues and came to realize that there are relationships that they did not consider before? The fact that the correlations for the second method are higher while for the third method are lower, could be seen as an *occasion effect* as well as a method effect.

In order to cope with this problem there are two possible strategies: (1) to try to reduce the number of repeated observations and (2) to separate the occasion effect from the method effect. In this chapter we will suggest several designs that can be used as alternatives to the classical MTMM design. These designs reduce the number of observations per person but compensate for the "missing data by design" by collecting data from different subsamples of the population. In doing so, the designs look very similar to the frequently used split-ballot experiments and hence are called the *split-ballot MTMM design* or SB-MTMM design.

11.1 THE SPLIT-BALLOT MTMM DESIGN

In the commonly used split-ballot experiments, random samples from the same population receive different versions of the same requests. In other words, each respondent group gets one method. The split-ballot design makes it possible to compare the response distributions of the different requests across their forms

[1] This chapter is based on a paper by W. E. Saris, A. Satorra, and G. Coenders (2004): A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design, *Sociological Methodology, 2004*.

and to assess their possible relative biases (Schuman and Presser 1981; Billiet et al. 1986).

In the split-ballot MTMM design also random samples of the same population are used but with the difference that these groups get two different forms of the same request. In total it is one less repetition than in the classical MTMM design and one more than in the commonly used split-ballot designs. We will show that our design, suggested by Saris (1998), combines the benefits of the split-ballot approach and the MTMM approach in that it enables researchers to evaluate measurement bias, reliability, and validity simultaneously, and that it does so, while reducing response burden. Applications of this approach can also be found in Saris (1998) and Kogovšek et al. (2001). A more complex alternative design has been suggested by Bunting et al. (2002). The suggestion to use split-ballot designs for structural equation models can be traced back to Arminger and Sobel (1991).

11.1.1 The two-group design

The two-group split-ballot MTMM design is structured as follows. The sample is split randomly into two groups. One group has to answer three survey items formulated by method 1 while the other group is given the same survey items presented in a second form, in the MTMM literature called "method 2." In the last part of the questionnaire all respondents are presented with the three items, which are now formulated in method 3 format. The design can be summarized as tabulated in Figure 11.1.

	Time 1	**Time 2**
Sample 1	Form 1	Form 3
Sample 2	Form 2	Form 3

FIGURE 11.1: *The two-group Split-ballot MTMM design.*

In summary, under the two-group design the researcher draws two comparable random samples from the same population and asks three requests about at least three traits in each sample: one time with the same and the other time with another form (method) of the same requests (traits) after sufficient time has elapsed. Van Meurs and Saris (1990) have demonstrated that after 20 minutes the memory effects are negligible. This time gap is enough to obtain independent measures in most circumstances.

The design in Figure 11.1 matches the standard split-ballot design at time 1 and provides information about differences in response distributions between the methods. Combined with the information obtained at time 2, this design provides extra information. The question still remains whether the reliability, validity and method effects can be estimated from this data, since each respondent answers only two requests about the same trait and not three, as is required from the classical MTMM design. The answer is not immediately

evident since the necessary information for the 9×9 correlation matrix comes from different groups and is by design incomplete (see Table 11.1). Table 11.1 shows the groups that provide data for estimating variances and correlations between requests using either the same or different forms (methods).

Table 11.1: Samples providing data for correlation estimation

	Method 1	Method 2	Method 3
Method 1	Sample 1		
Method 2	none	Sample 2	
Method 3	Sample 1	Sample 2	Sample 1+2

In contrast to the classical design, no correlations are obtained for form 1 and form 2 requests, as they are missing by design. Otherwise all correlations in the 9×9 matrix can be obtained on the basis of one or two samples, but the data come from different samples. This innovative technique was for the first time proposed and used by Saris (1998).

Each respondent is given the same requests only twice, reducing the response burden considerably. However, in large surveys the sample can be split into more subsamples and hence evaluate more than one set of requests. However, the correlations between forms 1 and 2 cannot be estimated, resulting in a loss of degrees of freedom when estimating the model on the now incomplete correlation matrix. This might make the estimation less effective than the standard design where all correlations are available, as in the three-group design.

11.1.2 The three-group design

The three-group design proceeds as the previous design except that three groups or samples are used instead of two, leaving us with the following scheme:

	Time 1	Time 2
Sample 1	Form 1	Form 2
Sample 2	Form 2	Form 3
Sample 3	Form 3	Form 1

FIGURE 11.2: *The three-group split-ballot MTMM design.*

Using this design, all request forms are treated equally: They are measured once at the first and later at a second point in time. There are also no missing correlations in the correlation matrix, as shown in Table 11.2.

Table 11.2: Samples providing data for correlation estimation

	Method 1	Method 2	Method 3
Method 1	Samples 1 and 3		
Method 2	Sample 1	Samples 1 and 2	
Method 3	Sample 3	Sample 2	Samples 2 and 3

Evidently, the major advantage of this approach is that all correlations can be obtained. A second advantage is that the order effects are canceled out because each measure comes once at the first position and another time at the second position within the questionnaire.

A major disadvantage, however, is that the main questionnaire has to be prepared in three different formats for the three different groups. In addition, the same measures are not obtained from all respondents. This may raise a serious issue in the analysis because the sample size is reduced with respect to its relationships with the other variables[2]. This design was for the first time used by Kogovšek et al. (2001).

11.1.3 Other SB-MTMM designs

Other methods that are guided by the principles discussed above can also be designed. The effects of different factors can be studied simultaneously and interaction effects can be estimated. However, an alternative to this type of study is to employ a meta analysis of many separate MTMM experiments under different conditions, which will be elaborated in the next chapter.

There is one other design that deserves special attention: the SB-MTMM design, which makes use of an exact replication of methods. In doing so, the occasion effects can be studied without placing an extra response burden on respondents. A possible design is illustrated in Figure 11.3:

	Time 1	Time 2
Sample 1	Form 1	Form 1
Sample 2	Form 1	Form 2
Sample 3	Form 2	Form 1
Sample 4	Form 2	Form 2

FIGURE 11.3: *A four-group split-ballot MTMM design with exact replications.*

[2] A possible alternative would be to add to the study a relatively small subsample. For the whole sample, one would use method 1, the method expected to give the best results, in the main questionnaire; method 2 for one subgroup, and method 3 for another subgroup in an additional part of the questionnaire that relates to methodology. With the subsample, one would use method 2 for the main questionnaire and method 3 in the methodological part. In this way method 1 is available for all people and all three combinations of the forms are also available. Also one could get an estimate of the complete covariance matrix for the MTMM analysis without harming the substantive analysis. But this design would cost extra money for the additional subsample. The appropriate size of the subsamples is a matter for further research.

Figure 11.3 models a complete four -group design for two methods and their replications. The advantage of this design is that the same information as with the other two designs is obtained and in addition to the previous design, the occasion-specific variance can be estimated. This is only possible if exact repetition of the same measures is included in the design. In order to estimate these effects the model specified in Chapter 10 has to be extended with an occasion-specific factor (Saris et al. 2004). This design can be reduced to a three-group design by leaving out sample 2 or 3 or alternatively sample 1 or 4, assuming that the order effects are negligible or that the occasion effects are the same for the different methods.

Another similar design can be developed including three different methods; however, it is beyond the scope of this chapter to discuss further possibilities. For further information we refer to Saris et al. (2004) and the first two large scale applications of this design in the ESS (2002).

We hope that we clarified that the major advantage of these designs is the reduction of the response burden from three to two observations. Furthermore, in order to show that these designs can be applied in practice, we need to discuss, based on the collected data, the estimation of the parameters.

11.2 ESTIMATING AND TESTING MODELS FOR SPLIT-BALLOT MTMM EXPERIMENTS

The split-ballot MTMM experiment differs from the standard approach in that different equivalent samples of the same population are studied instead of just one. Given the random samples are drawn from the same populations, it is natural to assume that the model is exactly the same for all respondents and equal to the model we have specified in Figure 10.4, which includes the restrictions on the parameters suggested by Saris and Andrews (1991). The only difference is that not all requests have been asked in every group.

Since the assignment of individuals to groups has been made at random, and there is a large sample in each group, the most natural approach for estimating is the *multiple -group SEM method* (Jöreskog 1971). It is available in most of the SEM software packages. We refer to this approach as multiple-groups structural equation model or MGSEM[3]. As indicated in the previous section, a common model is fitted across the samples, with equality constraints for all the parameters across groups. With the current software and applying the theory for multiple-group analysis, estimation can be made by using the maximum

[3] Because each group will be confronted with partially different measures of the same traits, certain software for multiple-group analysis will require some small tricks to be applied. This is the case for LISREL, where the standard approach expects the same set of observable variables in each group. Simple tricks to handle such a situation of the set of observable variables differing across groups were already described in the early work of Jöreskog (1971) and in the manual of the early versions of the LISREL program; such tricks are also described in Allison (1987). Multiple-group analysis with the software EQS, for example, does not require the same number of variables in the different groups.

likelihood (ML) method or any other standard estimation procedure in SEM. In the case of nonnormal data, robust standard errors and test statistics are available in the standard software packages. For a review of multiple-group analysis in SEM models as applied to all the designs enumerated in the present chapter, see Satorra (2000).

The incomplete data set-up we are facing could also be considered as a missing data problem (Muthen et al. 1987). However, the approach for missing data assumes normality, while this design does not provide the theoretical basis for robust standard errors and corrected test statistics that are currently available in MGSEM software. Thus, since the multiple-group option offers the possibility of standard errors and test statistics which are protected from nonnormality, we suggest that the multiple-group approach is preferable.

Given this situation, we suggest the MGSEM approach for estimating and testing the model on SB-MTMM data. In doing so, the correlation matrices are analyzed while the data quality criteria (reliability, validity coefficients and method effects) are obtained by standardizing the solution.

Although the statistical literature suggests that data quality indicators can be estimated using the SB-MTMM designs, we need to be careful while using the two group designs with incomplete data, because they may lead to empirical underidentification problems. Before addressing this issue, we will illustrate an application of the two designs based on data from the same study discussed in the previous chapters.

11.3 Empirical examples

In Chapters 9 and 10 an empirical example of the classic MTMM experiment was discussed. In order to illustrate the difference between this design and the SB-MTMM designs, we have randomly split the total sample of that study (n=428) into two (n=210) and three groups (n= 140). Thereafter, we took only those variables that would have been collected had the two- or three-group MTMM design been used, for each group. In this way, we obtained incomplete correlation matrices for each group. Next, we estimated the model, using the multiple-group approach. Now we will investigate the results, starting with the three-group design, where a complete correlation matrix is available for all groups. Later, we discuss the results for the two-group design, where the correlation information is incomplete.

11.3.1 Results for the three-group design

The random sampling of the different groups and selection of the variables according to the three-group design has led to the results summarized in Table 11.3. First, this table indicates that in each sample incomplete data are obtained for the MTMM matrix. The correlations for the unobserved variables are represented by zeros, the variances by ones. This presentation is necessary for the multiple-group analysis with incomplete data in LISREL but does not have to be used in general.

Table 11.3: Data for three-groups SB-MTMM analysis on the basis of three random samples from the British pilot study of the ESS

Correlations, means, and standard deviations of the first subsample

Correlations
1.00
.469 1.00
.250 .415 1.00
.0 .0 .0 1.00
.0 .0 .0 .0 1.00
.0 .0 .0 .0 .0 1.00
-.524 -.322 -.212 .0 .0 .0 1.00
-.313 -.523 -.273 .0 .0 .0 .509 1.00
-.244 -.313 -.517 .0 .0 .0 .442 .461 1.00
Means
2.39 2.69 2.41 .0 .0 .0 2.09 1.77 2.02
Standard deviations
.70 .71 .78 1.0 1.0 1.0 .71 .68 .73

Correlations, means and standard deviations of the second subsample

Correlations
1.00
.0 1.00
.0 .0 1.00
.0 .0 .0 1.00
.0 .0 .0 .598 1.00
.0 .0 .0 .601 .694 1.00
.0 .0 .0 .588 .398 .517 1.00
.0 .0 .0 .395 .690 .504 .547 1.00
.0 .0 .0 .397 .462 .571 .545 .564 1.00
Means
.0 .0 .0 5.22 4.30 4.98 1.91 1.69 2.00
Standard deviations
1.0 1.0 1.0 2.27 2.51 2.47 .69 .65 .71

Correlations, means and standard deviations of the third subsample

Correlations
1.00
.469 1.00
.393 .605 1.00
-.669 -.454 -.489 1.00
-.512 -.669 -.564 .707 1.00
-.495 -.508 -.742 .693 .729 1.00
.0 .0 .0 .0 .0 .0 1.00
.0 .0 .0 .0 .0 .0 .0 1.00
.0 .0 .0 .0 .0 .0 .0 .0 1.00
Means
2.41 2.65 2.50 5.18 4.32 4.99 .0 .0 .0
Standard deviations
.78 .77 .90 2.39 2.39 2.53 1.0 1.0 1.0

Keep in mind that these correlation matrices are incomplete because at each time interval one set of variables is missing. We see also that we have summarized the response distributions in means and standard deviations, which can be compared across groups as is done in the standard split-ballot experiments. However, in this case we also want estimates for the reliability, validity, and method effects. In estimating these coefficients from the data for the three randomly selected groups simultaneously, we have assumed that the model is the same for all groups except for the specification of variables selected for the three groups. The technical details of this analysis are given in Appendix 11.1, where the LISREL input is presented.

In Table 11.4 we provide the results of the estimation as provided by LISREL using the ML estimator[4]. The table also contains the full sample estimates for comparison. Given that on the basis of sampling fluctuations, one can expect differences between the different groups, the similarity between the results for the two designs indicates that the three-group SB-MTMM design can provide estimates for the parameters of the MTMM model that are very close to the estimates of the classical design. At the same time, the correlation matrices are rather incomplete since the respondents are asked to answer fewer requests about the same topic.

Table 11.4: Estimates of parameters for the full sample using three methods and for the three group design with incomplete data in each group

	Full sample			Three-group SB-MTMM design		
Reliability coefficient for	M1	M2	M3	M1	M2	M3
Q1	.79	.91	.82	.78	.91	.84
Q2	.85	.94	.87	.82	.97	.86
Q3	.81	.93	.84	.83	.95	.77
Validity coefficient for						
Q1	.93	.91	.85	.94	.91	.86
Q2	.94	.92	.87	.94	.93	.85
Q3	.95	.93	.88	.96	.93	.84
Method variance	.05	.73	.09	.04[a]	.73	.09

[a] This coefficient is not significantly different from zero, while all others are significantly different from zero.

Moreover, the fact that the program did not indicate identification problems suggests that the model is identified even though the correlation matrices in the different subgroups are incomplete. Let us now investigate the same example in an identical manner assuming that a two-group design has been used.

[4] In this case LISREL reports a chi^2 of 54.7 with df=111. However, the number of degrees of freedom is incorrect because in each matrix 24 correlations and variances were missing so the df should be reduced by 3x24 = 72 and the correct degrees of freedom are 39.

11.3.2 Two-group SB-MTMM design

Using the two-group design, the same model is assumed to apply for the whole group and the analysis is carried out in exactly the same manner. The data for this design are presented in Table 11.5.

The procedure for filling in the empty cells in the table was the same in Table 11.5 as in Table 11.3. An important difference between the two designs is that in the two-group design no correlations between the first and the second methods are available and so the coefficients have to be estimated on the basis of incomplete data.

Table 11.5: Data for the two-groups SB-MTMM analysis on the basis of two random samples from the British pilot study of the ESS

Correlations, means, and standard deviations of the first subsample
Correlations
1.00
.457 1.00
.347 .478 1.00
.0 .0 .0 1.00
.0 .0 .0 .0 1.00
.0 .0 .0 .0 .0 1.00
-.564 -.365 -.344 .0 .0 .0 1.00
-.366 -.597 -.359 .0 .0 .0 .546 1.00
-.350 -.386 -.530 .0 .0 .0 .512 .498 1.00
Means
2.42 2.75 2.43 .0 .0 .0 2.01 1.70 1.99
Standard deviations
.74 .76 .83 1.0 1.0 1.0 .71 .67 .73
Correlations, means, and standard deviations of the second subsample
Correlations
1.00
.0 1.00
.0 .0 1.00
.0 .0 .0 1.00
.0 .0 .0 .686 1.00
.0 .0 .0 .669 .742 1.00
.0 .0 .0 .585 .449 .441 1.00
.0 .0 .0 .464 .684 .546 .568 1.00
.0 .0 .0 .397 .516 .674 .516 .607 1.00
Means
.0 .0 .0 5.26 4.49 5.10 2.01 1.80 2.02
Standard deviations
1.0 1.0 1.0 2.38 2.40 2.51 .74 .73 .81

The first analysis of these matrices did converge, but the variance of the first method factor was negative. This issue may also arise in the classical MTMM approach when a method factor has a variance very close to zero.

In Table 11.4 we have seen that the method variance for the first factor was not significantly different from zero and rather small even though the estimate was based on two groups of 140 or 280 cases. In the two-group design, the variance has to be estimated on the basis of 210 cases and the program does not provide a proper solution. A common remedy is to fix one parameter on a value close to zero. If we fix the variance on .01, we get the result presented in Table 11.6[5]. With this restriction the estimates provided by the program are close to the estimates obtained in the classical MTMM design. The largest differences in the validity coefficients for the first method are a direct consequence of the restriction introduced.

Table 11.6: Estimates of the parameters for the full sample using three methods and for the two-group design with incomplete data

Reliability	Full sample			Two-group SB-MTMM design		
	M1	M2	M3	M1	M2	M3
Q1	.79	.91	.82	.80	.93	.83
Q2	.85	.94	.87	.87	.96	.86
Q3	.81	.93	.84	.83	.98	.82
Validity						
Q1	.93	.91	.85	.99	.90	.85
Q2	.94	.92	.87	.99	.91	.86
Q3	.95	.93	.88	.99	.92	.87
Method variances	.05	.73	.09	.01[a]	.86	.10

[a] this coefficient was fixed on the value .01 in order to avoid an improper solution.

On the whole, the estimates are such that regarding the reliability coefficients, the conclusion drawn from the estimates obtained by the two-group design would not differ from those using the estimates of the one-group design where the second method has the highest reliability. Given the restriction introduced on the method variance, one should be very cautious to draw a definite conclusion about the validity coefficients and hence about the method effects.

Clearly the fact that we had to introduce this restriction raises the question of whether the two-group design is identified and robust enough to be useful in practice. On the one hand, it would seem that the most natural approach is to reduce the response burden. On the other hand, when this approach is not

[5] In this case LISREL reports a chi² value of 12.7 with df=67 but also now the df has to be corrected in the way discussed above (footnote 4) and the correct df are 19.

robust enough to provide the same estimates as the classical or the three-group SB-MTMM design, then one of the other designs should be preferred.

With regard to the identification, we assert that the model is indeed identified under normal circumstances and the specified estimation procedure will provide consistent estimates of the population parameters.

Before proceeding to the next section, we should mention that the example above did not give a correct impression of the true quality of the different designs. The reason is that the quantity of data on which the parameters are based differed for the parameters in the different designs. The parameters of the classical design were based on approximately 420 cases. The parameter estimates in the three-group design are based on 280 respondents while some parameter estimates in the two-group design are based on either 210 or 420 cases. This consideration provides us with one explanation for the difference in performance between the designs. Hence, the topic of efficiency of the different designs is covered in the next section.

11.4 The empirical identifiability and efficiency of the different SB-MTMM designs

In order to study the robustness of these different designs, two different problems have to be evaluated. The first is that we would like to determine under what conditions the procedures break down even though the correct model has been specified. The second issue is what we can say about the efficiency of the different designs to estimate the parameters of the MTMM model. We will begin with addressing the first issue.

11.4.1 The empirical identifiability of the SB-MTMM model

Three aspects of these models require special attention after the model has been correctly specified:

1. Minimal variance of one of the method factors
2. Lack of correlation between the latent traits
3. Equal correlations between the latent traits

The first problem, of the minimal method variance, is a problem of overfitting. In this case a parameter is estimated that is not needed for the fit of the model to the data. If the model had been estimated with this coefficient fixed on zero, the fit would be equally good. This problem is not just an issue for SB-MTMM designs; it also occurs with the classical MTMM design. The solution to this problem, as mentioned above, is to specify the parameter that is not needed for the model on zero or a value close to zero. However, it is more of a challenge to detect where the actual problem in the model lies. Our experience with the analyses of MTMM data is that negative variances for the method variances are obtained in unrestricted estimation procedures if the variances are very close to zero. Therefore, in such cases, restricting the variances to a value very close to zero solves the problem. In the case where estimation procedures

include constraints on the parameter values, the value zero will automatically be obtained for the problematic method variance in order to avoid improper solutions.

The second condition, lack of correlations between the traits, can raise a problem because we know that the loadings of a factor model are identified if each trait has three indicators or two but then the traits have to be correlated with each other. If each trait has only two indicators and the correlation between the traits is zero, the situation is the same as for a model with one trait and two indicators, which is not identified. Applying this rule to the MTMM models, we can see that in the classical MTMM model each trait has three indicators and is therefore identified under normal circumstances even if the correlations between the traits are zero. In the different groups of the SB-MTMM designs, each trait has only two indicators. Therefore, if the correlation between two traits is zero, the model in the different subgroups will not be identified. Hence, if all three or two of the three correlations go to zero the standard errors of the parameters become very large. This is an indication that a problem of identification exists in the two- and in the three-group designs. Fortunately there is a simple solution to this issue – if the researcher has some freedom of choice in the selection of the traits for the experiments, then he/she can select traits for the experiments with sufficient correlation to avoid these problems. In other words, this type of problem can be prevented by choosing appropriate traits at the design stage.

The third condition that can cause problems was detected by chance while studying the identification of the two-group MTMM design. It was discovered that the basic model of the two-group SB-MTMM design is not identified if the correlations between the traits are identical. However, this is an uncommon scenario and it suffices to be aware of it. If confronted with a situation where the standard errors are rather large while the correlations between the traits are not close to zero, a possible explanation may be the equality of the correlations.

The discussion so far suggests that the SB-MTMM design with two groups can be used with traits that are correlated with each other but do not have equal correlations. Under these rather elementary conditions even the SB-MTMM designs with two groups will be identified and the multiple-group ML estimator will provide consistent estimates. For the three-group design, the previously mentioned requirements are not necessary.

11.4.2 The efficiency of the different designs

The second issue to be discussed is the efficiency of the different designs. This is a relevant issue because the reduction of the response burden might be gained at the expense of the efficiency of the methods. The efficiency of the different designs has been studied on the basis of the standard errors of the estimates of reliability and validity by Saris et al. (2004b). This study shows that for very small method variances the total sample of a two-group design has to

be very large. A much smaller total sample is needed for the three-group design. However, one should realize that the standard error for very small method variances is also minimal unless the variance is equal to zero as discussed above. This study also shows that the efficiency of the two- and three-group designs become quite similar if the method variance becomes larger. The researcher needs to keep in mind that for both designs the total sample sizes need to be considerably larger than 300, which is the chosen sample size for the one-group design.

With respect to reliability the study shows the same pattern, specifically, that the efficiency of the two- and three-group designs become nearly the same when the error variance becomes larger. The inefficiency of the two designs for very small error variances compared with the one-group design also becomes apparent. Fortunately or unfortunately, these very small error variances do not occur in survey research.

11.5 Conclusion and Discussion

We have pointed out that the classic MTMM design has its disadvantages because respondents have to answer approximately the same requests three times. As a solution, we have suggested an alternative known as the split-ballot MTMM design. This chapter has shown that the split-ballot MTMM experiment reduces the response burden, by decreasing the number of items to be asked in a questionnaire, without loss of information with regard to reliability and validity measures. Requests concerning the same trait need to be answered only twice and not three times as is required in the classical MTMM approach. The advantage is that it reduces the response burden effects. However, the effects of repeating requests concerning the same concept cannot be eliminated completely. Repeating the requests about the same concepts in different forms is necessary for estimating the reliability and validity of the measures. However, it was shown that two- or three-group designs with repeated observations of exactly the same measures can be used to estimate these effects. Also the meta-analyses discussed in the next chapter provide estimates of the effect of the repeated observations and allow for correction for this effect.

For the time being, we suggest analyzing the data of these multiple-group designs using the options available for MGSEM in standard software. With some programs, this may be a bit more complicated task when using the multiple group approach, but as compensation for this we can obtain corrections for the standard errors to cope with nonnormality (an option that is not available within the missing data approach).

Concerning the efficiency of the different designs it has been found that the three-group design is far more efficient than the two-group design at least for the small method variances and error variances. Thus the total sample sizes can be reduced by using three groups instead of two groups if the errors are rather small. If the errors become larger the designs become equally efficient, although, for all practical purposes, the three-group design remains a bit more

efficient than the two-group design. However, the disadvantage of the three-group design is that it requires more forms of the questionnaire. Furthermore, in the three-group design there will always be one group that does not have the same variables as the other two groups. This means that for these variables the group will not provide any data, being rather inefficient. Therefore, in this respect the two-group design is more attractive.

Whatever the design may be, it will be clear that MTMM studies on the quality of survey items will cost extra time and effort and provide information only about limited types of items. It would be attractive if one could use information about the quality of survey items on the basis of a sufficiently large quantity of experiments. In that case additional research is avoided, and this will translate into saved time and effort for both the researcher and the respondent. Suggestions on how to do this will be the topic of the next two chapters.

EXERCISES

1 Specify a MTMM design with one group for three requests of your own questionnaire:
   a. Specify the different requests.
   b. Specify the required sample size.
   c. Specify the correlation matrix that will be obtained.

2 Check how this design changes if you would use a two-group SB-MTMM design
   a. Specify the different requests.
   b. Specify the required sample size.
   c. Specify the correlation matrices that will be obtained.

3 Check how this design changes if you would use a three-group SB-MTMM design
   a. Specify the different requests.
   b. Specify the required sample size.
   c. Specify the correlation matrices that will be obtained.

4 Which of the three designs discussed would you prefer to collect information about the quality of requests?

APPENDIX 11.1: THE LISREL INPUT FOR THE 3 GROUPS SB-MTMM EXAMPLE

```
Analysis of the British satisfaction data with three-groups SB-MTMM model  group 1:
Data ng=3  ni=9 no=140 ma=cm
km
*
1.000
0.469   1.000
0.250   0.415   1.000
0.000   0.000    0.0000   1.000
0.0000   0.0000   0.0000   0.0000   1.000
0.0000   0.0000   0.0000   0.0000   0.0000   1.000
-.524   -.322    -.212    0.0000   0.0000   0.0000 1.000
-.313   -.523    -.273    0.0000   0.0000   0.0000 0.509 1.000
-.244   -.313    -.517    0.0000   0.0000   0.0000 0.442 0.461  1.000
me
*
2.39 2.69 2.41  0.0 0.0 0.0  2.09 1.77 2.02
sd
*
.70 .71 .78  1.0 1.0 1.0  .71 .68 .73
model ny=9 ne=9 nk=6 ly=fu,fi te=di,fr ps=di,fi be=fu,fi ga=fu,fi ph=sy,fi
value -1 ly 1 1 ly 2 2 ly 3 3
value 0 ly 4 4 ly 5 5 ly 6 6
pa te
15 16 17 0 0 0 18 19 20
value 1 te 4 4 te 5 5 te 6 6
value 1 ly 7 7 ly 8 8 ly 9 9
free ga 1 1 ga 4 1 ga 7 1 ga 2 2 ga 5 2 ga 8 2 ga 3 3 ga 6 3 ga 9 3
value -1 ga 1 4 ga 2 4 ga 3 4
value 1 ga 4 5 ga 5 5 ga 6 5 ga 7 6 ga 8 6 ga 9 6
free ph 2 1 ph 3 1 ph 3 2 ph 4 4 ph 5 5 ph 6 6
start .01 ph 4 4
value 1 ph 1 1 ph 2 2 ph 3 3
start .5 all
value .13 ph 5 5
value .18 ph 6 6
start .75 ga 1 1 ga 2 2 ga 3 3 ga 7 1 ga 8 2 ga 9 3
start .85 ga 4 1 ga 5 2 ga 6 3
out iter= 200 adm=off sc

Analysis of British satisfaction group 2
Data ni=9 no=150 ma=cm
```

```
Km
*
1.000
0.0   1.000
0.0   0.0   1.000
0.0   0.0   0.0   1.000
0.0   0.0   0.0   0.598  1.000
0.0   0.0   0.0   0.601  0.694  1.000
0.0   0.0   0.0   0.588  0.398  0.517 1.000
0.0   0.0   0.0   0.395  0.690  0.504 0.547  1.000
0.0   0.0   0.0   0.397  0.462  0.571 0.545  0.564  1.000
me
*
.0 .0 .0 5.22 4.30 4.98  1.91 1.69 2.00
sd
•
1.0 1.0 1.0 2.27 2.51 2.47 .69 .65 .71

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fr ps=in be=in ga=in ph=in
value 0 ly 1 1 ly 2 2 ly 3 3
pa te
0 0 0 21 22 23 18 19 20
value 1 te 1 1 te 2 2 te 3 3
value 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
out iter= 200 adm=off sc

Analysis of the British satisfaction group 3
Data ni=9 no=150 ma=cm
Km
*
1.000
0.469  1.000
0.393  0.605  1.000
-.669  -.454  -.489   1.000
-.512  -.669  -.564   .707    1.000
-.495  -.508  -.742   .693    .729    1.000
0.0   0.0   0.0   .000    .000    0.000 1.000
0.0   0.0   0.0   .000    .000    0.000 0.000  1.000
0.0   0.0   0.0   0.0000   0.0000   0.000 0.000  0.000 1.000
me

*
2.41 2.65 2.50 5.18 4.32 4.99 .0 .0 .0
sd
```

```
*
.78 .77 .90 2.39 2.39 2.53 1.00 1.00 1.00

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fr ps=in be=in ga=in ph=in
value 0 ly 7 7 ly 8 8 ly 9 9
pa te
15 16 17 21 22 23 0 0 0
value 1 te 7 7 te 8 8 te 9 9
value 1 ly 4 4 ly 5 5 ly 6 6
value -1 ly 1 1 ly 2 2 ly 3 3
out iter= 200 adm=off sc
```

This Page Intentionally Left Blank

CHAPTER 12

Estimation of the effects of measurement characteristics on the quality of survey questions[1]

The experiments presented in the previous chapter are typical of the MTMM experiments of the last 30 years. Such studies have been conducted by Andrews (1984) and Rodgers et al. (1992) in the United States. Költringer (1995) has conducted a similar study for German questionnaires, while Scherpenzeel and Saris (1996) in the Netherlands and Billiet and Waege in Belgium have conducted similar studies regarding Dutch questionnaires. In total, 87 MTMM studies are available containing 1023 survey items. All of these studies are based on, at least regional samples of the general population. In the United States, the Detroit area was studied, in Austria and the Netherlands national samples were used, while in Belgium random samples of the Flemish-speaking part of the population were taken. The topics in the different experiments are highly diverse. In general, the MTMM experiments are integrated into normal survey research where three or more questions of the survey are used for further experimentation. This approach guarantees that questions that are common to survey research are used. The same is true for the variation in the choices made in the design of survey items. The experiments are designed for the most commonly used methods (choices). More details on the studies are presented in the Appendix of this chapter and the previously mentioned publications.

12.1 A CROSS-CULTURAL STUDY OF DATA QUALITY

In order to integrate the 87 MTMM studies that were carried out in three languages they were reanalyzed, and the survey items were coded according to characteristics listed in Table 12.1. Scherpenzeel (1995) has indicated that without this recoding, the results of the different studies were incommensurable. Therefore, all survey items were coded in exactly the same manner. The code-book is available at the SQP website[2]. The data of the different studies

[1] This chapter is a reprint of a paper by Saris and Gallhofer (2007)Estimation of the effects of measurement characteristics on the quality of survey questions, *Survey Research Methods*, 31–46. The paper has been reprinted with permission of the editor of SRM.

[2] The codebook can be found at www.sqp.nl.

were pooled and an analysis conducted over all available survey items adding a variable "language" to it in order to take into account any effect due to differences in languages.[3]

Normally, multiple-classification analysis or MCA is applied (Andrews 1984; Scherpenzeel 1995; Költringer 1995) to meta-analysis, but the number of variables that need to be introduced in the analysis makes it impossible. A solution is (dummy) regression. The following equation presents the approach used:

$$C = a + b_{11}D_{11} + b_{21}D_{21} + \cdots + b_{12}D_{12} + b_{22}D_{22} + \cdots + b_3N_{cat} + \cdots + e \qquad (12.1)$$

In this equation, C represents the score on a quality criterion, which is either the reliability or validity coefficient. The variables D_{ij} represent the dummy variables for the j^{th} nominal variable. All dummy variables have a zero value unless a specific characteristic applies to the particular question. For all dummy variables, one category is used as the reference category which has received the value "zero" on all dummy variables within that set. Continuous variables, like the number of categories (N_{cat}), were not categorized, except when it was necessary to take nonlinear relationships into account. The intercept is the reliability or validity of the instruments if all variables have a score of zero. Table 12.1 shows the results of the meta-analysis over the available 1023 survey items. Table 12.1 indicates the effects of different survey design choices on the quality criteria of validity and reliability. The table contains also the standard errors (se) of these coefficients and their significance level (sign). The method effects were not indicated because they can be derived from the validity coefficients.

Each coefficient indicates the effect of a 1 point increase on each indicated characteristic while keeping all other characteristics constant. For example, all questions concerning "consumption," "leisure," "family," "personal relations" and "race" are coded as zero on all domain variables that can be seen as the reference category. For these questions the effect on reliability and validity is zero. Questions concerning other issues are coded further into several categories. If a question concerns "national politics" it belongs to the first domain category (D_{11}=1 for this category, while all other domain variables D_{i1}=0) and its effect on reliability and validity will be positive, .0528 and .0447, respectively as can be seen from the table. Note that all the effects in the table are multiplied by 1000. If a question concerns "life in general" then the fifth category applies (D_{51}=1) and the effects are negative: -.0768 and -.0159, respectively. From these results it also follows that questions concerning national politics have a reliability coefficient of .0528 + .0768 or .1296 higher than the questions about

[3] The analysis shows that the effect of language is additive, meaning that language affects only the absolute level of the quality indicators. If this were true for all languages, it would mean that comparisons of choices could be made for all languages and only the absolute level of the quality criteria could be incorrect.

life in general. This interpretation holds for all characteristics with a dummy coding such as "concepts," "time reference," and so on.

Other characteristics using at minimum an ordinal scale are treated as metric. For example, "centrality" is coded in five categories from "very central" to "not central at all." In this case an increase of one point gives an effect of -.0172 on reliability and the difference between a very central or salient item and a not at all central item is $5 \times -.0172 = -.0875$.

Furthermore, there are real numeric characteristics like the "number of interrogative sentences," "the number of words." In that case, the effect is an increase of one unit per word or interrogative sentence.

A special case in this category is the variable "position" because it turns out that while the effect of "position" on reliability is linear, for validity it is non-linear. To describe the latter relationship, the "position" variable is categorized, and the effects are determined within their respective categories.

Table 12.1: Results of the metaAnalysis

Variables	Number of measures	Effect on reliability			Effect on validity		
		Effect	se	sign	Effect	se	sign
				Domain			
National politics (0–1)	137	52.8	12.3	.000	44.7	10.9	.000
International politics (0–1)	64	29.4	18.1	.104	57.8	15.9	.000
Health (0–1)	82	16.9	13.9	.225	21.6	12.0	.073
Living condition/background (0–1)	223	21.4	8.7	.014	4.6	7.4	.541
Life in general (0–1)	50	-76.8	12.6	.000	-15.9	10.8	.139
Other subjective variables (0–1)	235	-66.9	14.2	.000	-1.0	12.4	.935
Work (0–1)	96	12.8	12.0	.287	28.2	10.4	.007
Others:	136	0.0	–	–	0.0	–	–
				Concepts			
Evaluative belief (0–1)	96	6.1	14.0	.669	13.8	12.3	.260
Feeling (0–1)	110	-4.2	10.9	.704	-7.5	9.4	.427
Importance (0–1)	96	35.9	15.6	.021	18.6	13.6	.171
Future expectations (0–1)	39	2.6	24.0	.913	-9.0	20.6	.662
Facts: background (18)							
Behavior (9) (0–1)	27	-126.2	21.8	.000	-150.5	19.2	.000
Other simple concepts	578	0.0	–	–	0.0	–	–
Complex concepts	1023	-72.3	17.4	.000	-47.2	15.2	.002

Table 12.1: continued

Variables	Number of measures	Effect on reliability Effect	se	sign	Effect on validity Effect	se	sign
		Associated characteristics					
Social desirability: no/ a bit/much (0–2)	1023	2.3	6.2	.709	8.0	5.3	.137
Centrality: very central not central (1–5)	1023	-17.2	5.2	.001	-8.9	4.4	.046
Time reference:							
Past (0–1)	106	43.9	15.0	.004	-1.6	12.9	.901
Future(0–1)	83	-13.3	16.1	.409	-10.1	13.8	.465
Present (0–1)	940	0.0	–	–	0.0	–	–
		Formulation of the requests: basic choice					
Indirect question							
Ex: Agree/disagree (0–1)	167	4.0	10.9	.713	41.6	9.5	.000
Other types: Direct request (190), More steps[i] (22)	212	0.0	–	–	0.0	–	–
Use of statements or stimulus (0–1)	317	-23.0	12.4	.065	-12.1	11.1	.275
Use of gradation (0–1)	809	79.6	14.1	.000	-22.8	12.4	.066
		Formulation of the request: other choices					
Absolute–comparative (0–1)	98	12.7	16.3	.436	-8.4	14.5	.564
Unbalanced (0–1)	411	-3.2	11.2	.772	-22.3	9.7	.022
Stimulance (0–1)	92	-11.1	13.3	.406	-11.7	11.5	.308
Subjective opinion (0–1)	86	-5.9	19.9	.767	-34.3	17.2	.047
Knowledge given (1–4)	358	-12.7	8.8	.145	-6.3	7.5	.401
Opinion given (0–1)	101	.653	14.5	.964	-10.3	13.1	.429
		Response scale: basic choice					
Yes/no (0–1)	3	-22.2	19.5	.254	-1.9	17.1	.911
Frequencies	23	120.8	24.8	.000	-95.9	21.5	.000
Magnitudes	169	116.2	20.8	.000	-115.5	18.3	.000
Lines	201	118.1	20.9	.000	-32.7	18.2	.073
More steps	26	48.7	27.3	.075	24.5	23.5	.297
Categories	630	0.0	-	-	0.0	-	

Table 12.1: continued

Variables	Number of measures	Effect on reliability Effect	se	sign	Effect on validity Effect	se	sign
		Response scale: other choices					
Labels: no/some/all (1–3)	1023	33.0	10.0	.001	-4.5	8.8	.605
Magnitudes	169	116.2	20.8	.000	-115.5	18.3	.000
Lines	201	118.1	20.9	.000	-32.7	18.2	.073
More steps	26	48.7	27.3	.075	24.5	23.5	.297
Categories	630	0.0	–	–	0.0	–	
Kind of label: short, sentence (0–1)	35	-47.5	16.0	.003	-9.1	13.7	.506
Don't know: present, registered, not present (1–3)	1023	-6.7	4.8	.165	-1.9	4.1	.647
Neutral: present, registered, not present (1–3)	1023	12.6	4.6	.007	8.4	4.0	.038
Range: Theoretical renge and scale unipolar Theoretical range and scale bipolar Theoretical range bipolar but scale unipolar (1–3)	1023	-15.1	9.6	.116	9.2	8.5	.277
Correspondence: high–low (1–3)	1023	-16.8	7.5	.025	1.1	6.5	.867
Symmetric labels (0–1)	195	25.5	11.8	.031	22.3	10.4	.033
First answer category: negative, positive (1–2)	358	-7.5	8.7	.387	14.7	7.6	.052
Fixed reference points (0– 3)	1023	14.7	4.3	.001	21.4	3.7	.000
Number of Categories (0–11)	1023	13.5	2.1	.000	-1.9	1.8	.298
Number of Frequencies (0–5000)	1023	-.068	.009	.000	-.065	.008	.000
		Survey item specification: basic choices					
Question present (0–1)	841	27.2	15.2	.074	11.5	13.1	.379
Instruction present (0–1)	103	-43.7	15.4	.005	-4.2	13.3	.753
No question or instruction	79	0.0	–	–	0.0	–	–
Respondent's instruction (0–1) interviewer's	492	-12.7	7.3	.083	-14.9	6.2	.017
instruction (0–1)	119	-.068	10.5	.995	5.7	9.0	.524
Extra motivation/information or definitions (0–3) >0	304	7.1	6.7	.296	-.3	5.7	.959
introduction (0–1)	515	5.7	12.1	.637	-10.5	10.3	.312

Table 12.1: continued

Variables	Number of measures	Effect on reliability Effect	se	sign	Effect on validity Effect	se	sign
		Survey item specification: other choices					
Complexity of the introduction							
Question in intro (0–1)	62	-44.6	16.3	.006	-21.3	14.1	.132
Number of subordinate clauses >0	129	29.3	9.8	.003	7.6	8.6	.377
Number of words per sentence >0	510	-1.3	.867	.134	1.4	.75	.063
Mean of words per sentence >0	510	.064	1.1	.954	-.373	.9	.699
Complexity of question							
Number of sentences (0–n)	192	12.7	9.8	.199	-8.3	8.6	.335
Number of subordinate clauses (0–n)	746	13.6	6.8	.048	-17.7	5.9	.003
Number of words (1–51)	1023	.809	.749	.280	-1.3	.644	.041
Mean of words per sentence (1–47)	1023	-2.2	.926	.014	1.1	.807	.161
Number of syllables per word (1–4)	1023	-32.5	9.6	.001	-10.4	8.2	.207
Number of abstract nouns on the total number of nouns (0–1)	1023	2.9	27.7	.917	-13.9	23.7	.558
		Mode of data collection					
Computer assisted (0–1)	626	-3.8	12.6	.760	-38.3	10.7	.000
Interviewer administered (0–1)	344	-50.8	22.9	.027	-104.1	19.5	.000
Oral (0–1)	219	10.4	12.2	.397	25.3	10.3	.014
		Position in questionnaire					
In battery (0–1)	225	-10.3	12.3	.403	28.9	10.7	.007
Position of question	1023	.304	.064	.000			
Position 25 (1–25)	396				1.5	.402	.000
Position 100 (26–100)	458				.420	.137	.002
Position 200 (101–200)	129				.267	.062	.000
Position 300(>200)	12				.098	.100	.333

Table 12.1: continued

Variables	Number of measures	Effect on reliability Effect	se	sign	Effect on validity Effect	se	sign
		Language used in questionnaire					
Dutch (0–1)	731	-20.3	22.8	.373	-76.0	19.8	.000
English (0–1)	174	-72.0	26.6	.007	-2.9	22.9	.899
German (0–1)	118	0.0	–	–	0.0	–	–
		Sample characteristics					
Percentage of less educated (3–54)	993	-.911	.596	.127	1.1	.511	.027
Percentage of high age (1–49)	1023	-.410	.560	.464	-.753	.488	.123
Percentage of males (39–72)	1023	-.030	.690	.966	.405	.596	.497
		MTMM design					
Design: one or more time points (0–1)	713	4.36	16.3	.790	-36.9	14.3	.010
Distance between repeated methods (1–250)	1023	-.169	.094	.072	-.249	.081	.002
Number of traits (1–10)	1023	-.370	2.0	.855	-1.7	1.7	.320
Number of methods (1–4)	1023	.959	2.6	.715	-2.3	2.2	.314
Intercept	825.2	69.5	.000	1039.4	60.4	.000	
Explained variance (adjusted)		.47			.61		
Correction for single item distance		-42.3			-62.25		
Starting point for single item		782.9			977.15		

Another exception is the "number of categories in the scale." For this variable we have specified an interaction term, because the effects were different for categorical questions versus frequency measures. Therefore, depending on whether the question is a categorical or a frequency question, a different variable is specified to estimate the effect on the reliability and the validity.

12.2 Results of the meta-analysis

Below we discuss the most important results presented in Table 12.1.

Domain, concept, and associated characteristics

- The research design determines the domain, concepts, and associated characteristics. Nevertheless, there are significant differences in reliability and validity for items from different domains, measuring different concepts or with different associated characteristics.
- Behavioral survey items tended to have a more negative effect than attitudinal questions, especially items concerning the "frequency of behavior." Although only a few items of this type were analyzed; therefore, the standard error of the effect is relatively large.

- Complex items should be avoided where ever possible, given their negative effect.
- It appears that reporting about the past is more reliable than reporting about the future or the present.

Formulation of the requests

In formulating the requests, the researcher has more freedom of design. We found that

- Indirect requests such as agree/disagree options perform similarly to direct requests on reliability and a bit better with respect to validity.
- The use of statements or stimuli has a small negative effect on reliability and validity; therefore, it is better to avoid them.
- On the other hand, the reliability improves with gradation requests, although they have a small negative effect on validity.
- A lack of balance in the formulation of the request has a significant negative effect on validity.
- Emphasizing subjective opinion has a significant negative effect on validity.

Response scale

- Use of response scales with gradation in the form of frequency, magnitude estimation or line production and the stepwise procedure has a positive effect on reliability, but is often associated with strong method effects such as rounding off errors, which reduces validity.
- Line production and stepwise procedures incur a relatively smaller method effect.
- Reliability is improved when labels instead of complete sentences are used.
- Not providing a neutral middle category improves both reliability and validity significantly.
- The use of fixed reference points has a quite large positive effect on reliability and validity. This approach is especially recommended for long scales with 7 or more categories.
- The effect of range is rather limited, which may be due to the selected categories.
- Making the numbers correspond with the labels has a significant positive effect on reliability.
- Symmetry within response categories has some positive effect on reliability and validity.
- The number of categories has an opposite effect for category and frequency scales. In the case of a category scale (2-points – 15-points and more steps procedures), reliability can be increased by more than .1 by going from a 2-point to an 11-point scale.
- In the case of a frequency scale, reliability and validity experience a large decrease if the range of the scale is too wide (i.e., if very high frequencies are possible).

- For magnitude estimation and line production, this effect does not apply. The number of categories seems to be integrated in the effect of the method itself.

Specification of the survey item as a whole

- The first item is more reliable if a normal request is asked and less reliable if an instruction is used, in comparison to subsequent items in a battery.
- Items in a battery without a request for an answer (almost all items except the first one) are better than items with an instruction but worse than items with a normal request for an answer. This may be due to the complexity of the procedure, which requires extra instruction, and not because of the effect of the instruction. The same may hold true for our discussion of the next effect.
- Respondents' instructions have a significant negative effect on reliability and validity. The item may be so difficult that it requires an explanation, and therefore the effect may be caused by the item and not by the instruction.
- Interviewer instructions, extra motivational remarks, definitions, and an introduction seem to have no significant effect on reliability or validity.
- Formulating general questions in the introduction, which are followed by the real request, should be avoided because they have a negative effect on both reliability and validity.
- On the other hand, a positive effect on reliability has been found if more explanation is given in subordinate clauses of the introduction.
- This effect holds true for the request itself, having also a positive effect on validity.
- However, there is a limit to the number of words in the request, if it becomes too long, it has a negative effect on validity.

The two indices for complexity of requests, the number of words per sentence (sentence length), and the number of syllables per word (word length), have a significant negative effect on reliability[4].

Mode of data collection

The mode of data collection can be analyzed by each basic method or by a general description.

- The CAI is as reliable as the non-CAI; however, it is less valid.
- A much stronger negative effect can be observed for interviewer-administered questionnaires than for the other methods.
- Oral questionnaires have a small but significant positive effect on the validity.

---

[4] The variables "syllables/word" and "proportion of abstract words" have been collected for the introduction and the question itself; however, in the introduction these variables correlated very highly with each other and with the variable "intro" and it was decided that these variables cannot be used together with the variable "introduction."

Position in the questionnaire

- The effect of the position of a request within a questionnaire is rather different for either reliability or validity.
- It seems that respondents continuously learn about how to fill in the questionnaire, causing the reliability of the response to increase linearly with its position. Over the range studied, the effect can be more than 100 points.
- On the other hand, the effect on validity is .037 point for the first 25 requests, followed by an effect of .031 for the 25th request until the100th, and for the 100th – 200th this effect is .026 while after the 200th request there is no further significant increase.

Basic choices for which correction is necessary

Some choices cannot be explicitly made such as language or the characteristics of a population. These choices can nevertheless have an influence on the quality criteria. In addition, the methodological experiments that form the basis for this meta-analysis also have some influence that has to be estimated and controlled for when the other effects are estimated.

- Unfortunately, compared with questionnaires in German, questionnaires in English are significantly less reliable, while Dutch questionnaires are significantly less valid.
- Of the three characteristics of the samples studied only the education level has a significant effect on the validity of responses. Samples with a high number of lower educated people may score in validity .050 lower than samples with few poorly educated people.
- The MTMM design used also has a significant effect on the data quality. As the distance in time between the items for the same trait increases, the reliability declines. For the largest distance found the reliability decreased by .042.
- The distance between the traits has an even larger effect on validity; for the largest distance found, the validity decreased by .062.

In a normal survey MTMM experiments are not present and one measure is available for each trait. Therefore, for predicting the quality of survey items, a correction for the fact that a survey item appears only once within the questionnaire has to be made. This correction is specified at the bottom of Table 12.1. We have corrected for the distance of the "previous measure of the same trait," where the intercept is adjusted by subtracting .0423 for reliability and .06225 for validity.

12.3 SPECIAL TOPICS

In this section, we will focus on the effects of certain choices that warrant further detail.

The choice of direct requests or agree/disagree requests

Agree/disagree requests score better on validity (.041) than do direct requests. However, agree/disagree requests are most commonly used in batteries, and we have found that compared with items presented later in a battery (with no question or instruction), a direct question is more reliable (.0272) while an instruction is less reliable (-.0437). Hence a difference in reliability between the two procedures of .0709 is compensated by .041 in validity. This difference is in favor of direct questions. Differences in reliability between these two types of questions also have been found in other studies (Saris and Galhofer 2006). However, it is somewhat surprising to find that agree/disagree procedures score higher on validity. It is anticipated that acquiescence would lead to the opposite effect (Krosnick and Fabrigar 1997); therefore this issue needs to be investigated further.

The effect of the number of categories

There is still no consensus about the effect of an increase in the number of categories in the scale on quality. Cox (1980), and Krosnick and Fabrigar (1997) defend the position that one should not use more than seven categories while Andrews (1984), Költringer (1995), and Alwin (1997) argue to the contrary that more categories lead to better results. Our analysis suggests that frequency scales, magnitude scales, and line scales are generally more reliable than category scales. However, frequency and magnitude scales especially pay the price for reliability by sacrificing validity. This phenomenon has two reasons. The first is that people round off their numeric values in a specific way. Some use numbers divisible by 25, others are more precise and use numbers divisible by 10, and others use even numbers divisible by 5. Such differences in behavior cause method effects. The other possible explanation is what Saris (1988) has called "variation in response functions." When respondents are allowed to specify their own response scales this will lead to method effects and as a consequence to lower validity coefficients. The solution suggested by Saris (1988) is confirmed by this analysis because better validity and reliability is obtained if the scales are made comparable through use of fixed reference points (see Chapter 7).

The reliability of category scales can also be improved by using more categories (so far up to 11 categories were studied) without decreasing validity. An alternative is to use a two-step procedure that improves both reliability and validity. Category scales can also be improved using labels for most categories as long as they are not in full sentence format. In summary, this analysis strongly suggests to use as many categories as possible in a category scale (more than seven) that are short and clearly labeled. Line production or magnitude estimation with fixed reference points are the optimal choice in most cases and should be used whenever possible.

Effects of the mode of data collection

On the basis of the choices specified in Table 12.1, the commonly used data collection methods can be constructed by combining different characteristics. Their results and the effects of their combinations on reliability and validity are presented in Table 12.2.

Table 12.2: Effects of modes of data collection on data quality, based on the combined effect of computer-assisted data collection and interviewer-administered data collection

	CAI	Not CAI
Interviewer-administered	CATI/CAPI	PAPI/TEL
Reliability coefficient	-.0538	-.050
Validity coefficient	-.1423	-.104
Self administered	CASI	Mail
Reliability coefficient	-.0038	.000
Validity coefficient	-.0383	.000

This presentation suggests the following order in quality with regard to validity and reliability:

1 Mail
2 CASI
3 PAPI/Telephone
4 CATI/CAPI

The differences between Mail and CASI are minimal, on the other hand, differences between these two and the PAPI/Telephone or CAPI /CATI are large. It should be mentioned that other quality criteria in the mode of data collection choice should also be considered, such as unit nonresponse and item nonresponse. In general, Mail surveys have lower response rates although the use of the total design method can reduce the problem (Dillman 1978, 2000). Therefore, the results suggest that a tradeoff between quality, with respect to reliability and validity, and item nonresponse has to be made.

12.4 Conclusions, Limitations, and the Future

Our results show that within and between questionnaires there is a wide variation in reliability and validity. In particular the following choices have a large effect on reliability and/or validity:

- The use of direct questions has a large positive effect on reliability and a smaller negative effect on validity when compared with batteries containing statements.

- The use of gradation has a large positive effect on reliability and a smaller negative effect on validity.
- The use of frequencies or magnitude estimation has a large positive effect on reliability and an almost equally large negative effect on validity.
- The use of lines as response modality has a large positive effect on reliability and a much smaller negative effect on validity.
- The more categories a response scale has, the greater the positive effect on reliability is. However, it also has a much smaller negative effect on validity.
- Allowing for high frequencies has both a large negative effect on reliability and validity.
- The use of interviewers has both a large negative effect on reliability and validity.

This analysis is an intermediate result; so far 87 studies have been reanalyzed with a total of 1023 survey items, which is not enough to evaluate all variables in detail. (The database is a work in progress that will be extended in the future with survey items that are at present underrepresented.) Important limitations to consider are listed below:

- Only the main categories of the domain variable have been taken into account.
- Requests concerning consumption, leisure, family, and immigrants could not be included in the analysis.
- The concepts of norms, rights, and policies have been given too little attention.
- The request types of open-ended requests and WH requests have not yet been studied.
- Mail and Telephone interviews were not sufficiently available to be analyzed separately.
- There is an overrepresentation of requests formulated in the Dutch language.
- Only a limited number of interactions and nonlinearities could be introduced.

Nevertheless, taking these limitations into account, the analysis can remarkably explain 47% of the reliability variance and 61% of the validity. In this respect, it is also relevant to refer to the standard errors of the regression coefficients which are relatively small, indicating that the correlations between the variables used in the regression as independent variables are relatively small.

If one considers that all estimates of the quality criteria contain errors while in the coding of the survey item characteristics errors are also made, the high explained variance is very promising.

This does not mean that we are satisfied with this result. Certainly, further research is needed, as we have indicated above, but for the moment Table 12.1 is the best summary of our knowledge about the effects of the questionnaire design choices on reliability and validity.

EXERCISES

1. Are there results that you did not expect in Table 12.1?
2. Can you imagine that there are effects on survey quality that depend on the value of other variables? For example, German sentences are in general much longer than Dutch sentences. So the effect of the number of words of the request for an answer is probably different in Dutch than in German questionnaires. Could you think of other variables with such conditional or interaction effects?
3. What are the best options for decisions that have to be made in survey design, given the results presented in Table 12.1?
4. Some choices cannot be optimized. Can you identify them?
5. The following request is presented:
   On the whole, how satisfied are you with the way democracy works in Britain? Are you... READ OUT
   1 Very satisfied?
   2 Fairly satisfied?
   3 Fairly dissatisfied?
   4 Very dissatisfied?
   (8 don't know)
   a Check which method options have been chosen in this case
   b Specify the effects they have on the reliability and validity according to Table 12.1
   c Determine the total reliability and validity for this question
6. Did the author of this question select the optimal choices?
7. Can you suggest improvements to the question?

APPENDIX 12.1 OVERVIEW OF THE EXPERIMENTS USED IN THE ANALYSES IN 2001

Country	Number	Year	Design	Data collection	Organization	Topics
NL	101	92	3×2×2	Mail/telep	STP	Seriousness of crimes
NL	102	91	4×2×2	Telep	STP	Political efficacy (europe)
NL	103	92	3×2×2	Mail/telep	NIMMO	Europe
NL	104	92	4×2×2	Tel	NIMMO	Satisfaction
NL	105	91	4×2×2	Mail	NIMMO	Satisfaction
NL	106	92	4×2×2	Mail	NIMMO	Satisfaction
NL	107	92	4×2×2	Mail/telep	NIMMO/STP	Satisfaction
NL	108	89	4×3	Telep	NIPO	Satisfaction
NL	109	91	4×2×2	Telep	STP	Satisfaction
NL	110	91	3×2×2	Telep	STP	Satisfaction
NL	111	92	3×2×2	Mail/telep	STP	Values
NL	112	91	3×2×2	Telep	STP	Values: comfort/self respect/status
NL	113	91	3×2×2	Telep	STP	Values: family/ambition/ independence
NL	114	91	3×2×2	Telep	STP	Values: comfort/self respect/status
NL	115	91	3×2×2	Telep	STP	Values: family/ambition/independence
NL	116	91	3×2×2	Telep	STP	Values: comfort/self/respect/status
NL	117	91	3×2×2	Telep	STP	Values: family/ambition/independence
NL	118	91	3×2×2	Telep	STP	Values: comfort/self respect/status
NL	119	91	3×2×2	Telep	STP	Values: family/ambition/independence
NL	120	91	3×2×2	Telep	STP	Seriousness of crimes
NL	124	91	3×2×2	Telep	STP	Seriousness of crimes
NL	121	91	3×2×2	Telep	STP	Seriousness of crimes
NL	122	91	3×2×2	Telep	STP	Seriousness of crimes
NL	124	91	3×2×2	Telep	STP	Seriousness of crimes
NL	125	91	3×2×2	Telep	STP	Seriousness of crimes
NL	–	90	–	Telep	STP	EU membership
NL	126	91	4×2×2	Telep	STP	EU membership
NL	127	91	3×3	Telep	STP	Crimes 1,2,3
NL	128	91	3×3	Telep	STP	Crimes 4,5,6
NL	129	91	3×3	Telep	STP	Crimes 7,8,9

Country	Number	Year	Design	Data collection	Organization	Topics
NL	–	88	–	Telep	NIPO	TV/Olympic games
NL	130	88	3×3	Telep	NIPO	Trade-unions
NL	131	88	3×3	Telep	NIPO	Trade-unions
NL	132	88	3×3	Telep	NIPO	Trade-unions
NL	133	88	3×3	Telepanel	NIPO	Trade-unions
NL	135	92	3×2×2	Telepanel	STP	Satisfaction
NL	136	92	3×2×2	Telepanel	STP	Satisfaction
NL	137	92	3×2×2	Telepanel	STP	Satisfaction
NL	138	92	3×2×2	Telepanel	STP	Satisfaction
NL	139	92	3×2×2	Telepanel	STP	Work conditions
NL	140	92	3×2×2	Telepanel	STP	Work conditions
NL	141	92	3×2×2	Telepanel	STP	Work conditions
NL	142	92	3×2×2	Telepanel	STP	Work conditions
NL	143	92	3×2×2	Telepanel	STP	Living conditions
NL	144	92	3×2×2	Telepanel	STP	Living conditions
NL	145	92	3×2×2	Telepanel	STP	Living conditions
NL	146	92	3×2×2	Telepanel	STP	Living conditions
NL	–	88	3×3	Telepanel	STP	TV watching
NL	147	88	3×3	Telepanel	STP	Eval. TV programs
NL	148	88	3×3	Telepanel	STP	Use of the TV
NL	149	88	3×3	Telepanel	STP	Reading
NL	150	88	3×3	Telepanel	STP	Eval. Policies
NL	151	88	3×3	Telepanel	STP	Estimate ages
NL	152	88	3×3	Telepanel	STP	Political participation
NL	153	88	3×3	Telepanel	STP	Estimation of income
NL	154	96	4×2×2	Telepanel	STP	Trust
NL	155	96	4×2×2	Telepanel	STP	F-scale
NL	156	96	3×2×2	Telepanel	STP	Threat
NL	–	96	–	Telepanel	STP	Ethno/wave 2
NL	–	96	–	Telepanel	STP	Ethno/wave 3
NL	–	98	Sbmt	Telephone	Nimmo	Voting
NL	157	96	4×2×2	Telepanel	STP	Outgroup
NL	158	96	4×2×2	Telepanel	STP	Ingroup
NL	159	96	4×2×2	Telepanel	STP	Trust
NL	–	96	–	Telepanel	STP	Ethno/wave 2
Belg	801	89	5×3	FTF[a]	KUL	Satisfaction
Belg	802	97	3×3	ftf/mail	KUL	threat

Country	number	year	design	data collection	organization	topic
Belg	803	97	3×3	FTF/mail	KUL	Outgroup
Belg	804	97	4×3	FTF/mail	KUL	Ingroup
Austria	1	92	4×3	FTF	IFES	Party politics
Austria	2	92	4×3	FTF	IFES	Econ. Expectations
Austria	3	92	4×3	FTF	IFES	Postmaterialism
Austria	–	92	4×3	FTF	IFES	Pschy problems
Austria	4	92	4×3	FTF	IFES	Social control
Austria	5	92	4×4	FTF	IFES	Party politics
Austria	6	92	4×3	FTF	IFES	Social control
Austria	7	92	4×3	FTF	IFES	Eu evaluation
Austria	8	92	3×3	FTF	IFES	Life satisfaction
Austria	9	92	3×3	FTF	IFES	Political parties
Austria	10	92	4×3	FTF	IFES	Conf in institutions
USA	1	79	4×3	FTF	ISR	Finances, business, health, news
USA	2	79	4×3	FTF	ISR	Finances, business, health,news
USA	3	79	4×3	FTF	ISR	Same as 1
USA	4	79	4×3	FTF	ISR	Same as 2
USA	5	81	3×3	FTF	ISR	Finance, business, Health, lastyear
USA	6	81	3×3	FTF	ISR	Finances, business, health, next year
USA	7	81	4×3	FTF	ISR	Satisfaction with life, etc
USA	8	86	2×2×3	FTF	ISR	Health/income
USA	9	86	3×2×2	FTF	ISR	Savings/transport/safety
USA	10	86	3×2×3	FTF	ISR	Restless/depressed/relaxed
USA	11	86	3×2×3	FTF	ISR	Exited/restless/energy
USA	12	86	4×2×2	FTF	ISR	Health/income
USA	13	86	5×2×2	FTF	ISR	Health/house/income/friends/life in general

[a] Face-to face interview.

This Page Intentionally Left Blank

Part IV

Applications in social science research

This part will recommend how the knowledge presented in this book can and also should be used. We will illustrate the following applications:

1. The prediction and improvement of survey requests by the survey quality predictor (Chapter 13)
2. Evaluation of the quality of concepts by postulation (Chapter 14)
3. Correction for measurement errors in survey analysis (Chapter 15)
4. Coping with measurement errors in cross-cultural research (Chapter 16)

This Page Intentionally Left Blank

CHAPTER 13

Prediction and improvement of survey requests by the survey quality predictor (SQP)[1]

Suppose that a survey designer would like to conduct a survey and, before going into the field, would like to evaluate the quality of the proposed survey items of the questionnaire using the information summarized for this purpose in Table 12.1. This would necessitate coding all the items on the variables in the classification system, and after that applying the prediction from Table 12.1 on all these items, in order to determine the total score for reliability and validity. This clearly is a great deal of work. It would, therefore, be advantageous to have a computer program that could evaluate all requests of a questionnaire automatically on a number of characteristics. The designer of the survey could then, on the basis of this information, determine which items require further study, in order to improve the quality of the data to be collected.

The *survey quality predictor* (SQP) has been developed by Van der Veld et al. (2000) as prototype of a program to help survey designers for Dutch-language questionnaires. The program contains the following functions: (1) reading and automatic coding of the survey items in a questionnaire, (2) prediction of the quality of proposed survey items on a number of criteria, (3) providing information about the effects of the different choices, and (4) providing suggestions for improvement of these items.

It is evident that it is complex and time consuming to develop such programs for many languages, since the automatic coding of the requests needs to be precise and in the context of the grammatical rules of the language in question. Therefore, a semiautomatic version of SQP has also been proposed by Oberski et al. (2005) where the coding is administered by the researcher answering requests about the characteristics of the survey item being addressed. The program then uses this information to provide an estimate of the reliability and validity and the total quality of a request on the basis of the results of Table

[1] This chapter is based on a paper by W.E. Saris, W. van der Veld and I.N. Gallhofer (2004) *Development and Improvement of questionnares using predictions of reliability and validity.* In Presser te al. *Methods for testing and evaluating survey questionnares.* Wiley, 275–299, the paper has been reprinted with permission of John Wiley Sons Inc.

12.1. The latter version of SQP can make quality predictions for requests in three different languages. Both procedures will be illustrated below.

13.1 THE AUTOMATIC SURVEY QUALITY PREDICTOR

Current questionnaires are written with text processors and are computer-readable. This, in principle, facilitates the automatic text analysis, which consists of automatically classifying different survey items with respect to their characteristics, where previous knowledge of how much they affect the quality of the data collected is available. In fact, for the survey item characteristics studied by Scherpenzeel and Saris (1997), an automatic coding procedure has been developed by Van der Veld et al. (2000). In the prototype of the program of SQP the proposed coding procedure was tested and implemented. After a file with survey items is read by the program it codes the survey items almost instantaneously. The characteristics that cannot be coded automatically generate request prompts that are presented to the user, whose answers are then stored. All these codes are then used in the next steps for survey quality prediction.

The prototype of SQP has successfully demonstrated that reliability, validity, method effect, and total quality of a survey item can be predicted in an automatic way. The initial computer screen image of the predictor is shown in Figure 13.1. It shows that for the survey item named 'tvsp1cat' [reading: "Hoeveel mensen denkt u dat gewoonlijk naar Studio Sport kijken?," i.e., "How many people, do you think, usually watch Studio Sport (on TV)?"] the reliability is .65, the validity is .81, the method effect is .59, and the score for the total quality of the item amounts to .53, which means that 53% of the variance in the observed variable is explained by the latent variable of interest.

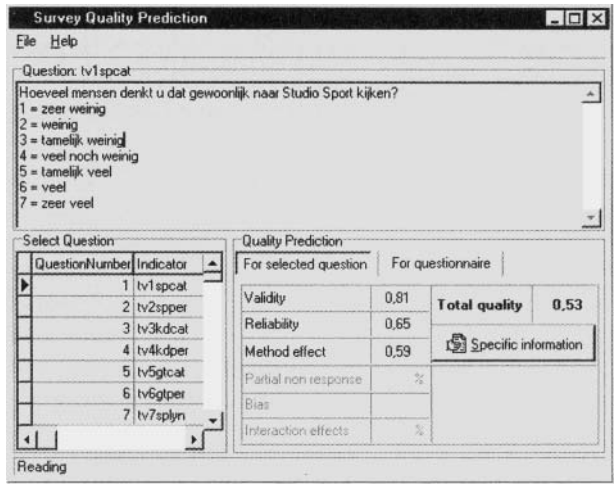


FIGURE 13.1: *Survey quality prediction using the MTMM data in SQP.*

These results are calculated[2] by automatically coding the survey item and applying the linear prediction equation discussed in Chapter 12. In the same way, nonresponse and bias could in principle be predicted on the basis of the study done by Molenaar (1986). The possibility of more quality indicators is limited only to the aspirations of future research.

On the basis of the information provided for each survey item, a user of the program can decide whether the survey item satisfies the required quality level regarding the different criteria. If not, the user can also obtain information about what might be the cause(s) of the problem. Given that information is available about the effect of the different item characteristics (or choices), the program can also indicate what the contribution of each of these choices is on the quality predictors and can also suggest improvements that require changes in these choices.

Although an automatic program is ideal, currently it is still not possible to realize a prediction program taking into account all the relevant characteristics with specific language features. Given these considerations, a semiautomatic program SQP has been developed in the interim.

13.2 THE SEMIAUTOMATIC SQP

At the beginning of the session, the user of the semiautomatic SQP program has to answer a series of requests about the choices he/she has made while developing the survey item. Next the program estimates the reliability, validity and total quality of the request on the basis of the information provided and the general knowledge that has been stored in Table 12.1. To date the program can predict the quality of requests in English, German, and Dutch. This approach has been applied to one of the requests proposed for the European Social Survey discussed in Table 10.2. The request used in this example reads as follows:

> *On the whole, how satisfied are you with the way democracy works in Britain?*
> *Are you ... READ OUT*
> 1. *Very satisfied*
> 2. *Fairly satisfied*
> 3. *Fairly dissatisfied*
> 4. *Very dissatisfied?*
>
> *(8 don't know)*

In the following sections we will first illustrate how to use the program and finish by presenting SQP results for the request.

---

[2] These calculations were done taking into account that the [method effect]^2 = 1- [validity coefficient]^2, and that the total quality coefficient = [reliability coefficient] × [validity coefficient].

13.2.1 How to use the semiautomatic program SQP

The program SQP and the associated codebook can be found on the Website www.sqp.nl. The start of the program is successful if your computer screen looks similar to Figure 13.2. At the top left-hand corner the program suggests to click on "Insert a new item" to begin. This text is mentioned in a box below the first text. By clicking on this box you will see a new dialog box as shown in Figure 3.3.

In this screen you have to specify a name of the item. Then you have to split the survey item into three parts: an introduction, a request, and answer categories. However, as was discussed in earlier chapters, it is not necessary that all three components be present. The only component that is necessary is the request. The other components that are not present can be left with an empty space.

FIGURE 13.2: *Screen 1 of the semiautomatic SQP.*

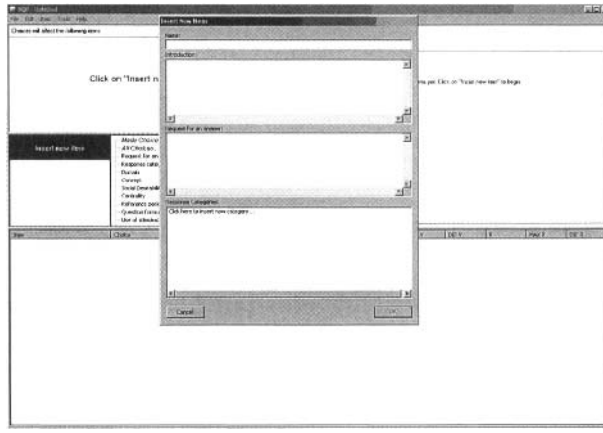


FIGURE 13.3: *Screen 2 of the semiautomatic SQP.*

We acknowledge that the distinction between the introduction and the request for an answer can be difficult, as has been discussed in previous chapters. We suggested that researchers also have the option of providing definitions or instructions before the request. Currently SQP makes a distinction only between the request and other parts of the survey item. Therefore all text before the request formally belongs to the introduction.

However, this rule may not be enough to resolve every type of request. For example, the previously mentioned one begins with the text *"And on the whole, how satisfied are you with the way democracy works in Britain?"* We suggest placing it in the introduction because it is not the request that will be answered. The correct request is

Are you ... (READ OUT)

1. *Very satisfied*
2. *Fairly satisfied*
3. *Fairly dissatisfied*
4. *Very dissatisfied?*

As you can see on the screen, the introduction of the request indicates what the object of the satisfaction should be. Technically, this request embeds the answer categories because the whole text has to be read aloud to the respondents. However, these answer categories also need to be specified in the third part of this screen page. Please take note that the request does not contain the response category "*don't know*," because that particular part of the text is not read aloud to the respondents. It also does not belong to the response scale if the number of categories is counted, since in this case we only count the number of categories presented on the ordinal scale, where "don't know" does not qualify. Whether this type of response category is or is not present is asked in a separate prompt. The different texts have to be typed into their respective prompts so that the program can automatically count the number of words and number of categories. After the texts have been typed into the appropriate prompts, click on "OK" in order to move on to the next screen, where the coding of the characteristics of the survey item starts.

After clicking "OK," you will be presented with an overview of the characteristics to be coded for the request (Figure 13.4). You will find a summary of the information you provided on this screen. In addition, you see at the right side of the screen the name of the first characteristic to be coded, called the "domain" of the survey item and its coding possibilities. Our example should be coded under the "national politics" category.

After indicating your choice, click on "Next" at the right side of the text for the next screen to provide the name and coding possibilities for the second characteristic called "the concept measured." Keeping with our example, we code the concept of the request as a "feeling." Click on "Next," and the program proceeds to the next characteristic. This is a process that continues until all characteristics are coded. At all times you can see which characteristics have or have not been coded. On the bottom half of the screen the names of the characteristics to

be coded are mentioned; in addition, the characteristics that have been coded are indicated by a green sign in front of them with the specified code at the right side. The characteristics still in need of coding are denoted by a red sign and are without any code, while gray signs mean that this choice does not apply. SQP always gives you the opportunity to go back and make corrections. If you click on the name of the characteristic chosen for correction, you will again be able to change the code for this characteristic.

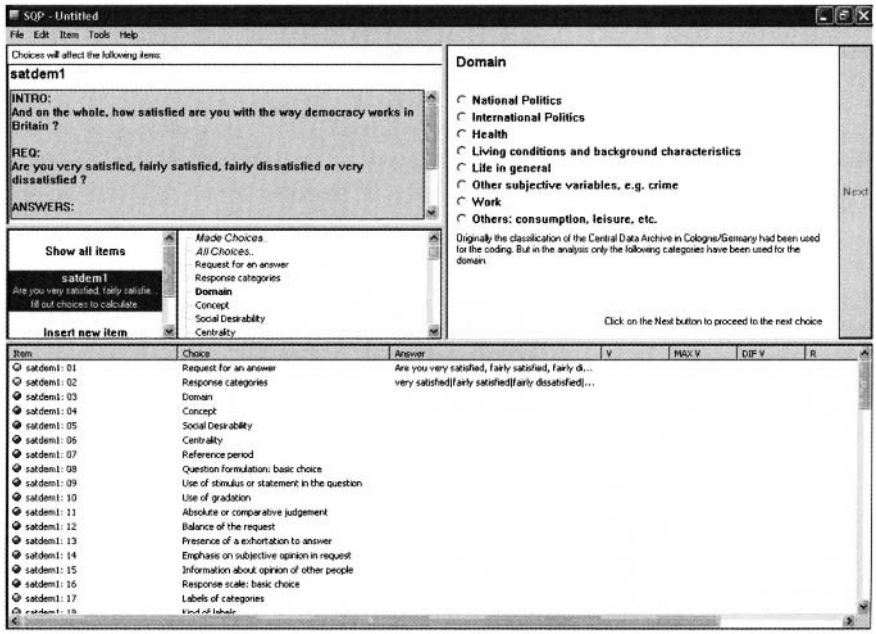


FIGURE 13.4: *Screen 3 SQP for the survey item satdem1.*

After all characteristics have been coded, SQP automatically computes the reliability and validity coefficients, as well as method effects. In order to get a summary of the data quality predictions, click on the phrase "Summary Survey," which is found on the last line of the screen in the middle; or on "Made Choices" at the top of the same screen. In our example we did the latter, and the result for the satdem1 request is presented in Figure 13.5.

13.2.2 Results obtained with SQP

The coding of the request and the addition of the effects of the different characteristics of the request leads to an estimate of the reliability coefficient of .735 and of .917 for the validity coefficient. It is interesting to compare this result with the results obtained for the same request (T_{31}) in the MTMM experiment summarized in Table 10.4. There, the reliability coefficient was estimated to be .81 and the validity coefficient, .95. These estimates are slightly higher

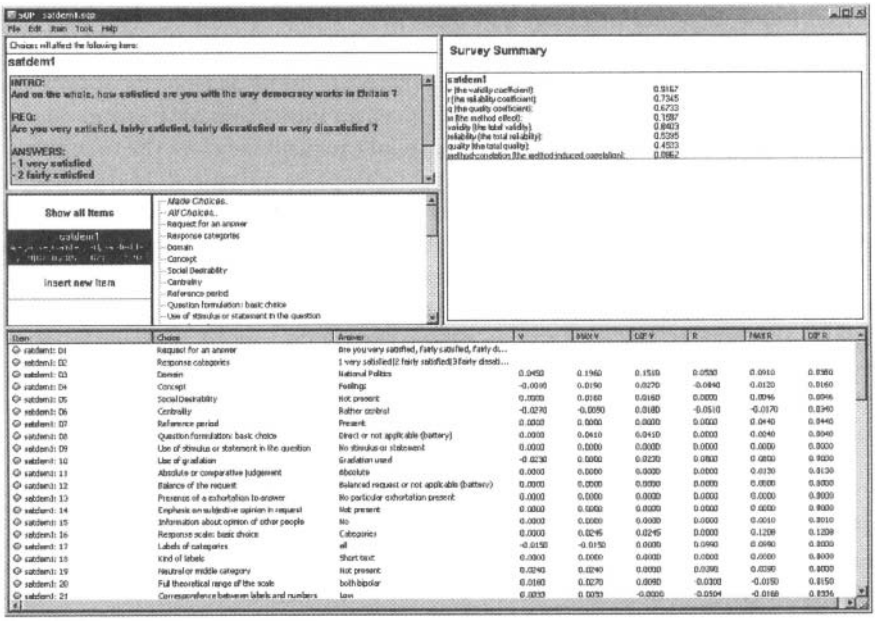


FIGURE 13.5: *Screen 4 of SQP for item satdem1.*

than the predictions in Figure 13.5, but it should be stated that the coefficients in these experiments are overestimated by .042 with respect to reliability and by .062 with respect to validity, because of the presence of repeated observations[3]. Correcting for overestimation, the values would be expected to be around .768 for reliability and .868 for validity. These corrected results of Table 10.3 are close to the values predicted on the basis of the MTMM experiment reported in Figure 13.5. It must be noted that now our prediction is based on the existing knowledge presented in Table 12.1 without collecting any new data. It should be clear that the agreement of the results between the two approaches will not always be this similar. There will certainly be cases where the differences are larger because the prediction cannot be perfect. The explained variance of 50% for reliability and 60% for validity indicates that the predictions will be quite good. However, a word of caution is appropriate, in that errors in the coding of the requests cannot be avoided and this is also true for the estimates of the reliability and validity.

[3] It may be observed that the correlations between variables in a questionnaire increase when the requests are repeated. This suggests that people realize better the relation between these requests. This could lead to higher estimates of the reliability and validity. This is one of the reasons that a part for design effects is included in Table 12.1. The effects in that part indicate to what extent the reliability and validity are too high because of design effects. Therefore, the correction factors are given in Table 12.1 for single requests. These correction factors have been applied here.

13.3 IMPROVEMENT OF SURVEY REQUESTS

In general one would prefer observed variables with reliability and validity coefficients, which are standardized coefficients, to be as close as possible to 1. The product of these two coefficients gives an indication of the total quality coefficient of the measure, and the square of the product gives the variance of the observed variable explained by the variable to be measured. In the specific case discussed above, the explained variance is 45.3%. This result is rather low but close to the mean value in survey research observed by Alwin and Krosnick (1991). Thus 45.3% of the variance in the observed variable comes from the variable of interest. The rest (54.7%) is error: the systematic error variance, due to the method effect, is 8.6 % ($.735^2 \times .4^2$) and the random error variance is 46.1 % ($1\text{-}.735^2$). This shows that for this request the random error is considerable and the quality of the measure is therefore rather low. It is highly desirable to try to improve at least the reliability of the measures and to keep the validity approximately the same or if feasible to improve that quality criterion as well.

There are two possible ways of improving a survey item: (1) by changing the characteristics which have the most negative effect on the quality criteria; and (2) by changing the characteristics that can lead to an improvement of the quality criteria.

In order to see where changes should be made, a table from SQP can be obtained that provides the effects of the different choices on the quality criteria, the maximal effects that could have been obtained, and the difference between the two. This table can be obtained in two ways; One possibility is to click on the text "Made Choices" at the top of the screen in the middle of the page of the SQP program (see Figure 13.5). The results are presented in the lower panel of Figure 13.5. The second way to obtain such a table is to export the file to HTML. This option can be found under the file menu. The file that is created by following the second option is Table 13.1. Both tables contain the same information. An advantage of the first choice is that the user is still in the program and can make corrections if a mistake has been made in the coding. The advantage of the second choice is that this table can be printed and therefore can be studied quietly without staying in the program. We now turn our attention to Table 13.1.

At the top of Table 13.1, we find the summary results that we discussed previously. Below these results, the effects of the different characteristics of the request are indicated. In the first column the names of the characteristics ("choice") of the request are given that were also used in Table 12.1, while the next column ("answer") lists the code for that characteristic. The following columns respectively present the contribution to the validity score ("v"), the maximal possible contribution ("max v") and the difference between the two previous columns ("Diff v"). These are followed by columns addressing the same information but now for the reliability coefficient: the contribution ("R"), the maximum value ("Max R"), and the difference between the two previous columns ("Diff R").

Table 13.1: WinSQP survey summary: satdem1.SQP

Item name:	validity coefficient (v)	reliability coefficient (r)	quality coefficient (q)	method effect (m)	total validity (validity)	total reliability (reliability)	total quality (quality)	method-induced correlation (method-correlation)
satdem1	0.9167	0.7345	0.6733	0.1597	0.8403	0.5395	0.4533	0.0862

satdem1

Choice:	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Request for an answer	"Are you very satisfied, fairly satisfied, fairly dissatisfied or very dissatisfied ?"	0.0000			0.0000		
Response categories	1 very satisfied 2 fairly satisfied 3 fairly dissatisfied 4 very dissatisfied 8 don't know	0.0000			0.0000		
Domain	National Politics	0.0450	0.1960	0.1510	0.0530	0.0910	0.0380
Concept	Feelings	-0.0080	0.0190	0.0270	-0.0040	0.0120	0.0160
Social desirability	Not present	0.0000	0.0160	0.0160	0.0000	0.0046	0.0046
Centrality	Rather central	-0.0270	-0.0090	0.0180	-0.0510	-0.0170	0.0340
Reference period	Present	0.0000	0.0000	0.0000	0.0000	0.0440	0.0440
Request formulation: basic choice	Direct or not applicable (battery)	0.0000	0.0410	0.0410	0.0000	0.0040	0.0040
Use of stimulus or statement in the request	No stimulus or statement	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Use of gradation	Gradation used	-0.0230	0.0000	0.0230	0.0800	0.0800	0.0000
Absolute or comparative judgement	Absolute	0.0000	0.0000	0.0000	0.0000	0.0130	0.0130
Balance of the request	Balanced request or not applicable (battery)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Presence of a exhortation to answer	No particular exhortation present	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 13.1: continued

Choice:	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Emphasis on subjective opinion in request	Not present	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Information about opinion of other people	No	0.0000	0.0000	0.0000	0.0000	0.0010	0.0010
Response scale: basic choice	Categories	0.0000	0.0245	0.0245	0.0000	0.1208	0.1208
Labels of categories	all	-0.0150	-0.0150	0.0000	0.0990	0.0990	0.0000
Kind of labels	Short text	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Neutral or middle category	Not present	0.0240	0.0240	0.0000	0.0390	0.0390	0.0000
Full theoretical range of the scale	both bipolar	0.0180	0.0270	0.0090	-0.0300	-0.0150	0.0150
Correspondence between labels and numbers	Low	0.0033	0.0033	-0.0000	-0.0504	-0.0168	0.0336
Symmetry of response scale	Symmetric / not applicable	0.0000	0.0220	0.0220	0.0000	0.0260	0.0260
Order of labels	First label positive	0.0280	0.0280	0.0000	-0.0160	-0.0080	0.0080
Number of fixed reference points	0	0.0000			0.0000		
Number of categories	4.000000	-0.0080			0.0560		
Don't know possibility	Only registered	-0.0040	-0.0020	0.0020	-0.0140	-0.0070	0.0070
Request form	request present	0.0120	0.0115	-0.0005	0.0270	0.0272	0.0002
Instruction for the respondent	No	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Instruction for the interviewer available?	Yes	0.0060	0.0060	0.0000	-0.0001	0.0000	0.0001
Extra motivation, info or definition available?	No	0.0000	0.0000	0.0000	0.0000	0.0070	0.0070
Introduction available?	Yes	-0.0110	0.0000	0.0110	0.0060	0.0060	0.0000

Table 13.1: continued

Choice	Answer	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Introduction	"And on the whole, how satisfied are you with the way democracy works in Britain?"	0.0000			0.0000		
Request in introduction available?	Yes	-0.0210	0.0000	0.0210	-0.0450	0.0000	0.0450
Number of words in introduction	16.000000	0.0224			-0.0208		
Number of sentences in introduction	1.000000	0.0000			0.0000		
Mean no. words per sentence in introduction	16.000000	-0.0059			0.0010		
Number of subordinated clauses in introduction	0	0.0000			0.0000		
Knowledge provided	No extra information provided	-0.0060	-0.0060	0.0000	-0.0130	-0.0130	0.0000
Number of sentences in the request	1.000000	-0.0083			0.0127		
Number of words in request	12.000000	-0.0156			0.0096		
Mean number of words per sentence in request	12.000000	0.0132			-0.0264		
Mean number of syllables per word in request	2	-0.0208			-0.0650		
Number of subordinate clauses in the request	0	0.0000			0.0000		
Total number of nouns in request	1	0.0000			0.0000		
Number of abstract nouns in request	0	0.0000			0.0000		
Mode of data collection	Paper and pencil	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 13.1: continued

Choice	Answer	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Interviewer administered or self completion	Interviewer	-0.1040	0.0000	0.1040	-0.0510	0.0000	0.0510
Presentation of information	Oral	0.0253	0.0253	0.0000	0.0104	0.0104	0.0000
Item placed in battery	No	0.0000	0.0289	0.0289	0.0000	0.0000	0.0000
Position of the item in the questionnaire	55	0.0231			0.0165		
Language used in the questionnaire	English	-0.0030	0.0000	0.0030	-0.0720	0.0000	0.0720

As we stated previously, the lack of quality of this request is due mainly to its low score on reliability. Therefore, we need to focus on possibilities to increase the reliability without decreasing validity. The first possibility that Table 13.1 suggests is changing the decision with respect to the characteristic "domain." One could get an improvement of .0380 in reliability with another choice. However, most of the time such a change is not an option. For example, if a request about "national politics" needs to be asked, another topic cannot be considered. For the same reasons, other options with respect to the concept, social desirability, centrality and reference period are also not possible because they are predetermined by the topic of research.

The next possibility to consider in order to increase reliability addresses the response scale. Our sample request employed a category scale. The Dif.R score of .1208 suggests that another scale could give a much better score, especially the frequency scale (see Table 12.1). There are, however, two reasons for not changing the scale: (1) this increase in reliability would come from a scale, which in our case is not appropriate for the request; (2) the new scale reduces validity by .0959, and that, in turn, counteracts the improvement of the reliability considerably. The option of employing magnitude estimation would have an effect similar to that of the frequency scale. Therefore, the line production would then be a better alternative (see Table 12.1). However, what we are looking for is a similar improvement without a negative effect on validity and without changing the scale type. For our sample request we see the following possibilities for improvements:

1. To use a category scale with categories 0 until 10 in order to increase the scale from 4 to 11 categories.
2. To use fixed reference points, such as by labeling the endpoints as "completely satisfied" and "completely dissatisfied" and the middle category as "neither satisfied nor dissatisfied."

Both corrections will result in a considerable improvement in reliability without having a significant negative effect on validity.

The corrections specified above lead to the following reformulation of our sample request:

> *And on the whole, how satisfied are you with the way democracy works in Britain?*
> *Choose a number on a scale from 0 and 10, where 0 means completely dissatisfied and 10 means completely satisfied and 5 means neither satisfied nor dissatisfied.*

Here it is important to note that the reliability will not increase by (11- 4) × .0135 because of the increase in the number of categories (see Table 12.1) plus 3 × .0147 due to the introduction of three fixed reference points. The request also changes because of differences in the number of words, the direction of

Table 13.2: WinSQP survey summary: satdem2.SQP

Item name:	validity coefficient (v)	reliability coefficient (r)	quality coefficient (q)	method effect (m)	total validity (validity)	total reliability (reliability)	total quality (quality)	method-induced correlation (method-correlation)
satdem2	0.8669	0.8992	0.7795	0.2485	0.7515	0.8086	0.6077	0.2010

satdem2

Choice:	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Request for an answer	"Choose a number on a scale from 0 – ` 10, where 0 means completely dissatisfied and 10 means completely satisfied and 5 means neither satisfied nor dissatisfied."	0.0000			0.0000		
Response categories		0.0000			0.0000		
Domain	National Politics	0.0450	0.1960	0.1510	0.0530	0.0910	0.0380
Concept	Feelings	-0.0080	0.0190	0.0270	-0.0040	0.0120	0.0160
Social desirability	Not present	0.0000	0.0160	0.0160	0.0000	0.0046	0.0046
Centrality	Rather central	-0.0270	-0.0090	0.0180	-0.0510	-0.0170	0.0340
Reference period	Present	0.0000	0.0000	0.0000	0.0000	0.0440	0.0440
Request formulation: basic choice	Direct or not applicable (battery)	0.0000	0.0410	0.0410	0.0000	0.0040	0.0040
Use of stimulus or statement in the request	No stimulus or statement	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Use of gradation	Gradation used	-0.0230	0.0000	0.0230	0.0800	0.0800	0.0000
Absolute or comparative judgement	Absolute	0.0000	0.0000	0.0000	0.0000	0.0130	0.0130

Table 13.2: continued

Choice	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Balance of the request	Balanced request or not applicable (battery)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Presence of a exhortation to answer	No particular exhortation present	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Emphasis on subjective opinion in request	Not present	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Information about opinion of other people	No	0.0000	0.0000	0.0000	0.0000	0.0010	0.0010
Response scale: basic choice	Categories	0.0000	0.0245	0.0245	0.0000	0.1208	0.1208
Labels of categories	some	-0.0100	-0.0150	-0.0050	0.0660	0.0990	0.0330
Kind of labels	Short text	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Neutral or middle category	Present/not applicable	0.0080	0.0240	0.0160	0.0130	0.0390	0.0260
Full theoretical range of the scale	both bipolar	0.0180	0.0270	0.0090	-0.0300	-0.0150	0.0150
Correspondence between labels and numbers	High	0.0011	0.0033	0.0022	-0.0168	-0.0168	0.0000
Symmetry of response scale	Symmetric / not applicable	0.0000	0.0220	0.0220	0.0000	0.0260	0.0260
Order of labels	First label negative or n/a	0.0140	0.0280	0.0140	-0.0080	-0.0080	0.0000

Table 13.2: continued

Choice	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Number of fixed reference points	3	0.0630			0.0450		
Number of categories	11.000000	-0.0220			0.1540		
Don't know possibility	Not present	-0.0060	-0.0020	0.0040	-0.0210	-0.0070	0.0140
Request form	instruction present	0.0040	0.0115	0.0075	0.0440	0.0272	-0.0168
Instruction for the respondent	Yes	-0.0150	0.0000	0.0150	-0.0130	0.0000	0.0130
Instruction for the interviewer	No	0.0000	0.0060	0.0060	0.0000	0.0000	0.0000
Response categories	No	0.0000	0.0000	0.0000	0.0000	0.0070	0.0070
Domain	Yes	-0.0110	0.0000	0.0110	0.0060	0.0060	0.0000
Concept	"And on the whole, how satisfied are you with the way democracy works in Britain ?"	0.0000			0.0000		
Social desirability	Yes	-0.0210	0.0000	0.0210	-0.0450	0.0000	0.0450
Centrality	16.000000	0.0224			-0.0208		
Reference period	1.000000	0.0000			0.0000		

Table 13.2: continued

Choice	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Request formulation: basic choice	16.000000	-0.0059			0.0010		
Use of stimulus or statement in the request	0	0.0000			0.0000		
Use of gradation	No extra information provided	-0.0060	-0.0060	0.0000	-0.0130	-0.0130	0.0000
Absolute or comparative judgement	1.000000	-0.0083			0.0127		
Balance of the request	27.000000	-0.0351			0.0216		
Presence of a exhortation to answer	27.000000	0.0297			-0.0594		
Emphasis on subjective opinion in request	1	-0.0104			-0.0325		
Information about opinion of other people	2	-0.0360			0.0280		
Response scale: basic choice	7	0.0000			0.0000		
Labels of categories	6	-0.0120			0.0026		
Kind of labels	Paper and pencil	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Neutral or middle category	Interviewer	-0.1040	0.0000	0.1040	-0.0510	0.0000	0.0510

Table 13.2: continued

Choice	Answer:	V:	Max. V:	Dif. V:	R:	Max. R:	Dif. R:
Full theoretical range of the scale	Oral	0.0253	0.0253	0.0000	0.0104	0.0104	0.0000
Correspondence between labels and numbers	No	0.0000	0.0289	0.0289	0.0000	0.0000	0.0000
Symmetry of response scale	55	0.0231			0.0165		
Order of labels	English	-0.0030	0.0000	0.0030	-0.0720	0.0000	0.0720

the order of the categories, and the instruction for the respondents. The SQP program can systematically and quickly evaluate the item again with respect to the quality, and the score results for the reformulated request are presented in Table 13.2.

Table 13.2 demonstrates that the program SQP can be used as a tool to improve the request with respect to quality; The reformulated request maintained approximately the same prediction validity, where the new score is .867 instead of the previous .917. However, a much higher prediction for the reliability of .899 is achieved, instead of the earlier score of .734. As a consequence, the total quality coefficient has also improved significantly from .673 to .779, this means an explained variance by the latent variable that changed 16% (from .45 to .61).

All the estimations and improvements are made on the basis of verified knowledge we have about the effect that choices have on the quality of survey requests with respect to reliability and validity. That is why we can call it "a scientific approach" to questionnaire development.

13.4 SUMMARY AND DISCUSSION

The purpose of the SQP program is to provide the user with a convenient method grounded in existing knowledge of the effects of survey item characteristics on the quality of survey items. This information is not summarized in tables, as is customary in academic literature, but in a user-friendly program for quality prediction. Furthermore, SQP goes beyond the traditional method of separately providing information for each item characteristic by treating all variables simultaneously. The SQP program thus provides the user with predictions about the quality of survey items and suggestions for the improvement of them. This is particularly useful since the user can apply the information in the design of his/her own questionnaire at a point when there is still time to improve it.

Such an expert system presents real advantages for survey researchers, since knowledge concerning the quality of survey items is dispersed in the methodological literature and it is too time-consuming to get at it during the questionnaire design phase. A program like SQP conveniently brings this information together and uses it to predict the quality of the survey items.

Although the above application of the program is very important, there are three more applications. The first one is that the estimates of the quality of the requests that measure the concepts by intuition can be used to estimate the quality of concepts by postulation. This will be discussed in the next chapter.

The second application is that the provided information can be used in the analysis of already collected data for the correction of measurement error on the basis of the best possible estimates of the measurement error provided by SQP. This approach has been discussed in Chapter 10, where we have indicated that the correlations between the concepts by intuition corrected for measurement error are obtainable. In Chapter 15 we will go further into the subject,

this time for a more general case such as for concepts by postulation.

The third application of the program and the correction for measurement error in general will be discussed in the last chapter, where problems encountered by cross-cultural comparisons will be discussed. In this context measurement errors play an important role and therefore information about the size of these errors is critical for a proper analysis of the data and its comparisons.

EXERCISES

1. Calculate with SQP the quality of the requests you have made for your own research project.
2. Determine which of the requests has a quality that is too low.
3. Specify alternative versions on the basis of the suggestions made by the program.
4. Determine of the new versions of the requests again their quality.
5. In the final report about your questionnaire you should indicate:
   a. Which requests have been asked for which concepts
   b. The original requests with their total quality scores
   c. Which requests have been changed and why
   d. The changed requests with their total quality scores.

CHAPTER 14

The quality of measures for concepts-by-postulation

In this chapter we pick up the discussion of the first chapter about concepts-by-postulation and concepts-by-intuition. This is important because often the concepts people want to study are not so simple that they can be operationalized by concepts-by-intuition. Several concepts-by-intuition are also combined into one concept-by-postulation in order to obtain a measure of the concept of interest with better reliability and/or validity. To date we have become familiar with the quality of measures for concepts-by-intuition. In this chapter we want to show how this information can be used to say something about the quality of measures for concepts-by-postulation. This is possible because a measure of a concept-by-postulation is an aggregate of several measures of concepts-by-intuition.

First we will introduce the possible structures of concepts-by-postulation. The logic is that the measures of concepts-by-postulation are based on concepts-by-intuition; and in order to determine the score of the concept-by-postulation, the relationships between the concepts-by-intuition and the concept-by-postulation need to be indicated. In some cases one can control whether the expected relationships indeed exist. Depending on the hypothesized relationships, different tests for the structure of the measures are performed. In this chapter we will discuss these different structures and indicate how the quality of the measures for the concepts-by-postulation can be determined on the basis of the estimated quality of the measures for the concepts-by-intuition.

14.1 THE STRUCTURES OF CONCEPTS-BY-POSTULATION

The structures of concepts-by-postulation and their tests have a strong research tradition. In fact the entire depth of this topic is beyond the scope of this chapter; for a more elaborate discussions of this topic we refer the reader to the following authors: Bollen (1989), Cronbach (1951), Guttman (1954), Hambleton and Swaminathan (1985), Messick (1989), and Nunnally and Bernstein (1994). We will proceed with the two most commonly used structures.

The first common structure type assumes that the concept-by-postulation is the variable that causes the correlations between the measures of the concepts-

by-intuition. This is, for example, the basic model for one of the definitions for attitude. Fishbein and Ajzen (1975) suggested that an attitude is a learned predisposition, where one consistently reacts in a positive or negative manner to an object. Therefore, a *factor analysis* or a more general *latent variable model* should apply, where the indicators are called *reflective* because they reflect the score of the latent variable.

The second most common structure is that the concept-by-postulation is an aggregate of several measures of different concepts-by-intuition. For example the concept socio-economic status (SES) is defined as a resultant of income, education, and occupational status. Since these indicators determine the definition of the concept, they are called *formative*. The initial formulation of this type of measurement model can be traced back to Blalock (1964). Other relevant sources are Bollen and Lennox (1991) and Edwards and Bagozzi (2000).

Many other measurement models have been developed for dichotomous or ordinal responses (see Guttman 1950; Mokken 1971; Rasch 1960). For a more general discussion of item response theory or IRT models, we can refer the reader to Hambleton and Swaminathan (1985). Other scales are developed for preference judgments, as, for example, the unfolding scale (Coombs 1964; Van Schuur 1988; Munnich 1998). These different scales do not fit very well the context of the models discussed here. So for these measurement models, we recommend the literature. Below we will concentrate on the first two models.

14.2 THE QUALITY OF MEASURES OF CONCEPTS-BY-POSTULATION WITH REFLECTIVE INDICATORS

If the concept-by-postulation is defined as a causal latent variable that affects a number of observable measures of concepts-by-intuition, then the latent variable models can describe the structure of the concept-by-postulation. This is a common assumption on which our example of the definition for attitude by Fishbein and Ajzen (1975) mentioned earlier rests. The frequently used concept of "political efficacy" has also been operationalized using this assumption. In the pilot study of the ESS the set of five questions presented in Table 14.1 is used to measure this concept.

Although some researchers speak of "political efficacy" as if it were one concept, most researchers assume that there are two concepts-by-postulation behind these items. The first three items are supposed to measure "subjective competence" or "internal efficacy, while the last two are supposed to measure "perceived system responsiveness", or "external efficacy" (Thomassen, 2002).

It is assumed in this case that the more "subjective competence" people have, the more likely it is that they will score lower on the first and higher on the second and third items. Therefore, it is also assumed that "subjective competence" explains the correlations between these three items. For the last two items, it is believed that people who think that the political system (via the politicians) facilitates them to influence it, will score higher on them. Also, the correlation between the last two items is assumed to be explained by a general

opinion about the responsiveness of the system. It is not clear if it can be expected that people with a higher "subjective competence" score, also perceive more "system responsiveness". If that were the case, there would be a strong indication that these two concepts-by-postulation are correlated, which is an important consideration for further analysis. Therefore it is necessary to test the factor model before further computations of the quality of the measures are made. Consequently we will proceed in the following manner. We will first test this type of model; Following this, we will discuss the way the measure for the concept-by-postulation is estimated and conclude with an estimation of the quality of the measures.

Table 14.1: Survey items for "political efficacy" in the first wave of the ESS

CARD C1: Using this card, how much do you agree or disagree with each of the following statements? First ... READ OUT

	Agree strongly	Agree	Neither agree nor disagree	Disagree	Disagree strongly	(Don't know)
"Sometimes politics and government seem so complicated that I can't really understand what is going on."	1	2	3	4	5	8
"I think I can take an active role in a group that is focused on political issues."	1	2	3	4		8
"I understand and judge important political questions very well."	1	2	3	4	5	8
"Politicians do not care much about what people like me think."	1	2	3	4	5	8
"Politicians are only interested in people's votes but not in their opinions."	1	2	3	4	5	8

14.2.1 Testing the models

Figure 14.1 represents the model mentioned above, where the unknown ("?") correlation between the latent traits that represent the two concepts-by-postulation ("subjective competence" and "perceived system responsiveness") are ready for testing. In this figure var1...var5 represent the responses to the requests 1 – 5, while $e_1 ... e_5$ represent random errors contained in these responses. Arrows represent direction of influence.

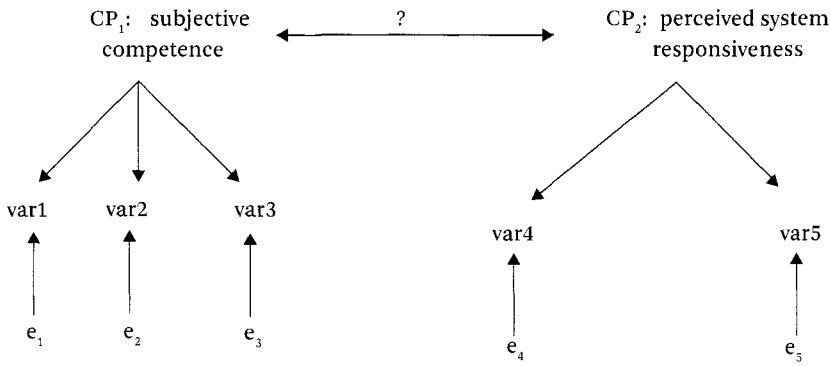


FIGURE 14.1: *The two-factor model for political efficacy as suggested in the literature.*

On one hand, if it is assumed that the two latent traits correlate perfectly (with correlation equal to 1), the model reduces to a one-factor model and then the factor can be called "political efficacy," and it does not make sense to speak about two separate concepts. On the other hand, if the correlation between the latent traits is 0, it means that knowing something about a respondent's "subjective competence" does not indicate anything about his/hers "perception of the system responsiveness." Both possibilities are theoretically acceptable; however, they lead to quite different measurement instruments. There is also the third possibility that there is a correlation between 1and 0.

In the Dutch pilot study of the ESS the correlations presented in Table 14.2 have been obtained for the five variables from Figure 14.1. The first three variables, that supposedly measure "subjective competence," correlate higher with each other than with the last two variables, which, in turn, correlate quite strongly with each other. However, there are also correlations different from zero between the two sets of variables. Using the procedures discussed in Chapter 10 different models can be estimated and tested. The model assuming that there is no correlation between the latent variables is rejected because this model cannot explain the correlations between the two sets of variables.

Table 14.2: Correlations between the 5 "political efficacy" variables in the Dutch ESS pilot study with a sample size of 230

	var1	var2	var3	var4	var5
var1	1.00				
var2	-0.33	1.00			
var3	-0.52	0.43	1.00		
var4	0.16	-0.13	-0.13	1.00	
var5	0.12	-0.18	-0.13	0.68	1.00

Also, the model assuming that that there is only one factor, "political efficacy" behind these observed variables is rejected because of the size of the residuals. However, the model allowing for the correlation between these two latent variables fits the data; this model is presented in Figure 14.2.

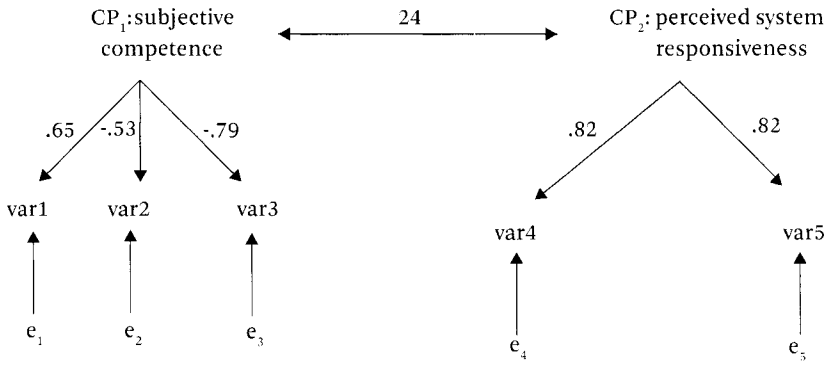


FIGURE 14.2: *The two-factor model for political efficacy as estimated on the basis of data from Table 14.1 assuming correlation between the two factors.*

Although the model in Figure 14.2, with the estimated values of the parameters neatly reproduces the correlation matrix, this does not mean that it necessarily is the correct model. In fact, after studying the errors in the data and finding that the batteries of agree/disagree items can produce quite large random and systematic measurement errors due to the method used, we think that the model is not correct. The systematic method effects also explain the correlations between these two sets of variables.

The alternative model for our data taking into account reliability, validity coefficients, and method effects is presented in Figure 14.3. In this model the lower part of the figure presents the result obtained by the MTMM experiment done in the Dutch pilot study of the ESS using three different methods. For our present purpose only the results for the A/D (agree/disagree) method are included. This part of the model is consistent with the model specified in

Chapter 9, where a distinction was made between the observed variables (var), the true scores (T), the variables of interest (F) and the "A/D method factor." New to this model is that above the variables of interest, another level of variables appears that represents the concepts-by-postulation (CP). Until now our analysis stopped at the level of the variables of interest, which represented concepts-by-intuition. Now we go further by looking at the concept-by-postulation that explains the correlations between the different concepts-by-intuition (corrected for measurement error). We assume that each of these observed variables has a unique component "u." This would not be the case if all the items only measured the same variable. Given that we know the reliability, validity, and method effects from earlier studies, the *consistency coefficients* for the relationships between the highest-order latent variables (representing the concepts-by-postulation) and the latent variables "F" (representing the concepts-by-intuition) can also be estimated.

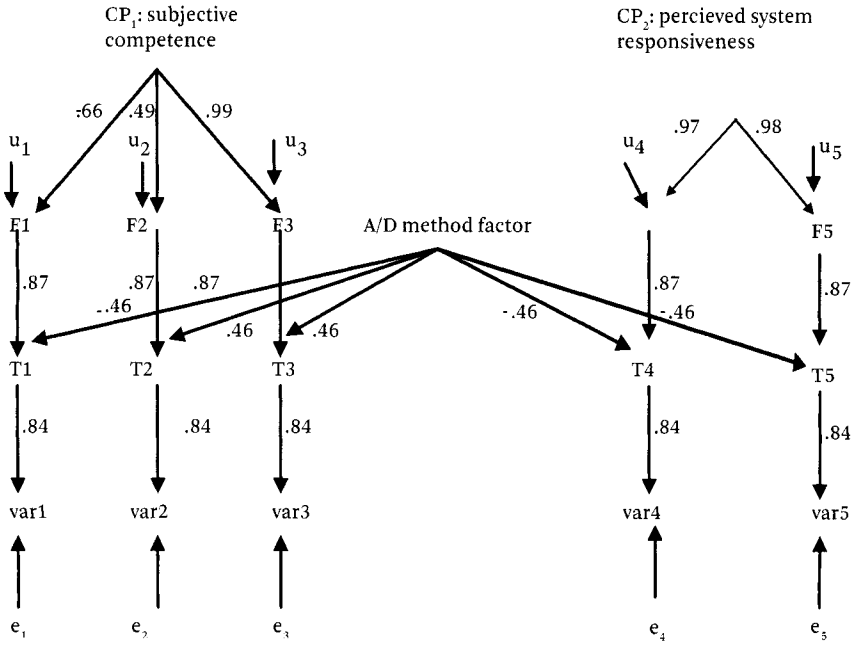


FIGURE 14.3: *An alternative factor model for political efficacy combining the two-factor model with information about the quality of the measurement of the concepts-by-intuition.*

If these effects are estimated, a very good fit is obtained where the effect of "subjective competence" on the lower level latent variables is respectively -.66 for F_1,.49 for F_2, and .99 for F_3, while the effect of "perceived system responsiveness" is .97 for F_4 and .98 for F_5. In this model it turns out that the correlation between the two highest latent variables is not significantly different from zero

(.04), and therefore it is reasonable to assume that the two latent variables vary independently from each other.

This model demonstrates that a different explanation for the correlations between the two sets of variables is possible. This is because in this particular model it is assumed that a part of the correlation between the observed variables is spurious, due to a factor that has no substantive meaning (in this case a method factor). In the model of Figure 14.2 this was not assumed, and the correlation between the latent variables was seen as substantive correlation. It is important to note that this difference between the models causes the estimates of the quality of the instruments to differ. Therefore it is absolutely necessary to test the model and to ensure that it is correct before starting to estimate the quality of the measures.

14.2.2 Estimation of the composite scores

Only after testing the latent variable model is it advisable to move to the next phase of constructing the measure of the concept-by-postulation. Equation (14.1) demonstrates a possibility for estimating the scores of the respondents on the latent variables by using the weighted average of observed variables (S) as a measure for the concept-by-postulation (CP_i):

$$S = \sum_{i=1}^{k} w_i \, vari \tag{14.1}$$

In this equation w_i is the weight for the ith observed variable. However, the question of choosing the weights is still left up to the researcher. Most of the time, this problem is avoided by choosing the value 1 for the weights, which leads to an unweighted sum score. However, this approach is rather inefficient if the different variables differ in quality. Therefore procedures have been developed to estimate weights that are optimal for some specific applications. Well-known criteria are as follows:

1. The sum of the squared differences in scores between the variable of interest and the sum score should be minimal. The weights derived using this criterion are known as *regression weights*. They are the most appropriate when trying to obtain scores for individual persons.
2. The sum score should be an unbiased estimate of the variable of interest and should satisfy criterion 1. The weights derived using these two criteria are known as the *Bartlett weights*. This would be the best procedure for comparison of means for different groups.
3. The relationships between the sum scores should be the same as the relationships between the different factors, and criterion 1 needs to be satisfied. The weights derived using these two criteria are known as the *Anderson and Rubin weights*. This is the preferred method when attempting to estimate the relationships between the latent variables in more detail.[1]

[1] In the next chapter we will show that this approach is not necessary because it is possible to directly estimate the relationships between the latent variables.

For further reading we recommend Lawley and Maxwell (1971) and Saris et al. (1978). In general it can be observed that the different methods generate only slightly different results if the different observed variables are approximately equally good. However, it will be shown below that unequal weights have important advantages if this condition is not satisfied.

If the observed variables are expressed in deviation or standardized scores and, therefore, have a mean of zero, then the sum score will also have a mean of zero. The variance of the sum can be calculated as follows:

$$\text{var}(S) = \sum_{i=1}^{k} w_i^2 \, \text{var}(\text{vari}) + 2 \sum_{i,j} w_i w_j \, \text{cov}(\text{vari},\text{varj}) \qquad (14.2)$$

Here "var(S)" stands for the variance of "S" while "cov(vari,varj)" is the covariance between the variables "i" and "j." By taking the square root of the var(S), the standard deviation of S can be obtained denoted by "sd(S)."

14.2.3 The quality of measures for concepts-by-postulation

In the previous chapters we have discussed the quality of only single items as indicators for concepts-by-intuition. Now we want to introduce the quality of the sum scores of concepts-by-intuition for the concepts-by-postulation. Until now quality, was always defined as the correlation squared between the theoretical variable of interest and the observed variable. This definition still holds true for our current case. We will also define method effects as the complement of validity and random error variance as the complement of unreliability.

If we combine the specification for generating a sum score with the knowledge we have about the model describing the relationship between the latent variables (which represent the concepts-by-postulation and the observed variables), we get, for the model presented in Figure 14.2, the results presented in Figure 14.4.In this figure we see the relationships between the latent variables representing the concepts-by-postulation (CP) and the sum scores (S) for these variables. The *quality of these sum scores as indicators for these latent concepts can be expressed in the correlation squared between these two variables.*

In Chapter 9 we mentioned that the correlation between two standardized variables is the sum of direct, indirect effects, spurious relationships, and joint effects. In this case there are only indirect effects, and therefore it follows that the correlation is equal to the sum of the indirect effects of "subjective competence" (CP_1) and S_1, and between "perceived system responsiveness" (CP_2) and S_2. This leads to the following general result for k observed variables[2]:

$$\rho(CP_1, S_1) = \sum_{i=1}^{k} \frac{q_i w_i}{\text{sd}(S_i)} = \frac{1}{\text{sd}(S_i)} \sum_{i=1}^{k} q_i w_i \qquad (14.3)$$

[2] Here it is assumed that all variables are standardized as is done in the whole text except the new variable S_i. For a more general formulation we mention Bollen (1989: 209-222)

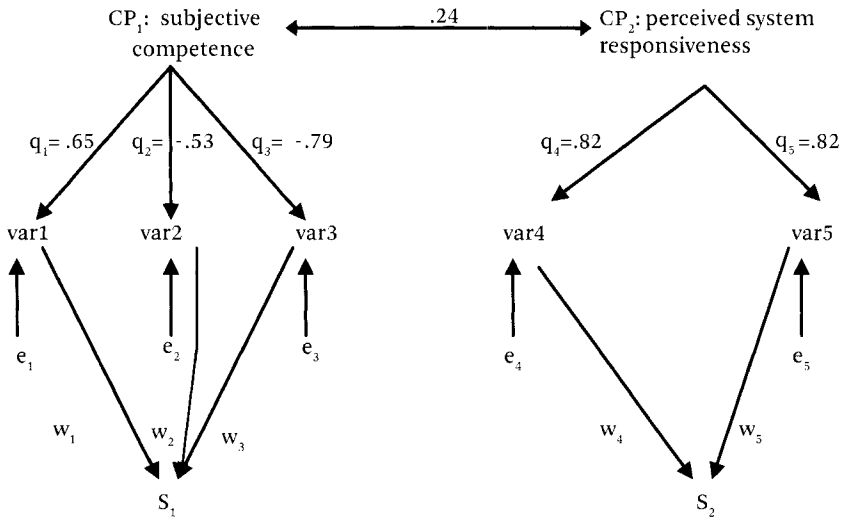


FIGURE 14.4: *The model for evaluation of the quality of the sum scores.*

The regression weights that minimize the sum of squared differences between the concept-by-postulation and the sum score are computed: for var1 it is .29, for var2 it is -.19 and for var3 it is -.55, while the weight for var4 and var5 turned out to be negligible. Using these weights the variance of the sum is calculated at .715 and the standard deviation is .845. Therefore, the correlation between CP_1 and S_1 becomes

$$\rho(CP_1, S_1) = \frac{1}{.845}\,[(.29 \times .65) + (-.19 \times -.53) + (-.55 \times -.79)] = .857$$

The strength of the relationship between this weighted sum score and CP_1 is this correlation squared, which equals .73.

In practice often the unweighted sum of the observed variables is often used. In that case the weights are all equal to 1/ sd(S). Hence formula 14.3 can be simplified to[3]:

$$\rho(CP_1, S_1) = \sum_{i=1}^{k} \frac{q_i}{sd(S_1)} = \frac{1}{sd(S_1)} \sum_{i=1}^{k} q_i \qquad (14.4)$$

If the unweighted sum[4] of the observed variables var1 through var3 is used then the variance of this sum is 7.0 and the standard deviation sd(S1) is equal to 2.646. From this it follows that:

[3] See also Heise and Bohrnstedt (1970) and Raykov (2001).

[4] In this calculation a correction for the direction of the scale is necessary; therefore the second and third weights are -1.

$$\rho\,(CP_1,\ S_1) = \frac{1}{2.646}\ [(1\times.65) + (-1\times-.53) + (-1\times-.79)] = .745$$

This means that the strength of the relationship between "subjective competence" (CP_1) and S_1 is

$$\rho(CP_1,S_1)^2 = .56$$

This correlation indicating the quality of the unweighted sum score is considerably lower than the quality for the weighted sum score, meaning that in this case the weighted sum is considerably better than the unweighted sum score.

In fact, it can be proved that the regression method provides weights that are optimal in the sense that it produces the highest possible correlation between the factor and sum score (Lawley and Maxwell 1971). *It produces the sum score with the best quality.* The major advantage of the regression method is that the quality of the sum score can never be lower than the quality of the best indicator. However, we have demonstrated with this example that this is not necessarily true for the unweighted procedure because variable 3 has better quality than the unweighted sum score.

In the literature people often use the so-called Cronbach α as a measure of the quality of the sum score. This quality index is calculated in many ways that are equal if the unweighted sum is used and all indicators have equal quality (q^2). Under this assumption it follows from equation (14.4) that

$$\rho\,(CP_1,\ S_1)^2 = \alpha = \frac{1}{var(S)}\ k^2q^2 \tag{14.5}$$

If all loadings are equal to q then the correlations between all observed variables is q^2 therefore it could also be said (see Bollen 1989) that:

$$\rho\,(CP_1,\ S_1)^2 = \alpha = \frac{1}{var(S)}\ k^2r \tag{14.6}$$

Since the correlations are normally not exactly equal, it is customary to take the mean of the correlations as an estimate for r. From our example we derive an estimate of the quality of the sum score of .56, which is much lower than the estimated reliability for the weighted sum using formula (14.3). It is known from the literature (Raykov 1997) that the Cronbach α is only the lower bound of the reliability. Only if the indicators satisfy the condition that they are all equally good indicators for the CP of interest and the unweighted sum score is used, is this estimate equal to the more general estimate presented in equation (14.3). We have seen above that this assumption is not necessarily true and that the quality of the sum scores can easily be calculated with the latter formula.

So far we have been talking about the estimation of the quality of the sum score as an indicator for the concept-by-postulation, now we would like to illustrate the consequences of the selected model on the quality of the sum score.

We will show what happens if we follow the same procedure with the model of Figure 14.3 as we did with the model of Figure 14.2. In order to proceed, we simplify the model. From Figure 14.3 we take that the effect of CP_1 and the method factor (A/D) on each observed variable is indirect. It can also be proved that the sizes of these indirect effects are equal to the products of the coefficients along the path from the causal variable to the effect variable (see Chapter 9). This leads to Figure 14.5.

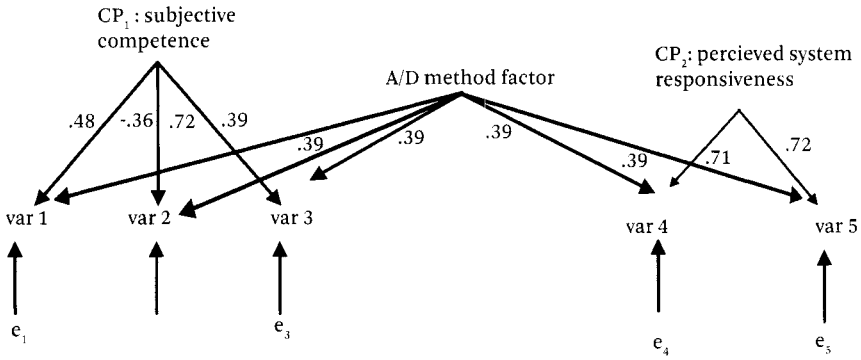


FIGURE 14.5: *The simplified model derived from Figure 14.3.*

This model differs quite a bit from the model in Figure 14.2 because in this case there is no correlation between the variables CP_1 and CP_2. The correlations between the observed variables are explained by the systematic effect of the method factor. We also see that by introducing this factor all measures for the quality of the different variables are lower than in Figure 14.2. These differences also affect the evaluation of the quality of the sum scores based on the observed variables. In Figure 14.6 we introduce the model for the estimation of the quality of the sum scores. The estimation of the quality of the sum score S_1 for CP_1 can be calculated by formula (14.3); however, the fundamental difference is that a second factor influences the sum score S_1 which is the A/D method factor. The effect of this A/D method factor, can be computed with the same formula but now substituting the quality coefficients (q_i) by the method effect coefficients (m_i):

$$\rho(AD,S_1) = \sum_{i=1}^{k} \frac{m_i w_i}{(sd(S_1)} = \frac{1}{sd(S_1)} \sum_{i=1}^{k} m_i w_i \qquad (14.7)$$

This correlation indicates the invalidity coefficient caused by the method in the sum score. If the unweighted procedure for estimation of the sum scores is used, the results of the estimation are for the quality of the sum score .35 and for invalidity .20. This result indicates that quite a large part of the systematic variance in the sum score S_1 is due to method effect (20%) and only 35% is due to

the variable it should represent (CP_1), while 45% is random error. These values differ significantly from the estimates derived on the basis of the model in Figure 14.2. There the quality was higher (58.3%) and invalidity was 0, because there was no method effect assumed. In principle the error variance of the sum score should remain the same along with the sum of the valid and the invalid variance, but minor deviations in the estimates can cause small differences.

The regression method improves the outcome: CP_1 explains 53.5% of the sum score and the method explains only 5.3% while the random error is 41.2%.

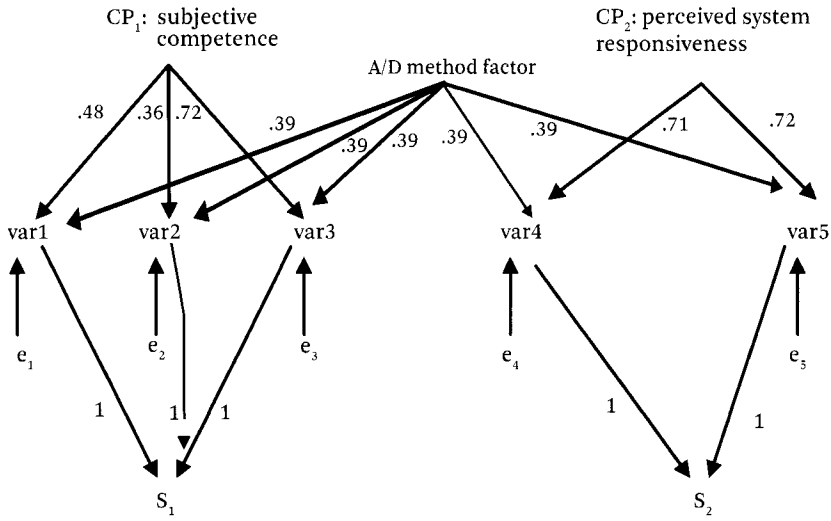


FIGURE 14.6: *Model for the calculation of the quality of the sum scores derived for the model in Figure 14.5.*

This example has illustrated how dependent estimates are on the specification of the model and on how the sum scores are computed. Our example illustrated that it is always safer to use one of the weighted procedures because they will give a sum score with better a quality. Also, our examples clearly showed the effect that the model has on the estimation. Again we specify how important it is to test the model before starting to evaluate the quality of sum scores.

After demonstrating the superiority of the second model with the method effects, we can state that the sum scores derived from the first model are biased. They overestimate the quality of the sum scores. However, for the second model we also have to conclude that a sum score with a quality of .535 is not sufficiently high. Therefore we think that indicators for the concept-by-postulation "subjective competence" need to be improved on.

14.2.4 Improvement of the quality of the measure

There are three reasons for the lack of quality of the sum score of "subjective competence" that we were discussing in section 14.2.3: (1) the lack of reliability of the three items; (2) the invalidity due to the method effect; and (3) at least two of the three variables have large unique components, making the link with the concept-by-postulation rather small (respectively $.66^2$ and $.49^2$). All three points can be improved on. Suggestions for the improvement of the reliability and validity can be obtained from analysis of the individual survey items using SQP. The program suggested that the results would be much better if the first three items would be reformulated into requests with trait-specific scales as shown in Table 14.3.

Table 14.3: The question format for "subjective competence" used in the first wave of the ESS

Var 1: *How often seem politics and government so complicated that you can't really understand what is going on?*

Never	*1*
Seldom	*2*
Occasionally	*3*
Regularly	*4*
Frequently	*5*
(Don't know)	*8*

Var 2: *Do you think that you could take an active role in a group that is focused on a political issue?*

Definitely not	*1*
Probably not	*2*
Not sure either way	*3*
Probably	*4*
Definitely	*5*
(Don't know)	*8*

Var 3: *How good are you at understanding and judging political questions?*

Very bad	*1*
Bad	*2*
Neither good nor bad	*3*
Good	*4*
Very good	*5*
(Don't know)	*8*

Using the MTMM estimates of the reliability and validity coefficients for the "subjective competence" questions the consistency coefficient have been estimated and are presented in Figure 14.7.

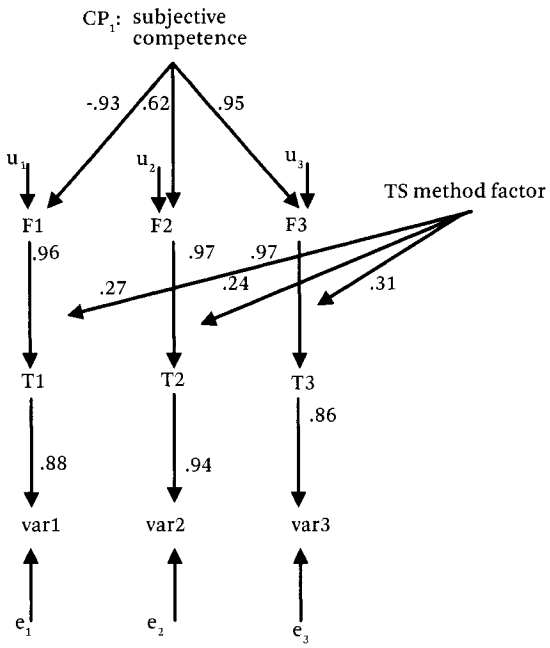


FIGURE 14.7: *The alternative factor model for "subjective competence" with information about measurement error.*

Applying the same simplification method as before, we get the result presented in Figure 14.8, where the derived effects are equal to the indirect effects of the model in Figure 14.7.

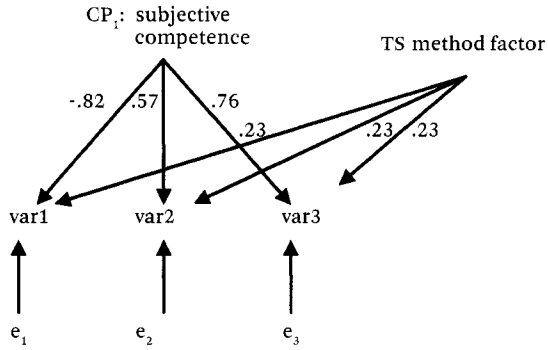


FIGURE 14.8: *The simplified model of Figure 14.7.*

Comparing this result with the result presented in Figure 14.5, we see that the strength of the relationships between CP_1 and all observed variables is stronger than in the A/D format. This improvement is due to higher reliability and validity derived from the changes we made. Given the stronger relationships the sum score of these variables should also be a better indicator for "subjective competence." In this case the regression weights are estimated at -.55 for var1, .16 for var2, and .38 for var3. The variance of the sum score derived, including the weights is .832 and the sd(S)= .911. While applying the procedure presented in equation (14.3), we see that the correlation coefficient between the concept-by-postulation and the sum score is .912, the quality of the sum score as an index for "subjective competence" is .83, while the method effect turns out to be .08. Our results indicate little allowance for random errors. It seems that this new sum score contains minor invalidity and only minimal random errors (9%), and that the quality of the measure is high with a 83% explained variance of the sum score by the factor of interest. This is because the TS format generates a higher data quality than does the A/D format. Our example also illustrates how the SQP predictions can be used to indicate the direction in which the quality of the measures for the concepts-by-intuition can be improved, thereby also improving the measures for the concepts-by-postulation.

In addition, the above illustration shows that the unweighted procedure for estimating the sum scores should not be automatically used. It can lead to a significant decrease in the quality of the measure of the concept-by-postulation compared with the measure obtained by the regression method. We have also shown that the Cronbach α is not always the best index by which to estimate the quality of a sum score; certainly not, if using weighted sums. We recommend the easy-to-use alternative discussed in this section, which results in better quality of the composite scores.

14.3 THE QUALITY OF MEASURES FOR CONCEPTS-BY-POSTULATION WITH FORMATIVE INDICATORS

A different model should be used if the variable of interest is the effect of several other variables such as when indicators, known as *formative*, determine or define the concept-by-postulation. The example in our introduction was the relationship between socio-economic status (SES) and the causal variables income, education, and occupation. The fundamental difference between the concept SES and the previous type of concept is that the observed variables are now the causal variables and not the effect variables. This has very different consequences. For example, in the previous section the model indicates that the correlations between the observed variables are spurious relationships, due to the unobserved causal variable. In the present case the unobserved variable has no effect on the observed variables. Whether there are correlations between the observed variables is not explained by the variable of interest. The model suggests only that there is an effect of each of the observed variables on the unobserved effect variable. Let us give two other examples of concepts-by-postulation with formative indicators.

Our first example, in Figure 14.9, is the measurement of "interest in political issues in the media" which is based of the time spent on politics in the media. Time spent watching TV programs or radio broadcastings or reading articles in the newspapers is observable. The "interest in political issues in the media" can be operationalized as the total time spent on political issues in the media and can logically be the sum of time spent on these three observable variables.

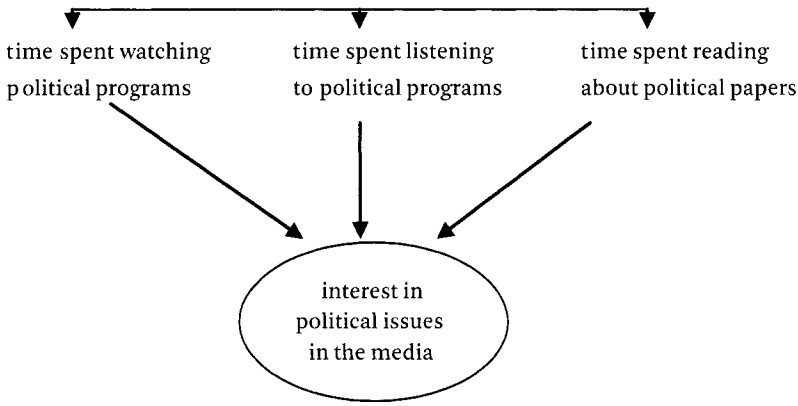


FIGURE 14.9: *The effect variable as concept-by-postulation.*

A second example is the measure of "social contacts" which is a key variable for research related to social capital and its effects. The measure includes "informal contacts" and "formal contacts" and an obvious measure for the concept-by-postulation is the sum of these two observable variables. In this case the causal structure for this concept is the same as indicated in Figure 14.9. Typical for both examples is that the observed causal variables do not have to correlate with each other. TV watching and reading newspapers can be done either in isolation or in combination.

A consequence of this model is, first that it is difficult to test, because the effect variable is not measured. Second, the weights of the different variables are left to the arbitrary choice of the researcher. A third issue is that the quality of the measure for the sum score as the correlation between the latent variable and the sum score cannot be determined. Therefore, different approaches have to be specified.

In the following section the solutions to the issues we raised will be discussed in the same sequence as was done for the concepts-by-postulation with reflective indicators.

14.3.1 Testing the models

A solution for testing this type of model and determining the weights can be to add extra variables to the models that are a consequence of the concept-by-postulation. In this way it becomes clear whether the effect really comes

from the concept-by-postulation or the concepts-by-intuition. Also it becomes possible to estimate the effects of the different components on the concept-by-postulation.

We will illustrate this procedure for our two examples. For the concept "social contacts" we add an effect on a latent variable "happiness" which has been measured by two variables: a direct question concerning "satisfaction" and a direct question concerning "happiness". It has been mentioned in the literature that socially active people are happier than socially inactive people. The theory does not state that this is due more to informal or formal interaction. Therefore, we assume that it is a consequence of the contacts in general.

For the measurement of "interest in political issues in the media" we can add the effect of "political interest in general" which will be operationalized by a direct question about political interest and by a measure of "knowledge of politics."[5] There is no doubt that these two variables are caused by the variable "political interest in general" and not by the different variables measuring "time spent on political issues in the media." Hence, we do not expect direct effects of these observed variables on the "political interest" indicators.

Taking into account that there is a difference between the concepts-by-intuition and the observed variables due to measurement error, we have created models for our two examples in Figures 14.10 and 14.11. These figures indicate the information that has been collected previously with respect to the quality of the requests. The information came from another source because the quality of the measures cannot be estimated by this type of model. The two sources for this information that have been discussed are the SQP program and the MTMM experiments, (which both can estimate the quality of single items). We know that the contact variables were asked only once in the first round of the ESS, therefore the quality was estimated by the SQP program. As it turns out, the quality coefficients are relatively good: .79 for informal contact and .68 for formal contacts. Since the measures are so different from each other, no correlation due to method effects is expected. Given the quality coefficients the error variances can also be calculated at $1\text{-}.79^2 = .38$ for "informal contact" and $1\text{-}.68^2 = .54$ for "formal contact."

The measures about "Time spent on programs in the media" were included in an MTMM experiment.[6] It turned out that the quality coefficients are .52 for TV, .73 for radio, and .48 for newspapers. Because these items had the same format; and were presented in a battery, a method effect of .09 was also found.

---

[5] This measure is based on the number of times the respondents answer "don't know" on questions concerning political issues in the ESS. Direct questions about political knowledge were not asked in the first round of the ESS.

[6] The MTMM experiments were conducted with the general questions about "media use" in the pilot study of the first round of the ESS. We assumed that the results for the questions about "political issues" will have the same quality characteristics since they share the same format.

Therefore, given these low-quality coefficients large error variances were found: .73 for TV, .47 for radio, and .77 for newspapers. In our models, the method effect is included as the correlated errors between the measurement error variables. This is a possible approach if the method factor itself is not specified.

These models are very different from the models for concepts-by-postulation with reflective indicators. Moreover, the measurement approach with formative indicators is more common in research than one would think. For example, in Likert scales different items are introduced to measure aspects or dimensions of a concept-by-postulation, and therefore there is no reason to expect correlations between the separate items (although such correlations cannot be ruled out). Hence, the quality of this type of model cannot be evaluated in the same way as we have elaborated in the previous section, but models like the ones specified for the two examples in Figures 14.10 and 14.11 can be used.

These two models have been estimated from round 1 data of the ESS. The LISREL input for the analysis is complex and available in Appendices 14.1 and 14.2 of this chapter. We will turn our attention to whether our analysis determines if the concepts-by-postulation are plausible. If the analysis shows that effects have to be introduced from the observed causal variables directly on the effect variables, it suggests that the concepts-by-postulation are not needed. Then the separate variables should be worked with as concepts-by-intuitions that have direct effects on other variables. On the other hand, if the effects are not needed, then the concepts-by-postulation are plausible because all effects go through them to other effect variables.

In our specific examples no direct effects were needed. It is a very convincing result because in both cases the effective sample size was 1500 cases (ESS 2002) making the power of such tests very high, meaning that even small effects would already lead to strong indications of misspecifications in the models and therefore to rejection of the models. Therefore, we can conclude that in our examples the concepts-by-postulation play the role that has been specified for them in their respective models.

14.3.2 Estimation of the composite score

In the above analyses the weights for the composite score were chosen to be equal to 1, making the composite score a simple sum of the different concepts-by-intuition. However, this is not necessarily the most accurate method. For example, it may be that "informal contacts" contribute more to the "happiness" of a person than do the "formal contacts" or vice versa. The same is true for the media attention. It may be that reading about political issues in the newspaper is a much better indicator of interest in politics than passively listening to radio or TV news.

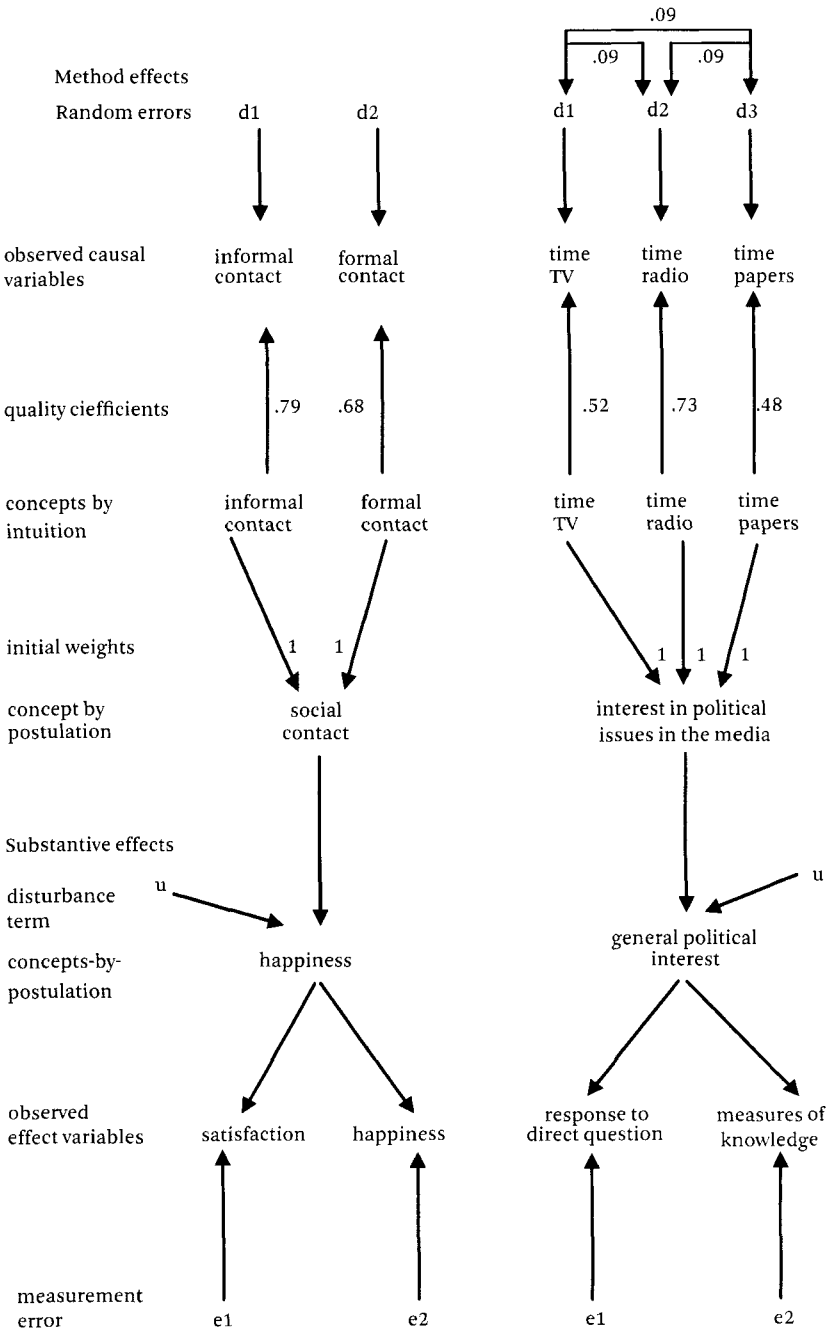


FIGURE 14.10

FIGURE 14.11

Structural equation model (SEM) programs show if the weights should be different from 1 in the expected parameter change (EPC) indices. These indices indicate the extent to which fixed coefficients will change if they are freely estimated. If these changes are substantively relevant, it would be wise to consider it.

In both our previous examples the EPCs for the weights were substantial. For the "social contact" variables the program suggested that the "informal contact" weighting would decrease by .87. After asking the program to freely estimate the coefficients the weights became .14 for "informal contacts" and .92 for "formal contacts." These differences are sufficiently large to be considered as substantively relevant. The results indicate that the social contact variable, with a much higher weight for "formal" than for "informal contacts," is a better causal variable for happiness than a "social contact" variable with equal weights.

For the concept-by-postulation "interest in political issues in the media" we see a similar phenomenon. Allowing a greater weighting for "reading about political issues in the newspapers," than for "radio" and "TV" the composite score better predicts "general political interest" than does a model with equal weights. The weights turned out to be .31 for TV, .1 for radio, and .8 for newspapers. Here, the differences in prediction quality also are substantively relevant.

Our analysis suggests that unequal weights should be used to estimate the scores of the concepts-by-postulation in both of our examples. The formula is the same as for calculating the concepts with reflective indicators [equation (14.1)]. However, below we will demonstrate that the evaluation of the quality of the composite scores is quite different in this particular case.

14.3.3 The estimation of the quality of the composite scores

So far we have evaluated measurement instruments by estimating the squared correlation between the observed variable and the latent variable of interest. There is, however, another equivalent way to evaluate measurement instruments. If the latent variable is called "F" and the observed variable "x" and the error variable "e" it has been shown by several authors (Bollen 1989) that

$$\text{Quality of } x = \rho_{Fx}^2 = \frac{\text{var}(F)}{\text{var}(x)} = 1 - \frac{\text{var}(e)}{\text{var}(x)} \qquad 14.8$$

In this situation we cannot use the squared correlation as a measure for the quality of the composite scores for the concepts-by-postulation with formative indicators, but we can use the last form. The quality of the sum score S can thus be defined as[7]

[7] This is true if it can be assumed that the concept-by-postulation is exactly defined as a weighted sum of the concepts-by-intuition. If that is not the case, and a disturbance term is specified, the result becomes more complex (Bollen and Lennox 1991).

$$\text{Quality of S} = 1 - \frac{\text{var}(e_S)}{\text{var}(S)} \qquad 14.9$$

where $\text{var}(e_S)$ is the variance of the errors in S and var(S) is the variance of the sum score S.

If for the different observed variables the weights (w), the error variances ($\text{var}(e_i)$), and covariances ($\text{cov}(e_i e_j)$) are known, we can estimate the error variance of the composite score "$\text{var}(e_S)$" as follows:

$$\text{Var}(e_S) = \Sigma\, w_i^2\, \text{var}(e_i) + 2\Sigma\, w_i w_j\, \text{cov}(e_i e_j) \qquad 14.10$$

The variance of the composite score can be obtained directly after calculating the composite score by asking for the variance of it. Formula (14.10) can be simplified to the first term if the error terms are not correlated (no method effects) and further reduced to the sum of the error variances if all weights for the components are equal to 1.

For the concept "interest in political issues in the media" we employ the complex formula because the error terms are correlated. On the other hand, for the concept "social contact," the second term can be ignored because the correlated error terms are equal to zero.

The results presented in the last two sections indicate that the variance of the errors for the concept "interest in political issues in the media" is:

$$\text{Var}(e_S) = .31^2 \times .7 + .1^2 \times .42 + .81^2 \times .75 + .31 \times .1 \times .09 + .31 \times .81 \times .09 + .1 \times .81 \times .09 = .60$$

The weights were estimated in such a way that the variance of the composite score is equal to 1. Hence, the quality of the composite score as an indicator for the concept "interest in political issues in the media" is

$$\text{Quality} = 1 - \frac{.6}{1} = .4$$

It will be clear that a quality score of .4 is not a very good result.

For the concept "social contact" the calculation simplifies because the correlations between the errors are zero and we have to evaluate only the first term:

$$\text{Var}(e_S) = .14^2 \times .384 + .92^2 \times .535 = .46$$

The weights were estimated in such a way that the composite score had a variance of 1 and the quality for this concept resulted as follows:

$$\text{Quality} = 1 - \frac{.46}{1} = .54$$

The quality of this score (.54) is better than of the previous one (.4), but it still is not very good. Both examples indicate that composite scores, as measures for concepts-by-postulation, can have considerable errors that should not be ignored. For both examples we recommend that researchers consider improving these measures before moving on with substantive research.

14.4 Conclusion

This chapter showed that there are several different models for representing the relationships between measures for concepts-by-intuition and concepts-by-postulation. In fact, the definition is the model. The testing of such models is essential. It is simpler if the model is a factor model. It becomes more difficult if the concept-by-postulation is the effect of a set of measures for concepts-by-intuition. In this chapter we have shown how these tests can be performed.

Since the concepts-by-postulation are defined as a function of the measures of the concepts-by-intuition, the quality of the composite scores can be derived directly from the information about the quality of the measures for the concepts-by-intuition. Therefore, evaluating the quality of concepts-by-intuition is very important and we have focused on this issue in this book.

We have also demonstrated that the composite scores (as measures of concepts-by-postulation) can contain considerable errors that can cause further substantive analysis to be biased. Therefore, the next chapter will show how to take these errors into account during the substantive analysis. In this context calculating the composite scores is highly advisable because we have seen that the models can become rather complex if substantive and measurement models need to be combined. Using composite scores simplifies the models. However, this should not be an excuse to ignore the measurement errors in the composite scores because they introduce considerable biases into the analyses.

EXERCISES

1. Choose the ESS data of one country for the following exercises:
   a. Compute the correlation matrix, means, and standard deviations for the indicators of the model of Figure 14.1.
   b. Estimate the parameters of the model on the basis of the estimated correlations.
   c. How high is the correlation between the factors?
   d. What do you conclude – can we speak of a variable "political efficacy" or should we make a distinction between two different variables?
2. For the same data set perform the following tasks:
   a. Estimate the regression weights for the indicators for the concepts found.
   b. Estimate the individual composite scores.
3. Evaluate the quality of the composite scores.
   a. Find the strength of the relationship between CP_1 and S_1 (the weighted sum score).
   b. Find the strength of the relationship between CP_2 and S_2 (the weighted sum score).
   c. Find the Cronbach α for the two relationships: CP_1-S_1 and CP_2-S_2.
4. From the ESS data of the same country, select the indicators for "formal" and "informal contact" and answer the following questions:
   a. Why are the indicators for "social contact" not reflective but formative indicators?
   b. Use the SQP program to determine the quality of the indicators.
   c. How large is the measurement error variance of these two variables?
   d. Now compute the unweighted composite score for "social contact".
   e. What is the variance of this variable?
   f. Calculate the quality of this composite score.
   g. Is the quality of the composite score good enough to use the composite score as an indicator for "social contact?"

APPENDIX 14.1: LISREL INPUT FOR FINAL ANALYSIS OF THE EFFECT OF "SOCIAL CONTACT" ON "HAPPINESS"

```
mimic partcip - satisfaction in The Netherlands
data ni=4 no=2330 ma=km
km
1.00
.660 1.00
.121 .134 1.00
.178 .181 .274 1.00
sd
1.647 1.416 1.356 .952
me
7.62 7.79 5.28 2.78
labels
satif happy infpart formpart
model ny=2 nx=2 ne=2 nk=2 ly=fu,fi te=di,fr lx=fu,fi td=di,fi ga=fu,fi be=fu,fi ps=sy,fi
ph=sy,fi
value 1.0 ly 1 2
free ly 2 2
value .785 lx 1 1
value .682 lx 2 2
value .384 td 1 1
value .535 td 2 2
value 1 ga 1 2
free ga 1 1
free be 2 1
free ps 2 2
value 1 ph 1 1 ph 2 2
free ph 2 1

start .5 all

out sc adm=of ns
```

APPENDIX 14.2: LISREL INPUT FOR FINAL ANALYSIS OF THE EFFECT OF "INTEREST IN POLITICAL ISSUES IN THE MEDIA" ON "POLITICAL INTEREST IN GENERAL"

```
Political interest in the Netherlands
data ni=5 no=2330 ma=km
km
1.00
 .215 1.00
-.056 -.262 1.00
-.046 -.126 .151 1.00
-.073 -.327 .247 .164 1.00
sd
1.249 .797 1.1356 1.56158 .93348
me
.401 2.28 2.28 1.366 1.054
labels
knowl polint tvtime radiotime paptime
select
1 2 3 4 5/
model ny=2 nx=3 ne=2 nk=3 ly=fu,fi te=di,fr lx=fu,fi td=sy,fi ga=fu,fi be=fu,fi ps=sy,fi
ph=sy,fi
value 1 ly 1 2
free ly 2 2
free be 2 1 ps 2 2

value .52 lx 1 1
value .73 lx 2 2
value .48 lx 3 3
value .64 td 1 1
value .38 td 2 2
value .69 td 3 3
value .09 td 2 1 td 3 1 td 3 2

value 1 ga 1 1
free ga 1 3 ga 1 2
value 1 ph 1 1 ph 2 2 ph 3 3
free ph 2 1
free ph 3 1 ph 3 2
start .5 all

out sc adm=of ns
```

This Page Intentionally Left Blank

CHAPTER 15

Correction for measurement error in survey data analysis

In this chapter we will discuss how to take measurement error into account during survey data analysis. In the previous chapters we have seen that random and systematic measurement errors in survey research can be considerable. It was also demonstrated that measurement error can have a considerable effect on the results and, therefore, that correction for measurement error is an essential part of survey data analysis. In fact, without first correcting for measurement error we cannot trust the results of the analysis.

There are many different ways to cope with measurement error in survey research. One approach is structural equation modeling. It is an approach that has potential, but, frequently too simple models are constructed that ignore parts of the problem. If this approach is used in the proper way, it can lead to rather complex models, which, in turn, cause technical problems.

An alternative approach is to use simple models where the variables are the composite scores for the concepts-by-postulation. This, however, is incorrect because, as demonstrated in the last chapter, it ignores the measurement errors that are still present in the composite scores.

We propose an alternative approach that consists of estimating simple substantive models that include the composite scores of the variables, but correct for the measurement errors that exist in the composite scores. In order to make our discussion practical we will illustrate the different approaches with one example that we will introduce in the next section.

15.1 A SIMPLE SUBSTANTIVE THEORY TO BE EVALUATED

During the last 15 years a lot of attention has been given to the theory of "social capital" (Coleman 1988; Putnam 1993; Newton 1997; Halpern 2005). This theory suggests that investment in "social contact" functions for people as an asset that results in trust in other people and in the political system. We take these hypotheses as the starting point for our model and add more variables to it because we think that not only "social contact" influences "social trust" and "political trust." We enrich the model by adding the variables "experience of discrimination" and "political interest" for explanation of "social trust" and

"political trust," and to explain "political trust" the variables "political efficacy" and "political interest" are added. Figure 15.1 incorporates these variables into a simple substantive model.

In this model it is assumed that the variables "social contact," "experience of discrimination" and "political interest" cause a correlation between "social trust" and "political trust" and that these two latter variables also have a reciprocal causal relationship. The reciprocal effect is included because it is plausible and to date has not been falsified.

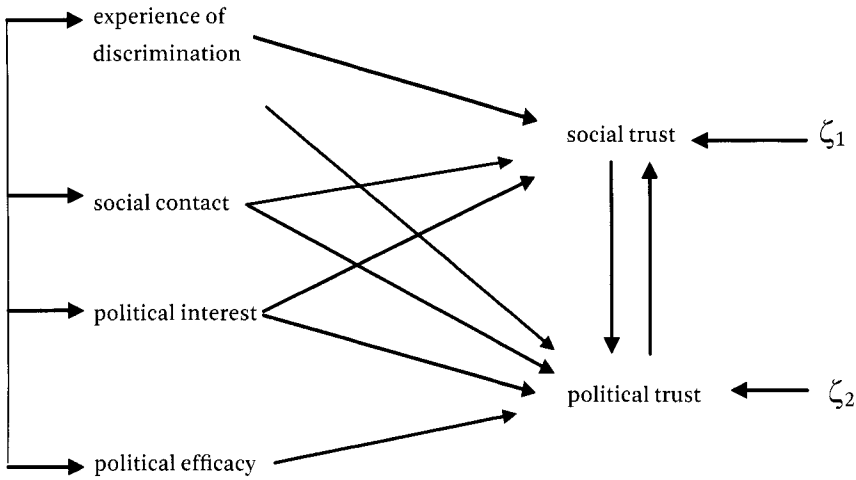


FIGURE 15.1 *A structural model of a simple theory about effects of "social contact" and other variables on "social trust" and "political trust."*

Given this model, there are sufficient reasons to assume that there must be a significant relationship between the two trust variables. However, in previous empirical studies of this theory using standard measures of "social trust" and "political trust" the correlations between these variables were not significant. Hence, scholars (Newton 1997) started to ask the following questions: Why should these variables be correlated? Moreover, should the correlation be present only at aggregate level? Even before these questions were answered, significant correlations were found in the first round of the ESS. The difference between the earlier studies and the ESS is that previously 4-point scales were applied, while the ESS used 11-point scales.

To be sure about this explanation, the ESS did an experiment in the pilot to the second round to see if it is indeed the difference in scale that causes the difference in results. The test consisted of asking the same people the same questions for "social" and "political trust" twice: once on a 4-point scale and once on an 11-point scale. The requests can be found in Appendix 15.1. The results obtained in Great Britain for the two different scales are listed in Table 15.1.

This table shows that the scales are responsible for the differences in the strength of the relationships. Using a 4-point scale only 1 out of 9 correlations is significant while all 9 correlations are significant if an 11-point scale is used.

However, what is more important is that we cannot state with certainty that the results of the 11-point scale are better than those using other scales. It is possible that the correlations are higher for this scale because they are increased by method effects. In fact, behind the different results only one set of correlations is possible because the data come from the same sample. Therefore the explanation for the differences has to be found in a combination of random and systematic errors that can differ for the two types of scales. This argument was already made in Part III of this book. Before we can draw any conclusions about these and other correlations or effects, corrections for random and systematic measurement errors have to be made.

Table 15.1 Correlations between "social" and "political trust" items for 4- and 11-point scales obtained from the British pilot study of round 2 of the ESS

	Political trust		
	Item 1	**Item 2**	**Item 3**
Social trust			
Item 1			
4 points	-.147[a]	-.030	-.094
11 points	.291[a]	.225[a]	.208[a]
Items 2			
4 points	-.060	-.070	-.005
11 points	.313[a]	.285[a]	.328[a]
Items 3			
4 points	-.074	-.064	-.041
11 points	.265[a]	.242[a]	.227[a]

[a] means significant on .05 level

For these corrections knowledge of the size of the random and systematic errors is needed. Therefore, we start with an evaluation of the operationalization of the variables of the example. After that we will discuss the correction for measurement error in the analysis.

15.2 Operationalization of the concepts

In Table 15.2 we give an overview of the operationalization of the different concepts defining the chosen approach of the ESS in the first round. Most concepts are concepts-by-postulation with several reflective indicators.

Table 15.2: The operationalization of the concepts in Figure 15.1

Concept name	Concept type	Observed indicator	Characterization of the indicators
Social contact	postulation	informal contact	formative
		formal contact	formative
Social trust	postulation	can be trusted	reflective
		fair	reflective
		helpful	reflective
Political trust	postulation	parliament	reflective
		legal system	reflective
		police	reflective
Political efficacy	postulation	complex	reflective
		active role	reflective
		understand	reflective
Discrimination	intuition	discriminated	direct question
Political interest	intuition	interested	direct question

"Social contact" is a concept-by-postulation with two formative indicators as has been discussed in Chapter 14. "Social trust," "political trust," and "Political Efficacy" are concepts-by-postulation with reflective indicators. "Experience of discrimination" is a concept-by-intuition measured by a direct question. "Political interest" could have been measured in different ways (see Chapter 1) but we opted for a direct question as a measure for the concept-by-intuition.

Part III of this book demonstrated how to estimate the size of the errors or the quality of a single question by using MTMM experiments; at least three forms of the same request for an answer are needed. In Chapter 13 we showed that an estimate of the size of the errors can also be obtained through the SQP program. It reduces the number of concepts to be measured to one for each indicator, which is more efficient than the MTMM approach.

Furthermore, in Chapter 14 we have already seen that the quality of a measure for a concept-by-postulation can be derived if the qualities of the measures for the concepts-by-intuition are known. Therefore, the number of observed variables can be reduced to 1 for each variable in the model.

Our overview of the different possibilities to evaluate the quality of the measures in a study leads to designs that differ with respect to the number of observed variables and complexity of the model. Table 15.3 summarizes the possibilities.

Table 15.3: Possible designs of a study with respect to the number of observed variables included in the model

Concept name	Number of observed variables			
	Composite scores	Indicators		
		Single	Multiple	In the ESS
Social contact	1	2	6	2
Social trust	1	3	9	9
Political trust	1	3	9	9
Political efficacy	1	3	9	9
Discrimination	1	1	3	1
Political interest	1	1	3	1
Number of observed variables	6	13	39	31

This table shows that only 6 observed variables are needed if composite scores for all concepts mentioned in the model are calculated while 13 variables are used if one form of each of the indicators for these concepts is employed. The option to combine the analysis with the evaluation of the data quality through MTMM analysis for this substantive research corresponds to the need for 39 observed variables. Finally, Table 15.3 informs us that there are 31 observed variables from the ESS: 13 from each indicator of the concepts within the model, which were collected in the main questionnaire. The remaining variables were collected in a methodological supplementary questionnaire that was answered by subgroups of the whole sample, using the two-group split-ballot MTMM design (Chapter 12). All 31 variables are not needed for the purpose of our analysis, but we see in this overview that some measures of the variables in the ESS design can also be evaluated by a MTMM analysis.

Our advice is to avoid making models with 31 or 39 variables, because it increases the risk of serious errors in the design and analysis. It calls for a complex model of a combination of MTMM models for each concept and the corresponding substantive model of Figure 15.1. Therefore, in the following discussion, we will concentrate on the use of composite scores (6 observed variables) and models with indicators for each concept-by-intuition (13 observed variables).

The following two steps are needed to reduce the design of the analysis while correcting for measurement error:

1. An evaluation of the measurement instruments
2. An analysis of the substantive model correcting for the detected errors

In the next section we will give an overview of the data quality of the possible observed variables.

15.3 THE QUALITY OF THE MEASURES

It is beyond the scope of this chapter to describe in detail how all the questions were evaluated. Some of the results of the studies of the quality of the measurement instruments have been presented previously. The results of these evaluations have been summarized in Table 15.4. The indicators for "social trust," "political trust," and "political efficacy" were evaluated by MTMM experiments,[1] while the other indicators have been evaluated by SQP.

Table 15.4: Quality estimates of the 13 indicators from the Dutch study in the ESS round 1. coefficient for

		Coefficient for				
Concept name	Indicator	Reliability	Validity	Quality	Consistency	Method used
Social contact	informal	.79	1.0	.79	-	SQP
	formal	.68	1.0	.68	-	SQP
Social trust	be trusted	.87	1.0	.87	.84	MTMM
	fair	.83	1.0	.83	.94	MTMM
	helpful	.84	1.0	.84	.66	MTMM
Political trust	parliament	.85	.95	.81	.66	MTMM
	legal system	.90	.96	.86	.99	MTMM
	police	.94	.96	.90	.66	MTMM
Political efficacy	complex	.88	.96	.85	.89	MTMM
	active role	.94	.97	.91	-.57	MTMM
	understand	.86	.97	.83	-.78	MTMM
Discrimination	direct request	.72	.72	.52	-	SQP
Political interest	direct request	.96	.80	.77	-	SQP

In the table we see that the quality of the indicators[2] evaluated by an MTMM experiment is much better than the quality of the indicators evaluated by the SQP program. Given that the SQP program is based on the MTMM experiments, there is no reason to think that this difference is due to the evaluation method used. The real reason is that MTMM experiments were done in the pilot study and the best method was selected for the main questionnaire in the definitive research. The results from the study confirm that this procedure is successful. The questions evaluated with the SQP program were not developed in the same

[1] In this chapter the Dutch data of the first official round of the ESS are analyzed, and not the data from the pilot study. As a consequence, the coefficients are slightly different from those presented in Chapter 14.

[2] The reader is reminded that the quality coefficient is the product of the reliability and the validity coefficient and the quality itself is the quality coefficient squared, which can be interpreted as the percentage explained variance in the observed variable by the concept-by- intuition.

way. They were not involved in a MTMM study in the pilot study and therefore were not improved upon.

This table also shows that for "social trust" and "social contact" the method effects are zero so that the validity coefficient, which is the complement, is equal to 1. For the concepts "political trust" and "political efficacy," this is not true; there the validity coefficients are not 1.

The low value of the quality of the "experience of discrimination" variable is of concern. The quality of this indicator is low because the explained variance in the observed score is only 27%. This is due partially to the lack of precision of the scale used, which is a yes/no response scale. Here a scale with gradation would result in a better quality measure. However, in the context of our illustration this lack of quality will serve to show just how large the effect of correcting for measurement error can be.

Table 15.4 also shows the size of the consistency coefficients of the different reflective indicators for the concept-by-postulation that they are supposed to measure. We have included these coefficients because they play a role in calculating the measures of the composite scores (Chapter 14). Such relationships do not exist for concepts with formative indicators or concepts-by-intuition.

Finally, we have to mention that we did not specify the method effects because they are the complement of validity (1 – validity coefficient squared). These effects are important because the method factors cause correlations between the observed variables, which have nothing to do with the substantial correlations. In this study such method effects can be found within sets of variables for the same concept, but not across the different concepts of the model, since the methods are too different for the different substantive variables.

Now that we have discussed the quality of the indicators, we can turn to the quality of the composite scores for the different concepts-by-postulation that have been included in Figure 15.1. Chapter 14 covered the quality of the composite scores for the "social contact" and "political efficacy" concepts. The measures for "social trust" and "political trust" are calculated using regression weights, followed by evaluation of the quality of these composite scores, using equation (14.3). The results for these four concepts-by-postulation have been summarized in Table 15.5.

In this table the construct validity coefficient represents the effect of the concept-by-postulation on the observed indicator. This coefficient is the product of the quality of the indicator and the consistency coefficient, which were presented in Table 15.4.

This table shows that the four concepts-by-postulation differ in quality. In the next section we will see that these differences play an important role when estimating the effects of the different variables on each other. Two concepts also contain invalidity due to method effects. However, we will not worry about this, because the methods were different across concepts and therefore the method effects could not affect the correlations between the different concepts.

Table 15.5: The quality of the measures for the concepts-by-postulation

		Construct		composite score	
Variable name	Indicator	Validity coefficient	Regression weights	Quality coefficient	Method effect
Social Contact				.74	.00
	informal	.79	.14		
	formal	.68	.92		
Social Trust				.81	.00
	can be trusted	.73	.35		
	fair	.81	.50		
	helpful	.55	.10		
Political Trust				.87	.31
	parliament	.53	.09		
	legal system	.86	.74		
	police	.59	.13		
Political Efficacy				.86	.22
	complex	.76	.53		
	active role	-.52	-.20		
	understand	-.66	-.34		

15.4 DIFFERENT WAYS TO CORRECT FOR MEASUREMENT ERROR IN ANALYSIS

Tables 15.4 and 15.5 summarize the quality of the variables that can be used as observed variables in the analysis to estimate the effects in the substantive model presented in Figure 15.1. Therefore we can now start with a discussion about the different possibilities to correct for measurement error in the analysis. Our example demonstrates that a fundamental choice is whether to use the composite scores as the observed variables in the analysis with a total of 6 observed variables or the indicators for the concepts-by-intuition, which leads to a total of 13 observed variables. Opting for using the indicators requires an extension of the model because in Figure 15.1 the indicators are not mentioned. Let us start with the slightly more complex approach, followed by the simpler one.

It should be mentioned that in principle it is also possible to estimate models with observed variables that represent all possible forms of the different indicators obtained from the MTMM experiments leading to 36 observed variables. However, this approach can create too many errors because of the complexity of the model. Therefore, for pragmatic reasons we will not discuss it any further in this book.

15.4.1 Models with indicators as observed variables

If indicators for the concepts-by-postulation are used as observed variables, the model in Figure 15.1 has to be extended to include the relationships between the substantive variables and the indicators for which scores have been obtained. Figure 15.2 illustrates this extension.

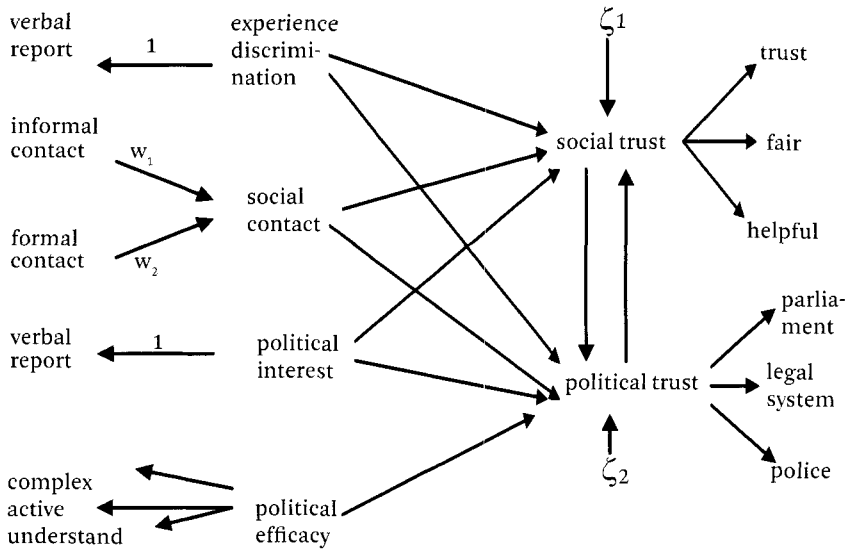


FIGURE 15.2: *Structural model that includes the indicators for the theoretical variables as observed variables.*

In this figure the measurement error variables have been omitted because of spatial constraints. They should be added to all indicators where the relationship between the theoretical variables and the observed variables is not specified. A "1" means that the observed and latent variable are seen as identical (no error). The two formative indicators for "social contact" are also assumed to have no errors and have an equal or unequal contribution to the theoretical variable "social contact" indicated by w_1 and w_2.

There are three different ways to estimate the coefficients for such models. All three approaches will be illustrated using the data from the first round of the ESS in the Netherlands (n=2300). The correlation matrix, applied in all three approaches, for all indicators is found in Appendix 15.2.

The first and most common approach is to directly estimate the parameters of the model from Figure 15.2. The LISREL input for this analysis is presented in Appendix 15.3.

A second approach is to correct for the measurement errors in the indicators by adding an extra layer of variables to the model, thereby making a distinction between the indicators as observed variables and the indicators corrected for measurement error. Here the quality estimates of the indicators from Table 15.4 are used for correcting for measurement error. The estimation of the parameters of this model is complex and goes beyond the scope of this book. The LISREL input for this approach is presented in Appendix 15.4.

A third approach is to correct for measurement error of the indicators by reducing the variances to the values of the quality coefficients squared

mentioned in Table 15.4. The covariances are corrected by subtracting the estimated method covariances from the observed covariances. After these adjustments the covariance matrix, asking for the analysis of the correlation matrix, the adjusted matrix is automatically corrected for the measurement errors[3] and the model from the first approach can be applied to estimate the parameters. The LISREL input for this approach is presented in Appendix 15.5.

Our analysis started with estimating the structural model presented in Figure 15.2. Parameters whose estimated values were not significantly different from zero were omitted in the model. This approach is common practice and in our example it demonstrates how large the differences between correction for measurement error and no correction for errors can be. Table 15.6 summarizes the results for the standardized coefficients, where only coefficients significantly different from zero are included.

The most remarkable result is that without correction for measurement error the effect of "social trust" on "political trust" was not significant at .19 (t=1.59) while the effect in the opposite direction was significant at .49 (t=2.38). Correcting for measurement error the results were exactly the opposite for both analyses; the effect of "political trust" on "social trust" was -.04 (t=-.09) and the effect of "social trust" on "political trust" was .42 (t=2.11). If the nonsignificant coefficients are omitted the results of Table 15.6 are obtained. These results show that considerably different conclusions will be drawn with or without correction for measurement error.

This also holds for other effect parameters. In general, not correcting for measurement error results in noticeably smaller effects than correcting for it. The difference increases as measurement quality decreases. In our example the "discrimination" variable has the lowest quality, and therefore correcting for errors has a considerable influence on the estimated values of the effects, especially the effect of "discrimination" on "social trust" increases. It is important to note that the quality should not be too low. If the relationship between the observed variable and the theoretical variable is too low, one does not know what the observed variable represents. In our example we have to be very careful with what conclusions we can draw for the "discrimination" variable.

[3] This is so because the program will calculate the correlation by dividing the provided covariance by the product of standard deviations of the related variables, namely the square root of the variances presented on the diagonal. But as the variance is equal to the quality of the measure and the square root of the quality is equal to the quality coefficient, this calculation is the same as dividing the provided covariance by the product of the quality coefficients, and this is exactly how correction for measurement error should be done, as was shown in Chapter 9.

Table 15.6: **Estimated values of the standardized parameters of the model presented in Figure 15.2 with and without correction for measurement error in the indicators**

	Correction for measurement errors in indicators		
Structural model	**No correction**	**Using quality coefficients**	**Using variance reduction**
Effects on social trust from			
Political Trust	.52	ns[a]	n.s.
Discrimination	.11	.36	.35
Social Contact	.05	.07	.08
Political Interest	ns	-.24	-.24
Political Efficacy	-	-	-
Effects on political trust from			
Social Trust	ns	.50	.48
Discrimination	.16	.14	.13
Social Contact	.05	ns	ns
Political Interest	-.07	ns	ns
Political Efficacy	-.28	.28	.29
Relationships with indicators			
For social trust			
Trust	.77	.89	.88
Fair	.73	.88	.89
Help	.56	.67	.67
For political trust			
Parliament	.63	.72	.72
Legal system	.87	.95	.95
Police	.67	.69	.70
For political efficacy			
Complex	.65	.84	.83
Active	-.60	-.63	-.64
Understand	-.65	-.78	-.81

[a] ns = not significantly different from zero

Finally, we see that the results from the analysis, using the quality coefficients for the indicators or reducing the variance of the observed indicators, are for all practical purposes identical. Any difference in results is mainly due to calculation accuracy. After comparing the LISREL inputs of Appendices 15.3 and 15.4, it can be concluded that the variance reduction method is much simpler and is preferable. A disadvantage is that the standard errors are not correct. They are underestimated.[4]

4 For proper estimates of the standard errors one has to use the alternative procedure.

15.4.2 Models with composite scores as observed variables

The most common method is to calculate the unweighted sum scores for all variables that have more than one indicator. However, it is also possible to use weighted sum scores, while keeping in mind the advice of Chapter 14 to apply the regression method for calculating weights. For the variables of Figure 15.1, we have presented the correlation matrices of both approaches in Appendices 15.6 and 15.7. Comparing these correlation matrices suggests that different results can be obtained. This is even more the case because sum scores are normally analyzed without correcting for measurement errors. We will do this, and we will compare the results against the two approaches where regression weights are used for calculating the composite score with correction for measurement error. Appendix 15.8 displays the LISREL input for the analysis of the sum scores is.

One way to correct for measurement error is to add 6 observed variables to the model and to specify that the relationship between the calculated composite scores and the theoretical variables is equal to the quality coefficient. The error variance of the observed variables is 1 minus the quality coefficient squared. The LISREL input for this approach is presented in Appendix 15.9.

A second way to correct for measurement error is to reduce the variances on the diagonal of the correlation matrix to the quality coefficient squared and to specify in the program that the matrix is a covariance matrix and that one would like to analyze the correlation matrix. The program will then automatically correct all correlations for measurement error and estimate the values of the parameters corrected for measurement error. The LISREL input for this approach is given in Appendix 15.10.

In the analysis we followed the previous procedure of beginning by estimating the model of Figure 15.1; depending on the output, only significant coefficients are included and nonsignificant ones are omitted. Table 15.7 presents the final result.

This table shows that the same results have been obtained as in the previous section; whether a correction for measurement error has been made has a significant impact on the final outcome. The greatest difference is found in the effects between "social trust" and "political trust."

We also see that the two approaches for correcting for measurement error, basically have the same results except for minor deviations due to how the errors were rounded off.

While comparing the two tables we see that the results after correcting for measurement error are similar, independent of whether indicators are used as observed variables or the composite scores of the theoretical variables. Therefore, there is no significant difference in whether the quality estimates of an extended model or the simple structural model are used for the analysis. However, an analysis without correcting for measurement error gives significantly different results depending on whether indicators or sum scores are employed as database. In our example the difference was most visible in the effect of the variable "political efficacy" on "political trust."

Table 15.7: Estimated values of the standardized parameters for the model presented in Figure 15.1 based on composite scores

	Correction for measurement errors in the composite scores		
Structural model	**No correction**	**using quality coefficients**	**using variance reduction**
Effects on social trust from			
Political trust	.42	ns	ns
Discrimination	.11	.38	.38
Social contact	.05	.08	.09
Political interest	-.06	-.26	-.25
Political efficacy	-	-	-
Effects on political trust from			
Social trust	ns	.49	.49
Discrimination	.15	.09	.10
Social contact	.05	ns	n.s
Political interest	-.14	ns	n.s
Political efficacy	-.15	-.27	-.27
Relationships with indicators			
For social trust	1	.81	.81
For political trust	1	.87	.87
For discrimination	1	.52	.52
Social contacts	1	.74	.74
For political interest	1	.77	.77
For political efficacy	1	.86	.86

The explanation is that while analyzing the data for the indicators some correction for measurement errors occurs, which results in a large effect between these two variables. In the analysis of the sum scores no correction for measurement error occurs, while Cronbach α for the sum score of "political efficacy" is only .655 and of "political trust" it is .724. Therefore, correcting for measurement error[5] would give an estimated effect of .22. This value is already much closer to the value obtained with correction for measurement error. It shows how important correcting for measurement error is when analyzing the relationships between variables.

[5] The correction can be approximated by dividing the effect by the square root of the two coefficients.

15.5 CONCLUSIONS

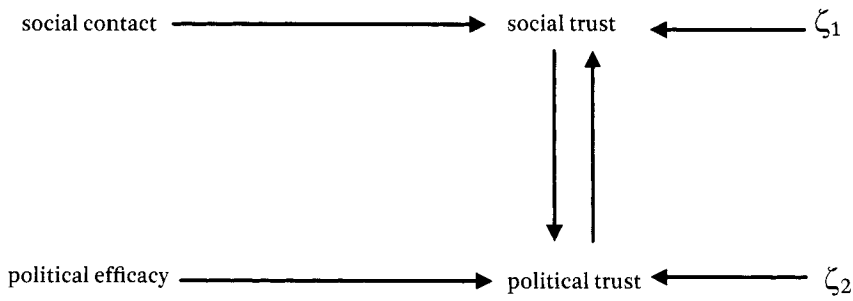
While using an example we showed in this chapter that whether corrections for measurement error are made makes a real difference in the final results. In order to be sure about the estimates of the relationships between variables we should correct for measurement error. This requires knowledge about the errors in the observed variables. In previous chapters we have shown that the more complex method used to obtain this information with respect to a single question is to do the MTMM experiments. A simpler way is to use the predictions that the SQP program provides. Given the information about the quality of the single questions, one can also estimate the quality of composite scores for concepts-by-postulation as we have shown in Chapter 14.

In this chapter we have shown how this information can be used to estimate the effects that the variables have on each other, corrected for measurement error. We demonstrated that this can be done in different ways and, if it is properly done, the different approaches should lead to similar results. Therefore, we recommend using the simplest method as demonstrated in the appendixes. This consists of using the calculation of composite scores using regression weights and applying the reduction of the variance to correct for measurement error.

A more commonly used, but wrong, approach is as follows. First indicators are developed for all theoretical variables, and the best are selected by factor analysis. Next unweighted sum scores are computed and the quality of the composite scores are evaluated with Cronbach α, hoping that the quality coefficients are close to .7 or higher. Then the model is estimated without any correction for measurement error. This approach leads to biased estimates of the relationships because correction for measurement error is not applied.

EXERCISES

1. Estimate the model presented below, which is a simplified version of the model used in this chapter.



The quality of the composite scores of "social contact" and "political efficacy" has been evaluated in the last chapter. Therefore your first task is to:

a. Test the measurement models for "political" and "social" trust for the same country that you used in the exercises of Chapter 14.
b. Compute the composite scores and determine the quality of the composite scores as was done in the last chapter.
c. Calculate the correlation matrix, means and standard deviations for the four composite scores.
d. Estimate the effects of the above model with and without correcting for measurement error on the basis of the correlation matrix.
e. Show that the two approaches to estimate the effects, correcting for measurement error, provide the same estimates, but different estimates for the standard errors of the parameters.
f. What is the consequence of this difference?

2. Given that we have the estimated values of the effects, the following questions can be asked:
   a. Does the model fit the data?
   b. If not, what has to be changed to fit the model?
   c. If so, which coefficients are not significant? Can these parameters be omitted while the model remains acceptable?
   d. What is your interpretation of the final model?

3. The parameters can also be estimated on the basis of the covariance matrix.
   a. How should one correct for measurement error if one uses the covariance matrix as the data for the estimation?
   b. Estimate the parameters on the basis of the covariance matrix.
   c. Why are the values of the parameters different?
   d. Ask for the completely standardized solution.
   e. If you did a correct analysis, the completely standardized solution should be approximately the same as the one in exercise 2.
   f. How can we explain minimal differences in the estimates and differences in chi^2?

APPENDIX 15.1: THE DIFFERENT VERSIONS OF THE REQUESTS CONCERNING "SOCIAL TRUST" AND "POLITICAL TRUST"

Measurement of "Social Trust" in the main questionnaire with an 11 point scale

A8 CARD 3: Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.

You can't be too careful										Most people can be trusted	(Don't know)
00	01	02	03	04	05	06	07	08	09	10	88

A9 CARD 4: Using this card, do you think that most people would try to take advantage of you if they would get the chance, or would they try to be fair?

Most people would try to take advantage of me										Most people would try be fair	(Don't know)
00	01	02	03	04	05	06	07	08	09	10	88

A10 CARD 5: Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves? Please use this card.

People mostly look out for themselves										People mostly try to be helpful	(Don't know)
00	01	02	03	04	05	06	07	08	09	10	88

Measurement of "political trust" in the main questionnaire with an 11 point scale.

CARD 8: Using this card, please tell me on a score of 0 to10 how much you personally trust each of the institutions. I read out: 0 means you do not trust an institution at all, and 10 means you have complete trust. Firstly...READ OUT...

		No trust at all										Complete trust	(Don't know)
B4	... [country]'s parliament?	00	01	02	03	04	05	06	07	08	09	10	88
B5	... the legal system?	00	01	02	03	04	05	06	07	08	09	10	88
B6	... the police?	00	01	02	03	04	05	06	07	08	09	10	88
B7	... politicians?	00	01	02	03	04	05	06	07	08	09	10	88

Measurement of "social trust" in the supplementary questionnaire with a 4 point scale.

S210. Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people?

Please indicate on a score of 1 to 4, where 1 means you can't be too careful and 4 means most people can be trusted.

You can't be too careful			Most people can be trusted
1	2	3	4

S211. Do you think that most people would try to take advantage of you if they would get the chance, or would they try to be fair?

Please indicate on a score of 1 to 4, where 1 means most people would try to take advantage of me and 4 means most people would try to be fair.

Most people would try to take advantage of me			Most people would try to be fair
1	2	3	4

S212. Would you say that most of the time people try to be helpful or that they are mostly looking out for themselves?

Please indicate on a score of 1-4 where 1 means people mostly look out for themselves and 4 means people mostly try to be helpful.

People mostly look out for themselves			People mostly try to be helpful
1	2	3	4

Measurement of "political trust" in the supplementary questionnaire with a 4 point scale.

S.213. Please indicate on a score of 1 to 4 how much you personally trust each of these institutions. 1 means you have a great deal of trust in them and 4 means you have none at all.

	A great deal	Quite a lot	Not very much	None at all
a) [Country]'s parliament?	1	2	3	4
b) The legal system?	1	2	3	4
c) Police?	1	2	3	4

APPENDIX 15.2: CORRELATION MATRIX FOR THE INDICATORS

	trust	fair	helpful	trparl	trlegal	trpolice
trust	1.00					
fair	0.57	1.00				
helpful	0.40	0.43	1.00			
trparl	0.33	0.26	0.26	1.00		
trlegal	0.38	0.29	0.27	0.54	1.00	
trpolice	0.31	0.28	0.29	0.39	0.59	1.00
discr	0.14	0.13	0.12	0.11	0.10	0.15
infcont	0.04	0.05	0.04	0.05	0.07	0.05
fromcont	0.09	0.10	0.06	0.06	0.09	0.05
polintr	-0.16	-0.12	-0.08	-0.23	-0.23	-0.09
compl	-0.15	-0.12	-0.04	-0.24	-0.27	-0.12
active	0.12	0.06	0.01	0.14	0.20	0.07
underst	0.06	0.03	-0.04	0.08	0.12	0.01

	discr	infcont	fromcont	polintr	compl	active
discr	1.00					
infcont	0.03	1.00				
fromcont	0.06	0.27	1.00			
polintr	0.02	-0.05	-0.10	1.00		
compl	0.01	-0.06	-0.09	0.43	1.00	
active	-0.09	-0.07	0.16	-0.41	-0.35	1.00
underst	-0.05	0.10	0.12	-0.42	-0.45	0.39

APPENDIX 15.3: LISREL INPUT FOR ESTIMATION OF THE MODEL WITHOUT CORRECTION FOR MEASUREMENT ERROR

```
Estimation of the effects without correction for measurement errors
data ni=13 no=2300 ma=km
km file=appendix15.2
labels
trust fair helpfull trparl trlegal trpolice discr infcont fromcont polintr compl active underst

model ny=6 ne=3 nx= 7 nk=5 ly=fu,fi te=di,fr lx=fu,fi td=di,fi be=fu,fi ga=fu,fi ps=sy,fr
ph=sy,fr
le
soctrust poltrust soccont
lk
discrim infcont forcont  polinterest poleff
free ga 1 1  ga 1 4 ga 2 4 ga 2 5
free  be 1 3 be 2 3
free be 1 2
value 1 ly 1 1 ly 4 2
free ly 2 1 ly 3 1 ly 5 2 ly 6 2
value 1 lx 1 1 lx 2 2 lx 3 3 lx 4 4 lx 5 5
fi ph 1 1 ph 2 2 ph 3 3 ph 4 4
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4
free  lx 6 5 lx 7 5
value 1.0 ga 3 2 ga 3 3
fi ps 3 3 ps 3 1 ps 3 2
free ga 2 1
fixed ps 2 1

free  td 5 5 td 6 6 td 7 7
start .5 all

out ss adm=off ns
```

APPENDIX 15.4: LISREL INPUT FOR ESTIMATION OF THE MODEL WITH CORRECTION FOR MEASUREMENT ERROR USING QUALITY ESTIMATES FOR ALL OBSERVED VARIABLES

```
Analysis using quality estimates of the indicators
data ni=13 no=2300 ma=km
km file=appendix1
labels
trust fair helpfull trparl trlegal trpolice discr infcont fromcont polintr compl active underst

model ny=9 ne=12 nx= 4 nk=5 ly=fu,fi te=sy,fi lx=fu,fi td=sy,fi be=fu,fi ga=fu,fi ps=sy,fi
ph=sy,fr
select
trust fair helpfull trparl trlegal trpolice compl active underst discr infcont fromcont polintr
le
soctrust poltrust soccont strust1 strust2 strust3 ptrust1 ptrust2 ptrust3 pole1 pole2 pole3
lk
discrim infcont forcont  polinterest poleff
free ga 1 1  ga 1 4  ga 2 1 ga 2 5
free  be 2 1 be 1 3

fixed ps 2 1

value 1 be 4 1
free be 5 1 be 6 1
value 1 be 7 2
free be 8 2 be 9 2
free ps 1 1 ps 2 2 ps 4 4 ps 5 5 ps 6 6 ps 7 7 ps 8 8 ps 9 9 ps 10 10
ps 11 11 ps 12 12

value .87 ly 1 4
value .83 ly 2 5
value .84 ly 3 6
value .243 te 1 1
value .311 te 2 2
value .294 te 3 3
value .808 ly 4 7
value .864 ly 5 8
value .902 ly 6 9
value .347 te 4 4
value .254 te 5 5
value .186 te 6 6

value .85 ly 7 10
value .91 ly 8 11
value .83 ly 9 12
value .28 te 7 7
value .17 te 8 8
value .31 te 9 9
fixed ph 5 5
value 1 ph 5 5
```

free ga 10 5 ga 11 5 ga 12 5

! method effect pol trust
value .07 te 5 4 te 6 4 te 6 5

value .52 lx 1 1
value .73 td 1 1
value .79 lx 2 2
value .68 lx 3 3
value .376 td 2 2
value .538 td 3 3
value .77 lx 4 4
value .41 td 4 4

! method effect for pol efficacy
value .05 te 8 7 te 9 7 te 9 8

fi ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
value .14 ga 3 2
value .92 ga 3 3

fi ps 3 3 ps 3 1 ps 3 2
start .5 all
out ss adm=off ns

APPENDIX 15.5: LISREL INPUT FOR ESTIMATION OF THE MODEL WITH CORRECTION FOR MEASUREMENT ERROR USING VARIANCE REDUCTION BY QUALITY FOR ALL OBSERVED VARIABLES

```
Analysis using the variance reduction of all indicators
data ni=13 no=2300 ma=km
cm
.757
.568 .689
.404 .429 .706
.326 .261 .256 .654
.383 .291 .271 .472 .740
.309 .275 .288 .318 .524 .810
.138 .132 .124 .106 .098 .148 .271
.043 .052 .041 .051 .070 .050 .026 .624
.087 .104 .062 .061 .091 .052 .059 .274 .463
-.158 -.121 -.082 -.228 -.225 -.092 .022 -.051 -.098 .593
-.153 -.123 -.039 -.235 -.273 -.123 .013 -.061 -.088 .426 .723
.116 .063 .012 .137 .195 .071 -.085 .073 .158 -.406 -.397 .828
.056 .034 -.037 .077 .120 .011 -.051 .095 .124 -.417 -.504 .344 .689
labels
trust fair helpfull trparl trlegal trpolice discr infcont fromcont polintr compl active underst

model ny=6 ne=3 nx=7 nk=5 ly=fu,fi te=sy,fi lx=fu,fi td=sy,fi be=fu,fi ga=fu,fi ps=sy,fr
ph=sy,fr
le
soctrust poltrust soccont
lk
discrim infcont forcont  polinterest poleff
free  ga 1 1 ga 1 4 ga 2 5
free  be 2 1  be 1 3
free ga 2 1
fixed ps 2 1
value 1 ly 1 1 ly 4 2
free ly 2 1 ly 3 1 ly 5 2 ly 6 2

value 1 lx 1 1 lx 2 2 lx 3 3 lx 4 4
fixed ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5

free lx 6 5 lx 7 5
free  lx 5 5
value .14 ga 3 2
value .92  ga 3 3
fi ps 3 3 ps 3 1 ps 3 2

free td 5 5 td 6 6 td 7 7

start .5 all
start .2 td 5 5 td 6 6 td 7 7
free te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
out  ss adm=off ns
```

APPENDIX 15.6: CORRELATION MATRIX FOR THE SUM SCORES

	soctr	*poltr*	*discr*	*socconr*	*poli*
soctr	1.00				
poltr	0.45	1.00			
discr	0.17	0.14	1.00		
socconr	0.10	0.10	0.05	1.00	
poli	-0.15	-0.22	0.02	-0.09	1.00
poleff	-0.10	-0.23	0.07	-0.15	0.54

APPENDIX 15.7: CORRELATION MATRIX FOR THE WEIGHTED COMPOSITE SCORES (REGRESSION METHOD)

	soctr	*poltr*	*discr*	*socconr*	*poli*
soctr	1.00				
poltr	0.41	1.00			
discr	0.16	0.12	1.00		
socconr	0.11	0.10	0.06	1.00	
poli	-0.15	-0.23	0.02	-0.10	1.00
polefr	-0.13	-0.26	0.05	-0.15	0.52

APPENDIX 15.8: LISREL INPUT FOR ESTIMATION OF THE MODEL WITHOUT CORRECTION FOR MEASUREMENT ERROR BASED ON UNWEIGHTED SUM SCORES

```
Analysis of the sum scores without correction for measurement error
data ni=6 no=2300 ma=km
km file=appendix15.6

labels
socts polts discr soccons poli poleff
select
1 2 6 3 5 4 /
model ny=2 nx=4 fixedx be=fu,fi ga=fu,fi ps=sy,fr

free ga 1 2 ga 1 3 ga 1 4 ga 2 2 ga 2 3 ga 2 4
free be 2 1
free ga 2 1
fixed ps 2 1
out ss
```

APPENDIX 15.9: LISREL INPUT FOR ESTIMATION OF THE MODEL WITH CORRECTION FOR MEASUREMENT ERROR USING QUALITY ESTIMATES FOR ALL OBSERVED COMPOSITE SCORES

```
Analysis of composite scores using quality estimates
data ni=6 no=2300 ma=km
km file=appendix15.7
labels
soctr poltr discr socconr poli polefr
select
1 2 6 3 5 4 /
model ny=2 ne=2 nx=4 nk=4 ly=fu,fi te=di,fi lx=di,fi td=di,fi be=fu,fi
ga=fu,fi ps=sy,fr ph=sy,fr
le
soctrust poltrust
lk
discr soccon polinterest poleff
free ga 1 1 ga 1 2 ga 1 3 ga 2 4
free be 2 1 be 1 2
free ga 2 1
fixed ps 2 1
value .81 ly 1 1
value .87 ly 2 2
value .52 lx 1 1
value .74 lx 2 2
value .77 lx 3 3
value .857 lx 4 4
value .344 te 1 1
value .243 te 2 2
value .73 td 1 1
value .452 td 2 2
value .407 td 3 3
value .265 td 4 4
fi ph 1 1 ph 2 2 ph 3 3 ph 4 4
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4
start .5 all

out ss adm=off ns
```

APPENDIX 15.10: LISREL INPUT FOR ESTIMATION OF THE MODEL WITH CORRECTION FOR MEASUREMENT ERROR USING VARIANCE REDUCTION BY QUALITY FOR ALL COMPOSITE SCORES[]

```
Analysis of composite score using the variance reduction method
data ni=6 no=2300 ma=km
cm
.656
.408 .757
.114 .099 .548
-.132 -.260 -.146 .735
-.155 -.227 -.101 .522 .593
.158 .115 .061 .049 .022 .270
labels
soctr poltr socconr poleff poli discr
select
1 2 6 3 5 4 /
model ny=2 nx=4 be=fu,fi ga=fu,fi ps=sy,fr

free ga 1 1 ga 1 2 ga 1 3 ga 2 1 ga 2 2 ga 2 3 ga 2 4
fixed ps 2 1
free be 2 1
out
```

This Page Intentionally Left Blank

CHAPTER 16

Coping with measurement error in cross-cultural research

In this chapter we will introduce the problem of measurement error in cross-cultural research. In particular, we will focus on comparative research across countries. In the last chapters we have established that measurement error has strong effects on results of research. Therefore, when the effects of measurement error differ in the individual countries of a study, comparisons across countries become quite challenging.

Two types of comparisons are most frequently made: comparison of *means* and comparison of *relationships* of different variables across countries. Often comparisons based on single requests or on composite scores of the latent variables are made. In this chapter we will add to this the comparisons based on latent variables.

The problem of such comparisons is that one can compare the results across different countries only if in fact the data are comparable, that is, if the measures used in the different countries have the same meaning. This topic is studied under the heading of functional equivalence or invariance of measures in different countries.

This chapter will concentrate on the procedures to determine equivalence of measurement instruments. But before we can introduce this topic, we have to introduce the notation for this topic. So far we used standardized variables and concentrated on the effects of these variables on each other. However, in cross-cultural research the means of the variables are frequently compared and therefore, we need to introduce unstandardized relationships, slopes, intercepts, and means.

16.1 Notations of response models for cross-cultural comparisons

In order to introduce the notations in this chapter, we will use the already familiar example of "political efficacy" and to a lesser extent "political trust." In Chapter 14 we introduced a measurement model for "political efficacy," here we will concentrate on the "subjective competence." Normally three requests are used to measure this concept. For the moment we will use only one request called "understand" with a 5-point response scale (see Appendix 16.1). The operationalization of this concept is illustrated in Figure 16.1.

At the top of the model the relationship between the concept by postulation "subjective competence" (CP_1) and the reaction to a specific request (f_1: "understand") is specified. This relationship represents a brain process that is triggered by the stimulus posed by the request. The output of this process is a variable that represents a possible reaction not yet expressed in the requested form (Van der Veld 2006). Assuming a linear relationship, we can expect the relationship that is presented in Figure 16.2 and equation (16.1):

$$f_1 = a + cCP_1 + u \tag{16.1}$$

Because "subjective competence" has been measured with different requests (see Chapters 14 and 15), "u" represents the unique component of the specific request (understand). We assume that the scores of this unique component vary around zero so that the mean score of this unique component over the whole population is zero.

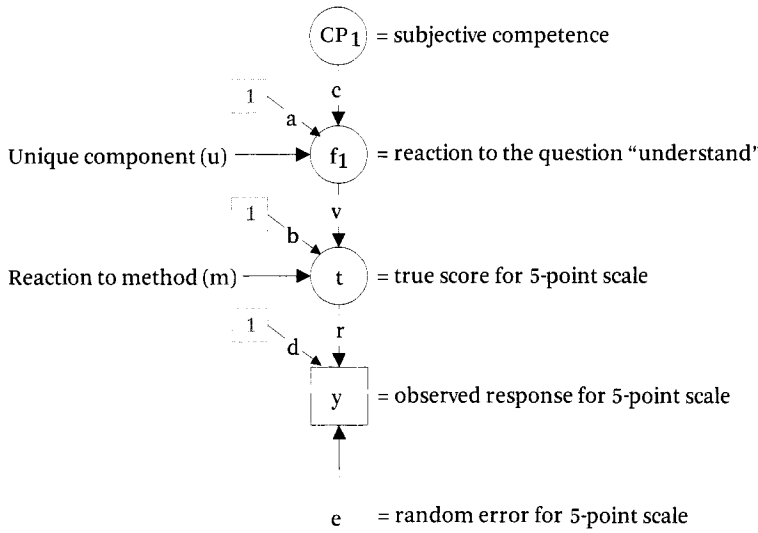


FIGURE 16.1: *A measurement model[1] for "subjective competence" using the request "understand" and a 5-point scale with a unique component (u), systematic method factor (m), and random error ("e").*

The unit of measurement of these variables is unknown, but what we can say is that "a" represents the mean value of f_1 if CP_1=0. If a=0, persons who have the impression that their "competence in politics" is 0 will also have a reaction that on average could be represented by 0. Furthermore, it is assumed that whatever

[1] New in this model is the specification of an effect of a "variable" called "1". This symbol is used to made a distinction between the effects of real variables and the effect of the intercepts.

the "subjective competence" of the person is, an increase in "subjective competence" of 1 unit will lead to an increase of "c" in the "reaction" (f_1). Therefore we can conclude that this relationship is linear, as can be seen in Figure 16.2 and equation (16.1), where the coefficient "a" is called the *intercept* and the "c" the *slope* of the function.

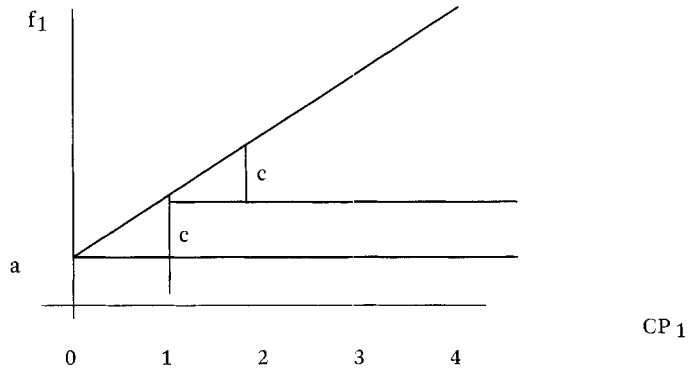


FIGURE 16.2: *The linear model presenting the cognitive response process.*

In the next step respondents have to express their "reaction to the request" in a certain format. Here we choose a 5-point scale; however, this choice is arbitrary, and for different choices we find the corresponding diverse results. Therefore in this relationship we introduce a method effect. Assuming again a linear relationship, the equation can be formulated as follows:

$$t = b + vf_1 + m \tag{16.2}$$

The interpretation is the same as the previous one: m is the random component and represents the different reactions respondents have to the method used (5-points scale). The intercept "b" will be 0, if the mean score of t=0 for those respondents for whom f_1=0. However, this may not always hold true because of the method effect. For example, an extreme opinion may have a value of 1 on the response scale and not 0 (see Appendix 16.1 for examples of this type) and then the intercept has to be equal to 1.

The final step is the real response selected on the 5-point scale. Assuming a linear relationship, we get

$$y = d + rt + e \tag{16.3}$$

Here "e" represents the disturbance due to random errors with a mean of zero. The coefficient "d" is the intercept that is zero, if the mean of y is zero for those respondents for whom t=0. Normally we can assume that r=1 and d=zero

because there is no reason to expect otherwise when the translation goes from a 5-point scale to a 5-point scale.

In the models used in the previous chapters the variables and coefficients were standardized and then the distinction between equations (16.2) and (16.3) makes sense because "v" would be the validity coefficient and "r," the reliability coefficient. However, in unstandardized forms these coefficients represent the slopes of their corresponding equations. Assuming that d=0 and r=1, we get that

$$y = t + e \tag{16.4}$$

This is the form commonly used in the classical test theory (Lord and Novick 1968).

Substituting equation (16.2) into equation (16.4) we get

$$y = b + vf_1 + m + e \tag{16.5}$$

Equations (16.1) and (16.5) together specify the response process, with the difference now being the use of unstandardized variables (variables expressed in their original units of measurement). Equation (16.1) presents the cognitive process started by the request for an answer and finishing with a preliminary reaction, while equation (16.5) represents the measurement process going from the preliminary reaction to an observed score. Both processes can differ between countries, and it may also be true that the differences come from only one of the two processes.

Most authors do not make this distinction (Grouzet et al. 2006; Davidov et al. 2006) and they only use the equation that can be obtained by substituting equation (16.1) into equation (16.5):

$$y = b + v(a + cCP_1 + u) + m + e$$

or

$$y = b + va + vcCP_1 + vu + m + e$$

This can be simplified to

$$y = \tau + \lambda CP_1 + \zeta \tag{16.6}$$

where

$$\lambda = cv \tag{16.6a}$$

$$\tau = b + va \tag{16.6b}$$

$$\zeta = e + m + vu \tag{16.6c}$$

Equation (16.6) is the equation frequently used for evaluation of measurement instruments also in cross-cultural research. This equation looks rather simple, but the coefficients are complex because they consist of different components. Unfortunately these components cannot be derived if equation (16.6) is used in the estimation of the response process as a whole.

On the other hand, we should say that the original model consisting of equations (16.1)–(16.3) is too complex to be estimated as it is. It requires more data to estimate all the parameters.

Assuming that the means of the disturbance term (u), the method effect variable (m), and the random error component (e) are equal to 0; then the mean over the observed responses (μ_y) of the respondents can be expressed as a function of the mean of the variable of interest (μ_{CP1}):

$$\mu_y = \tau + \lambda\mu_{CP1} \tag{16.7}$$

The result shows that the mean of the responses (μ_y) is not necessarily equal to the mean of the latent variable of interest (μ_{CP1}). This will normally only be the case if:

$$\tau=0 \text{ and } \lambda=1 \tag{16.8}$$

This requires that there be no systematic effects of both the request asked and the method used. It is unlikely that these conditions are fulfilled for all cases. However, the assumptions mentioned for equation (16.8) are commonly made because, normally, the observed mean is treated as the mean of the variable of interest.

So far we have formulated the response model in the original units of measurement. In the previous chapters we have always made use of standardized variables. The relationships between these different formulations can be found in the appendix of Chapter 9.

In cross-cultural research a translated request can be perceived differently between countries and languages. It is also possible that the use of a 5-points scale can create different reactions in different countries. Such between-country differences may change not only the correlations between the variables but also the slopes and intercepts of the different equations.

Normally it has been recommended that researchers compare responses across groups only if the requests can be seen as functionally equivalent. If *functionally equivalent* means that λ and τ in equation (16.6) are the same across countries, then the means across countries can be compared even though we are not sure that they represent the mean of the opinions of the sample on the latent variable of interest. But if these restrictions do not hold true what can we do? One question is whether we can make cross-cultural comparisons when these conditions are not satisfied. Another question is how we can evaluate whether requests are functionally equivalent. We will address these topics in the next sections.

16.2 TESTING FOR EQUIVALENCE OR INVARIANCE OF INSTRUMENTS

In this section we will present two definitions and approaches to evaluate the equivalence of indicators. First we will introduce the standard textbook approach. Then we will indicate the problem of this approach. After that we will formulate an alternative approach that is in line with the basic ideas in this book.

16.2.1 The standard approach to test for equivalence

Scholars commonly make a distinction between *configural*, *metric*, and *scalar invariance* (Horn et al. 1983; Meredith 1993; Steenkamp and Baumgartner 1998). While discussing the different forms of equivalence or invariance, they employ the model specification of equation (16.6) using several indicators for the same latent variable of interest. This is necessary, because with one observed variable the equivalence of the measures cannot be tested. In fact it has been discussed in Chapter 10 that three observed variables are needed to estimate the three effect parameters and error variances of a reflective measurement model with one latent variable. In case of the measurement of "subjective competence" there are indeed three indicators and the model can be estimated. The model would consist of three equations; one for each of the indicators. Using the formulation of equation (16.6) we get

$$y_1 = \tau_1 + \lambda_{11} CP_1 + \zeta_1 \quad (16.9a)$$
$$y_2 = \tau_2 + \lambda_{21} CP_1 + \zeta_2 \quad (16.9b)$$
$$y_3 = \tau_3 + \lambda_{31} CP_1 + \zeta_3 \quad (16.9c)$$

Since the scale of the latent variable needs to be fixed in one equation, the τ should be 0 and the λ equal to 1. This means that the latent variable CP_1 will be expressed in the same units as the observed variable, and that the observed score is 0 if the latent variable has a score of 0. If these restrictions are not made, the model is not identified. Further assumptions for the standard model are

$$\text{Covariance}\,(CP_1, \zeta_i) = 0, \quad \text{for all i} \quad (16.9d)$$
$$\text{Covariance}\,(\zeta_i, \zeta_j) = 0, \quad \text{for all } i \neq j \quad (16.9e)$$

This means that the model of Figure 16.3 is used in the test.

The literature says that comparison of means and relationships across cultures requires:

1. *Configural invariance*, meaning that the model of (16.9) holds for all the countries involved;
2. *Metric invariance*, meaning that, besides configural invariance, the slopes are the same in all the countries studied.

These two requirements are sufficient for comparison of relationships. The comparison of means requires:

3. *Scalar invariance*, meaning that, besides metric invariance, the intercepts are the same across all countries being compared.

These hypotheses can be tested using multiple-group analysis in SEM programs.

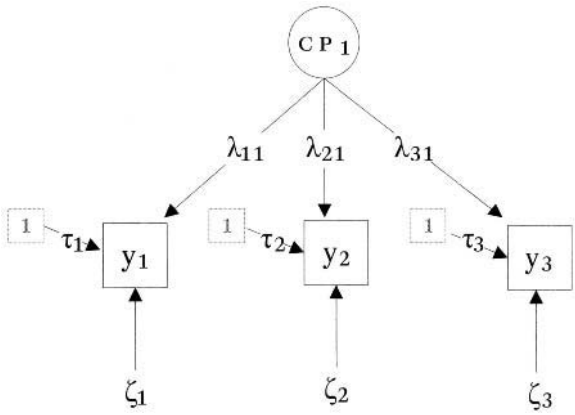


FIGURE 16.3: *The measurement model of "subjective competence."*

In this approach the same model (16.9) is simultaneously estimated in several samples, taking into account the invariance restrictions on the parameters from the different samples. Such restrictions can be tested using an overall test statistic, which is the sum of the test statistics for the different samples. The degrees of freedom of the test are equal to the sum of the degrees of freedom for the different samples meaning that

$$chi_k^2 = chi_{k1}^2 + chi_{k2}^2 + ... \quad (16.10a)$$

where

$$k = k1 + k2 + ... \quad (16.10b)$$

If the samples are independent, then chi_k^2 is also χ^2 distributed with df=k if the model is correct and distributed noncentral χ^2 if the model is incorrect. Here standard χ^2 tests can be used to test the hypotheses. The input of the scalar invariance test for the concept by postulation of "subjective competence" is presented in Appendix 16.3. The results of the tests of the different invariance restrictions for three countries, United Kingdom, The Netherlands, and Spain, are presented in Table 16.1.

This table shows that scalar invariance, where all loadings (slopes) and intercepts are equal across countries, must be rejected. The model, assuming that only the loadings are equal (metric invariance), fits much better to the data. However, this model can be improved upon if one loading in Spain is allowed to be different from those in the other countries. Then the model is acceptable, even though the sample sizes are large. Our results suggest that the indicators are not scalar-invariant, and according to the literature, this means that we

cannot compare the means across the different groups. In our case it is even questionable whether we can compare relationships given that one coefficient in Spain deviates from those in the other countries and, therefore, the measures are not metrically equivalent.

Table 16.1: Results of the tests of different requirements of invariance for the concept of "subjective competence" based on data from three countries

Invariance Restrictions	chi^2	df	Probability
Scalar	84.5	8	.00
Metric	13.3	4	.01
Metric except for λ_{21} in Spain	8.0	3	.39

We think that this test is too strict. The differences between the coefficients across countries may be due to the differences between measurement features of the requests, while the link between the concepts by postulation and the concepts by intuition is invariant. If this is the case, then we are able to correct for the measurement differences and have equivalent measures. How to do this will be illustrated in the next section.

16.2.2 An alternative test for equivalence

It is our opinion[2] that the invariance of the parameters in equation (16.1), the cognitive process equation, should be evaluated. This means that the parameters in the measurement equation can be different because the differences can be corrected. Therefore, we suggest that *cognitive equivalence of measurement instruments should be required, that is invariance after correction for differences in the measurement process.*

Consequently, the model specified in equations (16.1) and (16.5) should be used and not the derived model presented in equation (16.6). The model in Figure 16.4 gives us the second-order factor model that can be applied to test equivalence if the coefficients (v) and intercepts (b) of the measurement equation are known. If this information is available, the test of equivalence can be conducted in the same manner as previously indicated. This approach leaves us with the task to estimate the parameters of the measurement equations, and this can be done in a separate study using MTMM experiments or using predictions from SQP.

An alternative would be to simultaneously estimate and test the quality measures and the intercepts assuming equality of the consistency coefficients over the different countries. A model for this approach is shown in Figure 16.5.

[2] At first glance it appears as if Little (1997) and Grouzet et al. (2006) make the same observation but on a closer examination of their approach, we see that they share the opinion of other cited authors on this issue.

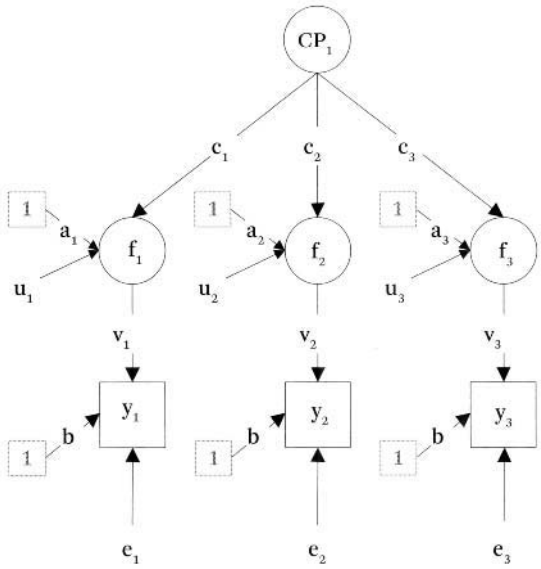


FIGURE 16.4: *The alternative measurement model[3] for "subjective competence" to be evaluated in cross-cultural research.*

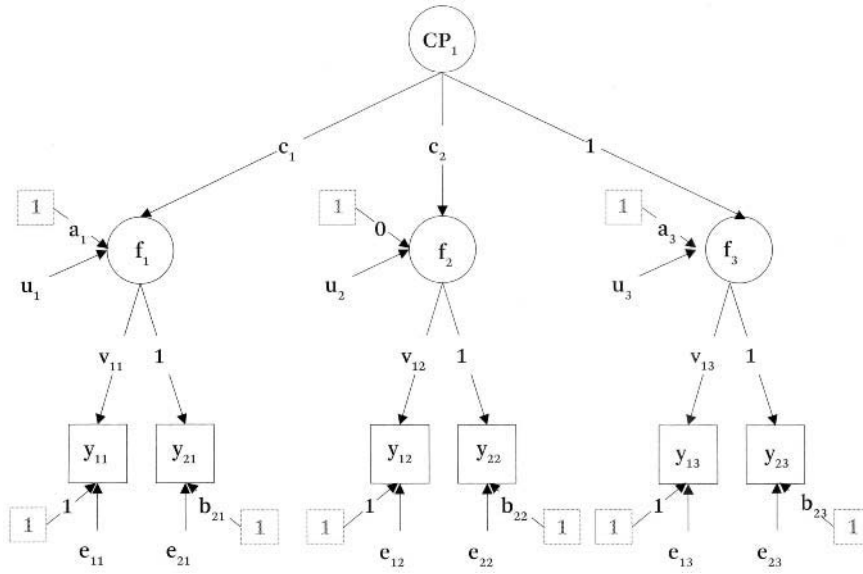


FIGURE 16.5: *The measurement model for "subjective competence" to be evaluated for three countries.*

[3] For simplicity sake we have ignored here method effects therefore the error terms are denoted by e and not by e + m as in 16.5

The above model illustrates that for identification reasons the slopes of the second indicator of each indicator are set to 1. This is possible because the same scale was applied in all three requests and, therefore, we expect the same unit of measurement. Also, the intercept of the first indicator of each concept by intuition is set to 1 because the scales have been specified[4] in such a way that an extreme value (fixed reference point) is equal to 1. Hence, we expect that a 0 value on the latent variable goes together with a score of 1 on the observed variable and not 0. Finally, the intercept for the concept "active role" was set to 0 because we expect that participants who have no competence, will respond that they cannot play a role in political activities and c_3 was fixed on 1 for identification of the variance of CP_1. For the LISREL input of this model, we refer the reader to Appendix 16.4.

The result of the analysis, assuming that the slope coefficients (c) and the intercepts (a) are equal across countries while all other coefficients are not fixed[5], was that chi^2 equals to 43 with 26 degrees of freedom (Pr=.01). Not fixing "c_2" in Spain reduced chi^2 to 34 with df=25 (Pr=.11). These results suggest that the model with one minor adjustment fits the data. The slopes of the three indicators were the same across countries except for the second one in Spain. Differences between the countries were found mainly in the measurement equations. Therefore, we can conclude that most of the differences in the parameters of the standard model (16.6) come from the measurement part of the model which we can correct. This can also be seen in Table 16.2 where the parameter values for the different countries are presented.

This table shows that the slopes of the measurement equations are quite similar in the three countries and that the intercepts are different. Furthermore, the coefficients characteristic for the relationships between the variable of interest, "subjective competence," and its three indicators are very similar, except for one deviation in Spain on the second indicator.

This result shows that the scale invariance test that is normally used to test for equivalence can lead to a rejection of the model, while the cause of the problem is not that the indicators have a different interpretation across countries, but that instead the respondents in the different countries employ the scales distinctively. However, if the differences due to the measurement procedure are corrected, rather good comparable indicators across countries may be obtained. Therefore, we prefer the less restrictive requirements of *cognitive equivalence* over the standard requirements.

---

[4] This, how ever, is not clear for the "understand" requests. In that case the extreme category cannot be called a fixed reference point. This may be the explanation for having to remove the restriction for the Spanish data set.

[5] To avoid a nonsignificant negative error variance, we set the variance of the latent variable "understand" in Spain to 0.001.

Table 16.2: Estimates of parameters of the model presented in Figure 16.5 for three countries[6]

parameter	UK	NL	Spain
Slopes of the measurement equations			
v_{11}	1.15	1.06	0.98
v_{12}	1.15	1.02	1.00
v_{13}	1.02	0.93	1.13
Intercepts of the measurement equations			
b_{21}	1.28	1.00	0.81
b_{12}	1.00	1.00	1.28
b_{22}	1.22	1.09	1.46
b_{23}	1.06	0.87	1.37
Slopes of the cognitive process equations			
c_1	-.94	-.94	-.94
c_2	.91	.91	.60
c_3	1.00	1.00	1.00
Intercepts of the cognitive process equations			
a_1	3.09	3.09	3.09
a_2	1.00	1.00	1.00
a_3	-.84	-.84	-.84

We will now discuss in the next sections whether minor deviations from cognitive equivalence (as defined above) prevent us from making comparisons of means across countries.

16.3 Comparison of means and relationships between single requests for answers

There is a strong research tradition that concentrates on differences in means [see, e.g. Torcal et al. (2005)] and relationships [see, e.g. Newton (1997)] between responses to single requests across countries. This, however, is a very questionable activity, if the requests are not previously checked for equivalence. As a consequence, we do not know the source of the response differences; it could be due to differences in measurement errors, cognitive processes, or substantive differences between countries or a combination thereof.

Previously we established that at least three observed variables for the same concept are needed in order to evaluate the quality of its measures. For separating measurement and cognitive processes, repeated observations are needed. However, most studies lack this type of information, and therefore their comparisons may be incorrect.

[6] We present only the estimated values of relevant parameters.

The magnitude of this problem can be illustrated with the items for "subjective competence": "complex," active," and "understand." It was found in the previous section that the second items were not scalar invariant for the countries United Kingdom, The Netherlands, and Spain. Given the estimated values of the parameters of the response process one can expect, assuming that the mean of the latent variable of interest is 3, that the mean for y_{12} in the UK is equal to 4.13, in The Netherlands it is 3.78, and in Spain it is 3.14. Without having collected the information about the differences in the response processes of the different countries, we could not have known that these differences have no substantive meaning, because in all three countries the mean on the latent variable is equal to 3.

For relationships between variables the same problems occur. The correlation between the variables "complex" and "understand" is -.44 in Greece and -.514 in the Czeck Republic, but after correction for measurement error, the difference is much larger. In Greece the correlation becomes -.59, and in the Czeck Republic it is -.77. In this case the differences would have been underestimated. However, the opposite can also occur. If we compare the correlation for the same variables in the Czeck Republic and Slovenia, we get for the observed correlations, respectively, -.514 and -.449. But after correction for measurement errors, these correlations are exactly equal and have the value -.77. So, in this case one could think that there are substantive differences while the differences are due merely to differences in data quality.

This overview demonstrated that the means and relationships of single requests in cross-cultural research cannot be compared unless the measurement instruments are equivalent or the differences in the response process are corrected. Given the usual lack of information, we advise to proceed with caution when attempting to compare results based on single requests.

16.4 Comparison of Means and Relationships Based on Composite Scores

In Chapter 14 we demonstrated how composite scores can be computed and evaluated. In Chapter 15 it was indicated how relationships between composite scores can be estimated. In this chapter we deal with the comparison of relationships across countries, which requires equivalence of the measurement instruments. Normally it is suggested that for comparison of relationships it is sufficient if metric invariance requirements are met while for comparison of means, scalar invariance has been required. The equality of slopes (λ) of factor models like the one in Figure 16.3 across the different countries would be required for comparison of relationships, and the equality of slopes and intercepts (τ) would be required for comparison of means. We suggested that these requirements are too strict and have proposed that these restrictions should be required for the cognitive part of the model n Figure 16.4. This means that the slopes (c) in the model of Figure 16.4 should be invariant across countries for comparison of relationships while for comparison of the means the coef-

ficients "c" and "a" should be identical across countries. The argument is that we can correct for measurement errors and that it can be shown that the slopes and intercepts of the cognitive process should be invariant.

After correction for differences in the measurement equations the composite score for CP_1 can be computed as a unweighted sum (C_1) of the three latent variables f_1–f_3:

$$C_1 = (f_1 + f_2 + f_3) \tag{16.11a}$$

or

$$C_1 = (a_1 + c_1\, CP_1 + u_1) + (a_2 + c_2\, CP_1 + u_2) + (a_3 + c_3\, CP_1 + u_3) \tag{16.11b}$$

or

$$C_1 = (a_1 + a_2 + a_3) + (c_1 + c_2 + c_3)\, CP_1 + (u_1 + u_2 + u_3) \tag{16.11c}$$

This shows that whenever one or more intercepts and/or slopes are different across countries, the computed composite score is most likely different for the different countries.[7] Even if two countries would have the same mean value on the variable of interest (CP_1), the means of the composite scores will usually be different if all coefficients in equation (16.11c) are not equal (scalar invariance) across countries. Hence, without scalar invariance the means of the computed composite scores cannot be used as indicators to compare means across countries. If scalar invariance holds true, the comparison can be made even though these means are not equal to the means of the variables of interest. In fact, the latter applies only if all the intercepts (a) are equal to 0 and the sum of the slopes (c) is equal to 1 which is quite rare.

A similar argument can be made for the comparison of relationships based on composite scores. The covariance between the variables of interest CP_1 and CP_2 is denoted by "σ_{CP1CP2}" and the composite scores for CP_1 and CP_2 are simple unweighted sums for the reaction variables (f). This means tha:

$$C_1 = f_{11} + f_{21} + f_{31} \text{ and } C_2 = f_{12} + f_{22} + f_{32} \tag{16.12a}$$

where f_{ij} represents the reaction to the ith request, which is an indicator for CP_j.

In Appendix 16.6 it is shown that:

$$\sigma_{C1,C2} = (c_{11} c_{21} + c_{31})\,(c_{12} c_{22} + c_{32})\,\sigma_{CP1CP2} \tag{16.12b}$$

This result shows that where composite scores are calculated for several different populations, the covariances across the countries cannot be compared if not all slope coefficients are invariant.[8] This is because differences between the covariances can stem from two sources: from the differences in

[7] Except for the unlikely case that some deviations cancel each other out.

[8] See note 7.

slopes or from substantive differences in covariances between the latent variables of interest. Therefore, our derivation shows that the minimum requirement for comparing relationships based on composite scores is that the slopes after correction for measurement errors are invariant.

In an earlier section we have shown how we think that one can test for scalar and metric invariance. In that analysis we have found that the indicators for "subjective competence" are not metric-invariant and therefore also not scalar-invariant. One of the items generated a different reaction in Spain than in the other two countries. Now we are left with the question of what to do. It is possible to leave one country out of the analysis. Another possibility is to omit one item and to reduce the number of indicators to two; however, in doing so, the concept by postulation will change. For our example, this option will not critically affect the analysis, because all indicators may not be necessary for the definition of the concept by postulation in a measurement model consisting of reflective indicators. Concepts defined by formative indicators are rather different. Therefore, for our example we would suggest reducing the number of indicators by one and continuing the analysis with the two items left.

We will not continue these computations here because in the two previous chapters we have already shown how to compute and evaluate composite scores, and how to estimate the relationship between them.

16.5 Comparison of Means and Relationships between Latent Variables

Composite scores are frequently used to compare the means of latent variables of interest. However, it is much easier and safer to use the estimated means of the latent variables for the comparison. In the estimation of the model of Figure 16.5 presented in Appendix 16.4 the means (ka) of the latent variable are also estimated. Even though the second indicator was not equivalent according to the popular definition of equivalence, the estimates of the means were correct because two indicators are sufficient to identify the means.

An interesting advantage of this approach is that we can test by specifying the restriction that the means in the different countries are identical, if the means are the same. Specifying this restriction we get a chi^2=53.8 with df=27 where Pr=.002. The difference with the model without the equality constraint was equal to 20.1 with df=2. So we can conclude that the means are significantly different from each other. When allowing the estimation of the means to be different they were calculated at 1.25 for the United Kingdom, 1.27 for The Netherlands, and 0.62 for Spain. This result indicates that for all three countries the mean of "subjective competence" is rather low when applying a 5-point scale. But the results also show that the "Subjective Competence" in Spain is significantly lower than in the United Kingdom or in The Netherlands.

This approach is much easier than the previous one and less prone to computation errors. Moreover, the fact that one of the indicators is not equivalent does not harm the estimates of the means and does not require any additional

effort. Scholars have described this approach as partial equivalence, stating that under this condition means of latent variables can be compared. Although this is statistically true, it should also be said that we changed the operationalization to determine the mean through two equivalent indicators, while the third indicator was treated as another effect variable of the latent variable. It is a theoretical question whether this correction of the interpretation of the measurement model is acceptable.

If the model is correctly specified, it is also possible to simultaneously estimate the relationships between the latent variables with the quality of the measurement instruments in the different countries with a minor extension of the input for the program, we have used before.

However, the problem is that the unique components of the variables for the different constructs may be correlated. These correlations will not be detected if composite scores for each concept by postulation are calculated separately, but they will be discovered when the two measurement models are combined in order to estimate a relationship between them.

Let us illustrate our point with the relationship between the variables "subjective competence" and "political trust." Both of them are defined as a concept with three reflective indicators. Appendices 16.1 and 16.2 present the requests for these two concepts. The equivalence of the measures for "subjective competence" has already been tested. The equivalence of the measures of "political trust" has been tested in the same way.

After correction for measurement error and assuming that the slopes are identical across countries, we get a chi^2 of 60.7 with 18 degrees of freedom. This is not a good fit. The program suggests that the slope for the second item in Spain should not be constrained to be equal to the same coefficients in the other countries. Repeating the analysis with this correction, improves the fit to 48.5 with 17 degrees of freedom. Without making any further improvements, we are satisfied with the results for the cognitive process part of the model. The final result is presented in Table 16.3.

Table 16.3: Unstandardized loading of the factor model for "political trust" for three countries

Indicator	UK	NL	ES
	slopes	slopes	slopes
parliament	1.00[a]	1.00[a]	1.00[a]
legal system	1.53	1.53	1.09
police	1.06	1.06	1.06

[a] These parameters have been fixed on 1 for identification.

Again, there is one item in Spain that is not invariant across countries. Previously this meant that we had to make the choice of omitting one country or one item. However, the attractive characteristic of directly estimating the relationships between latent variables is that we no longer have to make this

choice. This is because we can view the "not invariant" item as just another consequence of the latent variable defined by the other two equivalent items. In other words, allowing for a free estimation of the not invariant parameters will produce consistent estimates for the relationships between the latent variables.

Now we can specify the input for estimating the relationships between the two concepts by postulation, allowing for two deviations from metric equivalence after correction for measurement error. The model estimated is a combination of two measurement models like the one presented in Figure 16.3: one for "subjective competence" and the other for "political trust." The null model estimated is presented in Figure 16.6. Appendix 16.5 lists the LISREL input for this analysis.

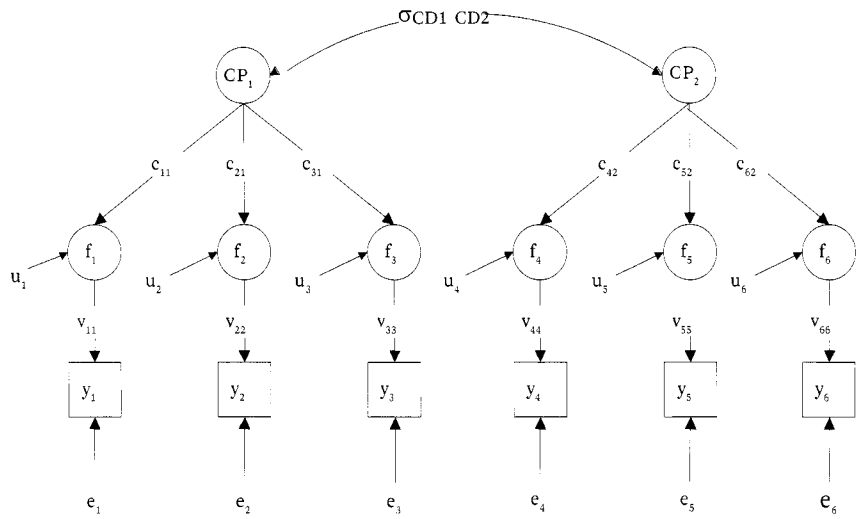


FIGURE 16.6: *The null model used to estimate the covariance between CP_1 and CP_2.*

The estimates of the parameters (v_{ij}) for the measurement equations have been inputted as fixed parameters. The results of this analysis are presented in the first row of Table 16.4 and they indicate that the fit of this model is not very good. The table shows that the covariances between "subjective competence" and "political trust" for the three countries could also be estimated in this analysis. Applying this approach, we detect that the model contains misspecifications. Therefore, before proceeding further, we need to check if the estimates are correct and to search for the misspecifications. It could be that correcting for misspecifications will change the values of the estimates as well.

Using the expected change of the parameters (EPC) in LISREL, we concluded that several coefficients have not been introduced in the model that should be there. So we introduced the following parameters in sequence for all three

countries: (c_{41}) for "trust in the government", (c_{12}) for "complexity of politics", and (c_{51}) for "trust in the legal system". These misspecifications could not have been detected without combining the two concepts into one model, because they represented effects from one concept on indicators of another concept. The specified effects suggest that:

- An increase in "subjective competence" will also increase the "trust in the government" (c_{41}) and the "legal system" (c_{51}).
- An increase in "political trust" will also increase the frequency with which people think that politics is too complex (c_{12}).

These effects suggest reduction of the unique components (u) for the different indicators that are related with variables of other concepts. Table 16.4 shows us how these corrections lead to better fit and after introducing these coefficients, only minor errors remain that will not affect the results significantly.

Table 16.4: Fit of the model and the estimated relationships between "subjective competence" and "political trust" for three countries

				Covariances in		
Model	chi^2	df	Prob	UK	NL	Spain
Null model	138.9	30	.000	.08	.25	.21
+ c_{41}	117.3	27	.000			
+ c_{12}	73.9	24	.000			
+ c_{51}	42.3	21	.004	-.11	.01	.14
Equal cov	47.4	23	.003	-.02	-.02	-.02
Cov=0	47.8	24	.004	.00	.00	.00

Abbreveations: cov= covariance, df= degrees of freedom, Prob=probability

For this final model the covariances between the concepts were estimated again. In row 4 of the table we see that these covariances differ considerably from the previous estimates, because a part of the covariances was due to misspecifications. The newly introduced effects absorb a large part of the relationships between the observed correlations, so that very little covariance is left between the two concepts. It turns out that these estimates are not significantly different from 0.

We tested this in two steps. First we tested the hypothesis that the covariances are the same across countries (Table 16.4, row 5). The test statistic does not change too much and the estimated covariance is not significantly different from 0. Second, we tested whether we could assume that there was no relationship in all three countries (Table 16.4, last row). The result also shows that this change in the model did not lead to a significant decrease in the fit.

This exercise illustrates several important points. Most importantly, this approach detects misspecifications within a model that cannot be detected

using composite scores. This is a critical point because a misspecified model can generate quite different estimates of the parameters than a correctly specified model as we have shown above.

Another important point is that we were able to perform the analysis even though in each operationalization one item in one country was not invariant, even after correction for measurement errors. We did not remove this one item. In fact, if we would have done so the last model could not have been identified anymore. Therefore, this approach is much more flexible than using composite scores.

We have also seen that we learned more about the measurement instrument with this approach than otherwise, because now we know more about the unique components of the different indicators.

Finally, we have shown that it is simple to test for the equality of the relationships of the concepts. Altogether, we hope to have shown that directly estimating the relationships between latent variables is a more efficient and therefore a better way of comparing relationships across countries than using composite scores. But, the composite scores can also be used if the model becomes too complex and one has some assurance that the model specified is correct.

16.6 Conclusions

In this chapter we have shown that cross-cultural comparisons are also affected by measurement error. They require different kinds of equivalences for measurement instruments that are sometimes not present or unknown. We have also insisted that comparing results across countries requires that the data are corrected for measurement error. Without the corrections, we run the risk of giving explanations for differences between countries on substantive grounds that could be due to differences in measurement quality of the instruments. Even though we agree with the requirements of metric and scalar equivalence, after correction for measurement error, we think that the commonly used requirements for equivalence are too strict.

A problem with this approach is that the needed information is seldom available. Especially in cases of comparing means, correlations using a single request or concept by intuition, we find that information about the quality of the requests is missing. In such instances the information about the quality can be derived from external sources such as MTMM experiments or SQP predictions.

In case of the use of composite scores for concepts-by-postulation, the comparison across countries requires perfect metric invariance for comparison of relationships and perfect scalar invariance for comparison of means. These requirements are very strict and will rarely be satisfied.

However, we have shown that comparing means and relationships between latent variables across countries does not have to require perfect invari-

ance. Consistent estimates of the means and relationships are also possible with partial equivalence. Therefore, using latent variables is a more flexible approach than employing composite scores. The only disadvantage is that the models become more complex.

Overall, the conclusion is that cross-cultural comparisons are not as simple as they seem and that the comparison of means is even more difficult than the comparison of relationships between variables.

Exercises

1. In these exercises we continue with the exercises of the previous chapter, and we add at least one more country to the analysis.
   a. Calculate the means for the variables "social trust" and "political trust."
   b. Are the means the same? What is your conclusion?
   c. Calculate the correlations between the indicators of "social trust" and "political trust" for both countries.
   d. Are these correlations similar or different?
   e. What can we say on the basis of these results about the relationship between "social trust" and "political trust"?
2. Let us now turn to measurement models for all four variables:
   a. Test the measurement models for all four variables on the basis of the data in the new country.
   b. Compare using multiple group analysis whether there is some level of invariance across the countries.
   c. Given the results, can you make comparisons across countries using these variables?
   d. If so, which comparisons can be made, and what is the result?
3. Let us now consider the use of latent variables.
   a. Given the invariance of the measurement instruments, can you make comparisons across countries using the latent variable means and/or the relationships between the latent variables?
   b. If comparisons are possible, make these comparisons and state your interpretation of the results.

APPENDIX 16.1: THE TWO SETS OF REQUESTS CONCERNING "SUBJECTIVE COMPETENCE"

The requests in the main questionnaire

C5 **CARD C4 How often do politics and government seem so complicated that you can't really understand what is going on?**

Never	*1*
Seldom	*2*
Occasionally	*3*
Regularly	*4*
Frequently	*5*
(don't know)	*8*

C6 **CARD C5 Do you think that you could take an active role in a group that is focused on political issues?**

Definitely not	*1*
Probably not	*2*
Not sure either way	*3*
Probably	*4*
Definitely	*5*
(don't know)	*8*

C7 **CARD C6 How good are you at understanding and judging political questions?**

Very bad	*1*
Bad	*2*
Neither good nor bad	*3*
Good	*4*
Very good	*5*
(don't know)	*8*

The set of requests in the supplementary questionnaire

How far do you agree or disagree with the following statements?

L4 **"Sometimes politics and government seem so complicated that I can't really understand what is going on."**

Please tick one box

Strongly disagree	☐
Disagree	☐
Neither disagree nor agree	☐
Agree	☐
Strongly agree	☐

L5 **"I think I could take an active role in a group involved with political issues."**

Strongly disagree	☐
Disagree	☐
Neither disagree nor agree	☐
Agree	☐
Strongly agree	☐

L6 **"I am good at making my mind up about political issues."**

Strongly disagree	☐
Disagree	☐
Neither disagree nor agree	☐
Agree	☐
Strongly agree	☐

APPENDIX 16.2: THE ESS REQUESTS CONCERNING "POLITICAL TRUST"

CARD C8:

Using this card, please tell me on a score of 0 to 10 how much you personally trust each of the institutions. I read out: 0 means you do not trust them at all, and 10 means you have complete trust. Firstly...**READ OUT**

		No trust At all										Complete trust	(Don't know)
C10	... the British government?	00	01	02	03	04	05	06	07	08	09	10	88
C11	... the legal system?	00	01	02	03	04	05	06	07	08	09	10	88
C12	... the police?	00	01	02	03	04	05	06	07	08	09	10	88
C13	... politicians?	00	01	02	03	04	05	06	07	08	09	10	88
C14	... the European Parliament?	00	01	02	03	04	05	06	07	08	09	10	88
C15	... the United Nations?	00	01	02	03	04	05	06	07	08	09	10	88

APPENDIX 16.3: THE STANDARD TEST OF EQUIVALENCE FOR "SUBJECTIVE COMPETENCE"

```
factor model test for Ned Spain and UK; netherlands
data ng=3 ni=3 no=1190 ma=cm
km
1.00
-.334 1.00
-.465 .389 1.00
me
3.00 2.18 2.96
sd
1.09 1.30 .986
model ny=3 ne=1 ly=fu,fi te=di,fr ps=fu,fr ty=fu,fi al=fu,fr
free ly 2 1 ly 3 1
free ty 2 ty 3
value 1 ly 1 1
out sc

Spain
data ni=3 no=280 ma=cm
km
1.00
-.306 1.00
-.508 .403 1.00
me
3.44 1.65 2.64
sd
1.20 1.044 1.175
model ny=3 ne=1 ly=in ty=in ps=sp al=sp te=di,fr
out sc

UK
data ni=3 no=885 ma=cm
km
1.00
-.317 1.00
-.381 .301 1.00
me
3.21 2.32 3.13
sd
```

```
1.122 1.355 1.069
model ny=3 ne=1 ly=in ty=in ps=sp al=sp te=di,fr
out sc
```

Appendix 16.4: The alternative equivalence test for "subjective competence" in three countries

```
Analysis of british efficacy experiment wave 1 data in ESS
Data ng=3 ni=6 no=885 ma=cm
Km
*
1.00
-.317 1.00
-.381 .301 1.00
 .538 -.328 -.373  1.00
-.330 .646 .303 -.362  1.00
-.390 .310 .489 -.401  .358  1.00

me
*
3.21 2.32 3.13 3.19 2.36 3.15
sd
*
1.122 1.355 1.069 1.049 1.121 .983
label
complex1 active1 understand1 complex2 active2 understand2
model ny=6 ne=3 nk=1 ly=fu,fi te=di,fr ps=di,fr be=fu,fi ga=fu,fi ph=sy,fr ty=fr
ka=fr al=fi
value 1 ly 4 1 ly 5 2 ly 6 3
free ly 1 1 ly 2 2 ly 3 3
value 1 ga 3 1
free ga 2 1 ga 1 1
fixed ty 1 ty 2 ty 3
value 1 ty 1 ty 2 ty 3
free al 1 al 3
start 1 ly 1 1
out
Analysis of Dutch efficacy experiment wave 1 data in ESS
Data ni=6 no=885 ma=cm
Km
```

```
1.00
-.334 1.00
-.465  .389 1.00
 .624 -.296 -.399  1.00
-.355  .671  .370  -.349   1.00
-.454  .341  .586  -.424    .420   1.00

me
*
3.00  2.18  2.96   2.89  2.25  2.98
sd
*
1.09  1.30  .986  1.07  1.133 1.055

label
complex1 active1 understand1 complex2 active2 understand2
model ny=6 ne=3 nk=1  ly=fu,fi te=di,fr ps=di,fr be=fu,fi ga=in ph=sy,fr ty=sp
ka=fr al=in
value 1 ly 4 1 ly 5 2 ly 6 3
free ly 1 1 ly 2 2 ly 3 3
fixed ty 1 ty 2 ty 3
value 1 ty 1 ty 2 ty 3
start 1 ly 1 1
out

Analysis of spanish efficacy experiment wave 1 data  in ESS
Data  ni=6 no=280 ma=cm
Km
1.00
-.306 1.00
-.508  .403 1.00
 .562 -.289 -.551  1.00
-.301  .643  .379  -.289   1.00
-.457  .311  .530  -.486    .297   1.00

me
*
3.44  1.65  2.64  3.31  1.83  2.83
sd
*
1.200 1.044 1.175 1.094  1.084  1.167
label
complex1 active1 understand1 complex2 active2 understand2
model ny=6 ne=3 nk=1  ly=fu,fi te=di,fr ps=di,fr be=fu,fi ga=in ph=sy,fr ty=sp
```

```
ka=fr al=in
value 1 ly 4 1 ly 5 2 ly 6 3
free ly 1 1 ly 2 2 ly 3 3
free ga 2 1
free ty 2
value 1 ty 1 ty 2 ty 3
fixed ps 3 3
start 1 te 6 6 te 5 5
start 1 ly 1 1
out adm=off
```

APPENDIX 16.5: THE LISREL INPUT TO ESTIMATE THE NULL-MODEL FOR ESTIMATION OF THE RELATIONSHIP BETWEEN "SUBJECTIVE COMPETENCE" AND "POLITICAL TRUST"

```
estimation of relations between pol eff and pol trust
!first group UK
data ng=3 ni=6 no=880 ma=cm
km
1.00
-.317 1.00
-.382 .301 1.00
-.153 .104 .093 1.00
-.093 -.001 .027 .456 1.00
-.027 -.034 -.063 .353 .529 1.00
me
3.21 2.32 3.13 4.75 5.16 6.16
sd
1.122 1.355 1.069 2.339 2.357 2.374
label
complex active understand parliament juridical police

model ny=6 ne=6 nk=2 ly=fu,fi te=sy,fi ga=fu,fi ph=fu,fr ps=sy,fi
value 1 ga 3 1 ga 4 2
free ga 1 1 ga 2 1 ga 5 2 ga 6 2

!later corrections
!free ga 4 1
!free ga 1 2
!free ga 5 1
```

free ps 1 1 ps 2 2 ps 3 3 ps 4 4 ps 5 5 ps 6 6

value 1.15 ly 1 1
value 1.15 ly 2 2
value 1.02 ly 3 3
value 1.17 ly 4 4
value 1.03 ly 5 5
value 1.06 ly 6 6
value .56 te 1 1
value .61 te 2 2
value .46 te 3 3
value .73 te 4 4
value 1.18 te 5 5
value .40 te 6 6

start .42 ph 1 1
start 1.14 ph 2 2
fr ph 2 1
out adm=ofns

netherlands
data ni=6 no=1150 ma=cm
km
1.00
-.334 1.00
-.465 .389 1.00
-.264 .132 .052 1.00
-.299 .170 .112 .597 1.00
-.122 .057 -.001 .445 .606 1.00
me
3.00 2.18 2.98 5.18 5.34 5.84
sd
1.093 1.295 .986 2.025 2.216 1.934
label
complex active understand parliament juridical police
model ny=6 ne=6 nk=2 ly=fu,fi te=sy,fi ga=in ph=sp ps=sp
value 1 ga 3 1 ga 6 2

!free ga 4 1
!free ga 1 2
!free ga 5 1
value 1.06 ly 1 1
value 1.02 ly 2 2
value 0.93 ly 3 3

value 1.15 ly 4 4
value 1.16 ly 5 5
value 1.11 ly 6 6
value .40 te 1 1
value .67 te 2 2
value .41 te 3 3
value .79 te 4 4
value .79 te 5 5
value .32 te 6 6
start .51 ph 1 1
start 1.32 ph 2 2

out

Spain
data ni=6 no=281 ma=cm
km
1.00
-.306 1.00
-.508 .403 1.00
-.147 .076 .158 1.00
-.064 .005 .121 .617 1.00
-.030 .080 .068 .526 .561 1.00
me
3.44 1.65 2.64 4.96 4.48 5.69
sd
1.200 1.044 1.175 2.265 2.328 2.359
label
complex active understand parliament juridical police
model ny=6 ne=6 nk=2 ly=fu,fi te=sy,fi ga=in ph=sp ps=sp
value 1 ga 3 1 ga 6 2
free ga 2 1 ga 5 2

value 0.96 ly 1 1
value 1.00 ly 2 2
value 1.13 ly 3 3
value 1.20 ly 4 4
value 1.14 ly 5 5
value 1.12 ly 6 6

value .70 te 1 1
value .36 te 2 2
value .52 te 3 3
value .16 te 4 4

value .56 te 5 5
value .55 te 6 6
start .67 ph 1 1
start 2.05 ph 2 2

out

APPENDIX 16.6: DERIVATION OF COVARIANCE BETWEEN COMPOSITE SCORES

The covariance "σ_{CP1CP2}" between the variables of interest CP_1 and CP_2 as expressed in deviation from their mean is defined for the population as follows:

$$\sigma_{CP1CP2}=\frac{1}{N}\sum (CP_1.CP_2) \text{ summed over all people (N) for the population} \quad (16.6A.1)$$

When the indicators for each latent variable are expressed in deviation from their means the relationships between the latent variables of interest and the indicators corrected for measurement error can be formulated as

$$F_{11} = c_{11}CP_1 + u_{11} \quad (16.6A.2a)$$
$$F_{21} = c_{21}CP_1 + u_{21} \quad (16.6A.2b)$$
$$F_{31} = c_{31}CP_1 + u_{31} \quad (16.6A.2c)$$
$$F_{12} = c_{12}CP_2 + u_{12} \quad (16.6A.2d)$$
$$F_{22} = c_{22}CP_2 + u_{22} \quad (16.6A.2e)$$
$$F_{32} = c_{22}CP_2 + u_{32} \quad (16.6A.2f)$$

$$\text{Assuming } cov(CP_i,u_j) = 0 \text{ for all i and j, and } cov(u_i,u_j) = 0 \text{ for all } i \neq j \quad (16.6.2g)$$

and that the means of all disturbances (u) are equal to 0.

Based on the scores from the respondents on the indicators, we can calculate composite scores of C_1 for CP_1 and C_2 for CP_2. The covariance between the composite scores is not necessarily the same as the covariance between the latent variables CP_1 and CP_2 assuming that the model is correct.[9] In order to proceed, we take a simple unweighted sum for the composite scores. This means that

$$C_1 = F_{11} + F_{21} + F_{31} \text{ and } C_2 = F_{12} + F_{22} + F_{32} \quad (16.6A.3a)$$

9 Note that this is not tested and it can lead to biased estimates.

By substituting (16.6A.2a)–(16.6A.2f) into (16.6A.3a) we get

$$C_1 = c_{11}CP_1 + u_{11} + c_{21}CP_1 + u_{21} + c_{31}CP_1 + u_{31} \quad (16.6A.3b)$$

$$C_2 = c_{12}CP_2 + u_{12} + c_{22}CP_2 + u_{22} + c_{32}CP_2 + u_{32} \quad (16.6A.3c)$$

This can be rewritten as

$$C_1 = (c_{11+}c_{21} + c_{31})CP_1 + (u_{11} + u_{21} + u_{31}) \quad (16.6A.3d)$$

$$C_2 = (c_{12+}c_{22} + c_{32})CP_2 + (u_{12} + u_{22} + u_{32}) \quad (16.6A.3e)$$

Given that the means of CP_1 and CP_2 and all disturbance terms (u) are 0, the means of C_1 and C_2 are also equal to 0; therefore the covariance of C_1 and C_2 is defined as

$$\sigma_{C1,C2} = \frac{1}{N}\sum (C_1.C_2) \quad (16.6A.4a)$$

$$\sigma_{C1,C2} = \frac{1}{N}\sum [(c_{11+}c_{21} + c_{31})CP_1 + (u_{11} + u_{21} + u_{31})][(c_{12+}c_{22} + c_{32})CP_2 + (u_{12} + u_{22} + u_{32})] \quad (16.6A.4b)$$

However, (16.6A.4b) can be simplified after multiplying it out, using (16.6A.2g) to

$$\sigma_{C1,C2} = (c_{11+}c_{21} + c_{31})(c_{12+}c_{22} + c_{32})\frac{1}{N}\sum CP_1CP_2 \quad (16.6A.4c)$$

or

$$\sigma_{C1,C2} = [(c_{11+}c_{21} + c_{31})(c_{12+}c_{22} + c_{32})\,\sigma_{CP1CP2} \quad (16.6A.4d)$$

This Page Intentionally Left Blank

References

ABELSON R. P., D. R. KINDER, M. D. PETERS, AND S. T. FISKE 1982. Affective and semantic components in political person perception. *Journal of Personality and Social Psychology*, 42, 619–630.

AJZEN I., AND M. FISHBEIN 1980. Understanding attitudes and predicting social behaviour. The expectancy-value model. *Actes du Congrès de l'AFM (Poitiers), 681–695.*

AJZEN I. 1989. Attitude structure and behavior. In A. R. Pratkanis, S. J. Breckler and A. G. Greenwald (eds.), *Attitude Structure and Function*, Hillsdale, NJ: Erlbaum, 241–247.

AJZEN I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179–211.

ALLISON P. D. 1987. Estimation of linear models with incomplete data. In C. C. Clogg (ed.), *Sociological Methodology*, Washington DC: American Sociological Association, 71–103.

ALTHAUSER R. P., T. A. HEBERLEIN, AND R. A. SCOTT 1971. A causal assessment of validity: The augmented mulitrait-multimethod matrix. In H. M. Blalock Jr. (ed.), *Causal Models in the Social Sciences*. Chicago: Aldine, 151–169.

ALWIN D. F. 1974. An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimehod matrix. In H. L. Costner (ed.), *Sociological Methodology*, San Francisco: Jossey Bass, 79–105.

ALWIN D. F., AND J. A. KROSNICK 1991. The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, 20, 139–181.

ALWIN D. F. 1997. Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods and Research*, 25, 318–341.

ANDREWS F. M., AND S. B. WITHEY 1974. Developing measures of perceived life-quality: Results from several surveys. *Social Indicators Research*, 1, 1–26.

ANDREWS F. M. 1984. Construct validity and error components of survey measures: A structural equation approach. *Public Opinion Quarterly*, 48, 409–442.

AQUILINO W. S. 1993. Effects of spouse presence during the interview on survey responses concerning marriage. *Public Opinion Quarterly*, 57, 358–376.

AQUILINO W. S. 1994. Interview mode effects in surveys of drug and alcohol use. *Public Opinion Quarterly*, 58, 210–240.

AQUILINO W. S., AND L. LOSCIUTO 1990. Effect of interview mode on self-reported drug use. *Public Opinion Quarterly*, 54, 362–395.

ARMINGER G., AND M. E. SOBEL 1991. Pseudo-maximum likelihood estimation of mean and covariance structures with missing data. *Journal of the American Statistical Association*, 85, 195–203.

BAGOZZI R. P. 1989. An investigation in the role of affective and moral evalua-

tions in the purposeful behavior model of attitude. *British Journal of Social Psychology,* 28, 97–113.

BAGOZZI R. P., AND Y. YI 1991. Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research,* 17, 426–439.

BARKER M. 1981. *The New Racism: Conservatives and the Ideology of the Tribe.* London: Junction Books.

BARTELDS J. F., E. P JANSEN, AND TH. H. JOOSTEN 1994. *Enquêteren: Het pstellen en gebruiken van vragenlijsten.* Groningen:Wolters-Noordhoff.

BELSON W. 1981. *The Design and Understanding of Survey Questions.* London: Gower.

BEMELMANS-SPORK M., AND D. SIKKEL 1986. Data collection with handheld computers. *Proceedings of the International Statistical Institute,* 3, Voorburg: International Statistical Institute.

BILLIET J. G. LOOSVELDT, AND L. WATERPLAS 1986. *Het Survey-interview onderzocht:Effecten van het onderwerp en gebruik van vragenlijsten op de kwaliteit van antwoorden.* Leuven: Sociologisch Onderzoeksinstituut KU Leuven.

BILLIET J. G., AND J. MCCLENDON 2000. Modelling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling,* 7, 608–629.

BISBE J. J., M. BATISTA-FOGUET, AND R. CHENHALL *(forthcoming). What do we Really Mean by Interactive Control Systems? The Risk of Theoretical Misspecification.* Barcelona: ESADE.

BLALOCK H. M. JR. 1964. *Causal Inferences in Nonexperimental Research,* Chapel Hill, NC: University of North Carolina Press.

BLALOCK H. M. JR. 1968. The measurement problem: A gap between languages of theory and research. In H. M. Blalock, and A. B. Blalock (eds), *Methodology in the Social Sciences.* London: Sage, 5–27.

BLALOCK H. M. JR. 1990. Auxilary measurement theories revisited. In Hox J. J., and J. de Jong-Gierveld (eds.), *Operationalization and Research Strategy.* Amsterdam: Swets and Zeitlinger, 33–49.

BOBO L., J. R. KLUEGEL, AND R. A. SMITH 1997. Laissez faire racism: The crystallization of a kinder, gentler anti-black ideology. In S. A. Tuch, and J. K. Martin (eds.), *Racial Attitudes in the 1990's: Continuity and Change.* Westport CT: Praeger, 76–90.

BOLLEN K. A. 1989. *Structural Equations with Latent Variables.* New York: Wiley

BOLLEN K. A., AND R. LENNOX 1991. Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin,* 110, 305–314.

BRADBURN N. M., AND S. SUDMAN 1988. *Polls and Surveys. Understanding what they tell us.* San Francisco: Jossey Bass.

BRANNICK M. T., AND P. E. SPECTOR 1990. Estimation problems in the block-diagonal model of the multitrait-multimethod matrix. *Applied Psychological Measurement,* 14, 325–339.

BROWNE M. W. 1984. The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology,* 37, 1–21.

BUNTING B., AND G. ADAMSON 2000. Assessing reliability and validity in the context of planned incomplete data structures for multitrait-multimethod models. In A. Ferligoj, and A. Mrvar (eds.), *Developments in Survey Methodology, Metodološki Zvezki,* 15. Ljubljana: FDV, 37–53.

BÜTSCHI D. 1997. *Informations et opinions: Promesses et limites du questionnaire de choix. Ph.D. thesis,* University of Genève.

CAMPBELL A., P. E. CONVERSE, W. E. MILLER, AND D. E. STOKES 1960. *The American Voter.* New York: Wiley.

CAMPBELL, D. T., AND D. W. FISKE 1959. Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin,* 56, 81–105.

CAMPBELL D. T., AND E. J. O'CONNELL 1967. Method factors in multitrait-multimethod matrices: multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409–426.

CARPENTER E. H., AND M. JUST 1975. Sentence Comprehension: A psychological model of verification. *Psychological Review*, 82, 45–73.

CHISHOLM W. E., L.T. MILIC, AND J. A. GREPIN 1984. *Interrogativity: A Colloquium on the Grammar Typology and Pragmatics of Questions in Seven Diverse Languages*. Amsterdam: J. Benjamins.

CLARK H. H., AND E. V. CLARK 1977. *Psychology and Language*. New York: Harcourt.

COCHRAN W.G. 1977. *Sampling Techniques*. New York: Wiley.

COENDERS G., AND W. E. SARIS 1998. Relationship between a restricted correlated uniqueness model and a direct product model for multitrait-multimethod data. In A. Ferligoj (ed.), *Advances in Methodology, Data Analysis and Statistic., Metodološki Zvezki* 14. Ljubljana: FDV, 151–172.

COENDERS, G., W. E. SARIS, J. M. BATISTA-FOGUET, AND A. ANDREENKOVA 1999. Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling*, 6, 135–157.

COENDERS, G., AND W. E. SARIS 2000. Testing nested additive, multiplicative and general multitrait-multimethod models. *Structural Equation Modeling*, 7, 219–250.

COLEMAN J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology*, 94, Supplement S95–S120.

COLEMAN J. S. 1990. *Foundations of Social Theory*. Cambridge MA: Belknap Press of Harvard University.

CONVERSE P. 1964. The nature of belief systems in mass publics. In D. A. Apter (ed.), *Ideology and Discontent*, New York: Free Press, 206–261.

CONVERSE J. M., AND H. SCHUMAN 1984. The manner of inquiry. An analysis of survey questions from across organizations and over time. In C. F. Turner, and E. Martins (eds.), *Surveying Subjective Phenomena*, 2, New York: Russel Sage Foundation, 283–316.

CONVERSE J. M., AND S. PRESSER 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage.

COOMBS C. H. 1964. *A theory of data*. New York: Wiley.

CORNELIUS R. R. 1996. *The Sciences of Emotion. Research and Tradition in the Psychology of Emotions*. New Jersey: Prentice Hall.

CORTEN I., W. E. SARIS, G. COENDERS, W. VAN DER VELD, C. ALBERS, AND C. CORNELIS 2002. The fit of different models for multitrait-multimethod experiments. *Structural Equation Modeling*, 9, 213–232 .

COUPER M. P., S. E. HANSEN, AND S. A. SADOVSKI 1997. Evaluating interviewer use of CAPI. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewind (eds.), *Survey Measurement and Process Quality*. New York: Wiley, 267–285.

COUPER M. P., R. P. BAKER, J. BETHLEHEM, C. Z. F. CLARK, J. MARTIN, W. L. NICHOLLS II AND J. M. O'REILLY (eds.) 1998. *Computer Assisted Survey Information Collection*. New York: Wiley.

COUPER M. P. 2000. Web Surveys. A review of issues and approaches. *Public Opinion Quarterly*, 64, 464–494.

COX E. P. 1980. The optimal number of response alternatives for a scale. *Journal of Marketing research*, 17, 407–422.

CUDECK, R. 1988. Multiplicative models and MTMM matrices. *Journal of Educational Statistics*, 13, 131–147.

CRONBACH L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 294–334.

DANIEL F. 2000. *QUAID Question Taxonomy*. Internet site wwww.psyc.memphis.edu/quaid. html.

DANIELSON L., AND P. A. MAARSTAD 1982. *Statistical Data, Collection with*

Hand-held Computers: A Consumer Price Index. Orebro: Statistics Sweden.

DAVIDOV E. P. SCHMIDT, AND S. SCHWARTZ 2007. Bringing values back in. The adequacy of the European Survey to measure values in 20 countries. *Sociological Methods and Research*.

DE GROOT A. D., AND F. L. MEDENDORP 1986. *Term, Begrip, Theorie: Inleiding tot Signifische Begripsanalyse*. Meppel: Boom.

DE HEER W. 1999. International response trends: results of an international survey. *Journal of Official Statistics*, 15, 129–142.

DE PIJPER W. M., AND W. E. SARIS 1986. *The Formulation of Interviews Using the Program Interv*. Amsterdam: Sociometric Research Foundation.

DIJKSTRA W., AND J. VAN DER ZOUWEN 1982. *Response Behaviour in the Survey-Interview*. London: Academic Press.

DILLMAN D. A. 1978. *Mail and Telephone Survey: 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. 2000. *Mail and Internet Surveys. The Tailored Design Method*. New York: Wiley.

DUNCUN O. D. 1975. *Introduction to Structural Equation Models*. New York: Academic Press.

EAGLY A. H., AND S. CHAIKEN 1993. *The Psychology of Attitudes*. New York: Harcourt-Brace -Jovanovich.

EDWARDS J. R., AND R. P. BAGOZZI 2000. On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5, 155–174.

EID M. 2000. Multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.

EMANS B. 1990. *Interviewen, Theorie, Techniek en Training*, Groningen: Stenfert Kroese.

ERICSSON K. A., AND H. A. SIMON 1984. *Protocol Analysis: Verbal Reports as Data*. Cambridge MA: MIT Press.

ESPOSITO J. P. C. CAMPANELI, J. ROTHGEB, AND A. E. POLIVKA 1991. Determining which questions are best: Methodologies for evaluating survey questions. In *Proceedings of the section on survey methods research*, American Statistical Association, 46–55.

ESPOSITO J. P., AND J. M. ROTHGEB 1997. Evaluating survey data: Making the transition from pretesting to quality assessment. In P. Lyberg, P. Biemer, L. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: Wiley, 541–571.

ESSED P. 1984. *Alledaags Racisme*. Amsterdam: Feministische Uitgeverij SARA.

EUROPEAN SOCIAL SURVEY 2002. *European Social Survey Round 1: Report of the First Year*, London: NatCen.

FELDMAN S. 1989. Reliability and stability of policy positions: evidence from a five-wave panel. *Political Analysis*, 1, 25–60.

FERLIGOJ A., AND V. HLEBEC 1999. Evaluation of social network measurement instruments. *Social Networks*, 21, 111–130.

FISHBEIN M., AND I. AJZEN 1975. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading MA: Addison Wesley.

FORSYTH B. H., J. T. LESSLER AND M. L. HUBBARD 1992. Cognitive evaluation of the questionnaire. In C. F. Tanur and R. Tourangeau (eds.), *Cognition and Survey Research*, New York: Wiley, 183–198.

FOWLER F. J., AND T. W. MANGIONE 1990. *Standardized Survey Interviewing. Minimizing Interview-Related Error*. Newbury Park: Sage.

GAERTNER S. L., AND J. F. DOVIDIO (eds.), 1986. *The Aversive Form of Racism. Prejudice, Discrimination and Racism*. New York: Academic Press.

GALLHOFER I. N., AND W. E. SARIS 1995. *Foreign Policy Decision-Making. A Qualitative and Quantitative Analysis of Political Argumentation*. Westport CT: Praeger.

GFÖRER J. C., AND A. L. HUGHES 1992. Collecting data on illicit drug use by phone. In C. F. Turner, J. T. Lessler, and J. C. Gförer (eds.), *Survey measurement of drug use: Methodological studies.* Rockville MD: National Institute on Drug Abuse, 277–295.

GINZBURG J. 1996. Interrogatives: Questions, facts and dialogue. In S. Lappin (ed.), *The Handbook of Contemporary Semantic Theory.* Cambridge MA: Blackwell, 385–421.

GIVON T. 1984. Syntax. *A Functional-Typological Introduction* Vol. I–II. Amsterdam: J. Benjamin.

GRAESSER A. C., C. L. MCMAHEN, AND B. K. JOHNSON 1994. Question asking and answering. In M. Gernsbacher (ed.), *Handbook of Psycholinguistics*, San Diego, CA: Academic Press, 517–538.

GRAESSER, A.C., P. K. WIEMER-HASTINGS, R. KREUZ, AND P. WIEMER-HASTINGS, 2000 a. QUAID: A questionnaire evaluation aid for survey methodologists. *Behavior Research Methods, Instruments, and Computers, 32, 254 – 262.*

GRAESSER, A.C., K. WIEMER-HASTINGS, P. WIEMER-HASTINGS, AND R. KREUZ 2000 b. The gold standard of question quality on surveys: Experts, computer tools, versus statistical indices. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, Washington DC: American Statistical Association, 459–464.

GROENENDIJK J., AND M. STOKHOF 1997. Questions. In J. van Benthem and A. ter Meulen (eds.), *Handbook of Logic and Language*, Amsterdam: Elsevier, 1055–1124.

GROUZET F. M. E, N. OTIS, AND L. G. PELLETIER 2006. Longitudinal cross-gender factorial invariance of the academic motivation. *Structural Equation Modeling*, 13, 73 – 98.

GROVES, R. M. 1989. *Survey Errors and Survey Costs.* New York: Wiley.

GROVES R. M., AND M. P. COUPER (1998). *Nonresponse in Household Interview Surveys.* New York: Wiley.

GUTTMAN L. 1954. A new approach to factor analysis: The Radex. In P. F. Lazersfeld (ed.), *Mathematical Thinking in the Social Sciences.* Glencoe: The Free Press, 258–348.

GUTTMAN L. 1981. Definitions and notations for the facet theory of questions. In I. Borg (ed.), *Multidimensional Data Representations: When and Why.* Ann Arbor: Mathesis Press, 95–125.

GUTTMAN L. 1986. Science and empirical scientific generalizations. In H. Gratch (ed.), *Research Report 1984–1985.* Jerusalem: Israel Institute of Applied Research.

HALPERN D. 2005. *Social Capital.* Malden MA: Polity Press.

HAMBLETON R. K, AND H. SWAMINATHAN 1985. *Item Response Theory. Principles and Applications.* Boston: Kluwer-Nijhoff.

HAMBLIN R.L. 1974. Social attitudes: Magnitude measurement and theories. In H.M. Blalock (ed.), *Measurement in the Social Sciences.* Chicago: Aldine, 61–120.

HARARY F. 1971. *Graph Theory.* London: Addison-Wesley.

HARKNESS J. A., F. J. R. VAN DE VIJVER, AND P. PH. MOHLER (eds.), 2003. *Cross-Cultural Survey Methods.* Hoboken NJ: Wiley.

HARKNESS J. A. 2003. Questionnaire translation. In Harkness J. A., F. J. R. Van de Vijver, and P. Ph. Mohler (eds.), *Cross-Cultural Survey Methods.* Hoboken NJ: Wiley, 35–56.

HARREL L., AND R. CLAYTON 1991. *Voice Recognition Technology in Survey Data Collection: Results of the first field tests.* Paper presented at the National Field Technologies Conference, San Diego CA.

HARRIS Z. 1978. The interrogative in a syntactic framework . In Hiz H. (ed.), *Questions.* Dordrecht: Reidel, 37–89.

HARTMAN H., AND W. E. SARIS 1991. *Data Collection on Expenditures.* Paper presented at the Workshop on Diary Surveys, Stockholm Sweden.

HELMERS H. M., R. J. MOKKEN, R. C. PLIJTER, AND F. N. STOKMAN 1975.

Graven naar Macht. Op zoek naar de Kern van de Nederlandse Economie. Amsterdam: Van Gennep.

HEISE D. R. 1969. Separating reliability and stability in test-retest-correlation. *American Sociological Review*, 34, 93–101.

HEISE D. R., AND G. W. BOHRNSTEDT 1970. Validity, invalidity and reliability. *Sociological Methodology*, 2, 104–129.

HIPPLER H. J., AND N. SCHWARZ 1987. Response effects in surveys. In H. J. Hippler, N. Schwarz, and S. Sudman (eds.), *Social Information Processing and Survey Methodology*. New York: Springer Verlag, 102–122.

HOLSTI O. R. 1996. *Public Opinion and American Foreign Policy*. Ann Arbor: The University of Michigan Press.

HOMANS G. C. 1965. *The Human Group*. London: Routledge- Kegan.

HORN J. L., J. MCARDLE, AND R. MASON 1983. When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist* 1, 179–188.

HOX J. J. 1997. From theoretical concept to survey questions. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*. New York: Wiley, 47–70.

HUDDLESTON R. 1988. *English Grammar: An outline*. Cambridge: Cambridge University Press.

HUDDLESTON R. 1994. The contrast between interrogatives and questions. *Journal of Linguistics* 30, 411–439.

HURFORD J. R., AND B. HEASLEY 1994. *Semantics: A Course Book*. Cambridge: Cambridge University Press.

HYMAN H. H., AND P. B. SHEATSLEY 1950. The current status of American public opinion. In J. C. Payne (ed.), *The Teaching of Contemporary Affairs, Twenty-first Yearbook of the National Council of Social Studies*, Princeton NJ: Princeton University Press, 11–34.

JOBE J. B., W. F. PRATT, R. TOURANGEAU, A. BALDWIN, AND K. RASINSKI 1997. Effects of interview mode on sensitive questions in a fertility survey. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, New York: Wiley, 322–329.

JÖRESKOG K. G. 1971. Simultaneous factor analysis in several populations, *Psychometrika* 34, 409–426.

JÖRESKOG K. G., AND D. SÖRBOM 1989). *LISREL 7. A Guide to the Program and Applications*. Chicago: SPSS Inc.

KAASE M., AND S. BARNES. 1979. *Political Action: Mass Participation in Five Western Democracies*. Beverly Hills: Sage.

KALFS P. 1993. *Hour by Hour. Effects of the Data Collection Mode in Time Use Research. PhD. thesis*, University of Amsterdam.

KALTON G. 1983. *Introduction to Survey Sampling*. Newbury Park: Sage.

KAPER E. 1999. *Panel Effects in Consumer Research. Statistical Models for Underreporting*. Amsterdam: Tinbergen Institute Research Series.

KAPLAN D. 2000. *Structural Equation Modeling: Foundations and Extensions*. London: Sage.

KAY A. F. 1998. *Locating Consensus for Democracy. A Ten-Year U.S. Experiment*. St. Augustine FL: American Talks Issues.

KELLEY H. H., AND J. L.MICHELA 1980. Attribution theory and research. *Annual Review of Psychology*, 31, 475–501.

KENNY D. A. 1976. An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247–252.

KENNY D. A., AND D. A. KASHY 1992. Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112 , 165–172.

KIESLER S., AND L. S. SPROULL 1986. Response effects in the electronic survey. *Public Opinion Quarterly*, 50, 402–413.

KINDER D. R., AND D. O. SEARS 1981. White opposition to busing: On

conceptualizing and operationalizing group conflict. *Journal of Personal and Social Psychology*, 40, 414–431.

KISH L. 1965. *Survey Sampling*. New York: Wiley.

KLINGEMANN H. D. 1997. The left-right self-placement question in face to face and telephone surveys. In W. E. Saris, and M. Kaase (eds.), *Eurbarometer Measurement Instruments for Opinions in Europe, Zuma- Nachrichten Spezial*, Mannheim: ZUMA, 2, 113–123.

KNOKE D., AND J. H. KUKLINSKI 1982. *Network Analysis. Beverly Hills: Sage.*

KOGOVŠEK T., A. FERLIGOJ, G. COENDERS, AND W. E. SARIS 2001. Estimating reliability and validity of personal support measurements: full information ML estimation with planned missing data. *Social Networks*, 24, 4- 20.

KÖLTRINGER R. 1993. *Gültigkeit von Umfragedaten*. Wien: Bohlau.

KÖLTRINGER R. 1995. Measurement quality in Austrian personal interview surveys. In W. E. Saris, and A. Münnich (eds.), *The Multitrait-Multimethod Approach to evaluate measurement instruments*, Budapest: Eötvös University Press, 207–225.

KONING P. L., AND P. J. VAN DER VOORT 1997. *Sentence Analysis*. Groningen: Wolters- Noordhoff.

KOVEL J. 1971. *White Racism: A Psychohistory*. London: Allen- Lane.

KRECH D., AND R. CRUTCHFIELD 1948. *Theories and Problems in Social Psychology*. New York: McGraw Hill.

KRECH D., CRUTCHFIELD R., AND E. BALLACHEY 1962. *Individual in Society*. New York:McGraw Hill.

KRIESI H. 1993. *Political Mobilization and Social Change. The Dutch Case in Comparative Perspective*. Aldershot: Averbury.

KROSNIK J. A., AND H. SCHUMAN 1988. Attitude intensity, importance and certainty and susceptibility to response effects. *Journal of Personality and Social Pschology*, 54, 940–952.

KROSNICK J. A. 1991. Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 201–219.

KROSNICK J. A., AND R. P. ABELSON 1991. The case for measuring attitude strength in surveys. In J. M. Tanur (ed.), *Questions about Questions. Inquiries into the Cognitive Bases of Surveys*, New York: Russel Sage Foundation, 177–203.

KROSNICK J. A., AND L. R. FABRIGAR 1997. Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, New York: Wiley, 141–164.

KROSNICK J. A. AND L.R. FABRIGAR *(forthcoming). Designing Good Questionnaires: Insights from Cognitive Psychology.*

LAMBRECHT K. 1995. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

LASS J., W. E. SARIS, AND M. KAASE 1997. Sizes of the different effects: coverage, mode and non-response. In W. E. Saris, and M. Kaase (eds.), *Eurbarometer Measurement Instruments for Opinions in Europe*, ZUMA–Nachrichten Spezial 2, Mannheim: ZUMA, 73–86.

LAWLEY D. N., AND A. E. MAXWELL 1971. *Factor Analysis as a Statistical Method*. London: Butterworth.

LEHNERT W. G. 1977. Human and Computational Question Asking. *Cognitive Science*, 1, 47–73.

LESSLER J. T., AND B. H. FORSYTH 1966. A coding system for appraising questionnaires. In N. Schwarz, and S. Sudman (eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, San Francisco: Jossey Bass, 259–292.

LITTLE T. D. 1997. Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.

LODGE M. , J. TANNENHAUS, D. CROSS, B. TURSKY, M. A. FOLEY, AND M. FOLEY 1976. The calibration and cross model validation of ratio scales of political opinion in survey research. *Social Science Research*, 5, 325–347.

LODGE M. 1981. *Magnitude Scaling. Quantitative Measurement of Opinions.* Beverly Hills: Sage.

LODGE M., AND B. TURSKY 1981. On the magnitude scaling of political opinion in survey research. *American Journal of Political Science*, 25, 376–419.

LORD F., AND M. R. NOVICK 1968. Statistical Theories of Mental Test Scores. Reading MA: Addison–Wesly.

MARSH H. W. 1989. Confirmatory factor analysis of multitrait-multimethod data: many problems and few solutions. *Applied Psychological Measurement* , 13, 335–361.

MARSH, H. W., AND L. BAILEY 1991. Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement,* 15, 47–70.

MCCONAHAY J. B., AND J. C. HOUGH JR. 1976. Symbolic racism. *Journal of Social Issues*, 32, 23–45.

MARTINI M. C. 2001. *Assessment of Erroneous but Compatible Answers in Social Surveys. PhD thesis* , University of Padua.

MEREDITH W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.

MESSICK L. 1989. Validity. In R. L. Linn (ed.), Educational measurement, New York: Macmillan, 13–103

MIETHE T. D. 1985. The validity and reliability of value measurements. *Journal of Psychology,* 119, 441–453.

MILLER G.A. 1956. The magical number seven plus or minus two - some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.

MOKKEN R. J. 1971. *A Theory and Procedure of Scale Analysis with Applications in Political Research*. New York: Walter–Gruyter–Mouton.

MOLENAAR. N. J. 1986. *Formuleringseffecten in Survey-Interviews. PhD thesis*, Amsterdam: Free University.

MÜNNICH A. 2004. *Judgement and Choice. PhD thesis*, University of Amsterdam.

MUTHEN B., D. KAPLAN, AND M. HOLLIS 1987. On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462.

NEIJENS P. 1987. *The Choice Questionnaire. Design and Evaluation of an Instrument for Collecting Informed Opinions of a Population. PhD thesis*, Amsterdam: Free University.

NEWTON K. 1997. Social capital and democracy. *American Behavioral Scientist*, 40, 575–586.

NEWTON K. 1999. Social trust and political trust in established democracies. In P. Norris (ed.), Critical Citizens, Oxford: Oxford University Press, 169–187.

NEWTON K. 2001. Social *trust and political disaffection: Social Capital and Democracy.* Paper presented at the EURESCO Conference on Social Capital: Interdisciplinary perspectives, Exeter, UK.

NISBETT R. E., AND T. D. WILSON 1977. Telling more than we know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.

NORRIS P. 1999. *Critical Citizen.* Oxford: Oxford University Press.

NORTHROP F. S. C. 1947. *The Logic of the Sciences and the Humanities.* New York: World Publishing Company.

NUNNALLY J. C, AND L. H. BERNSTEIN 1994. *Psychometric Theory.* New York: McGraw Hill

OBERSKI D., L. KUIPERS, AND W.E. SARIS 2005. *SQP Survey Quality Predictor.* www.sqp.nl

OPPENHEIM A.N. 1966. *Questionnaire Design and Attitude Measurement.* London: Heinemann.

OSKAMP S. 1991. *Attitudes and Opinions.* Englewoods Cliffs NJ: Prentice Hall.

PARSONS T. 1951. *The Social System.* Glencoe: Free Press.

PAYNE S. 1951. *The Art of Survey Questions*. Princeton: Princeton University Press.

PETTIGREW T. F., AND R. W. MEERTENS 1995. Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, 25, 57–75.

PHIPPS P. A., AND A. R. TUPEK 1991. Assessing measurement errors in a touchtone recognition survey. *Survey Methodology*, 17, 15–26.

PIAZZA T., AND P. M. SNIDERMAN 1991. Incorporating experiments into computer-assisted surveys. In Couper M. P., R. P. Bakker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II., and J. M. O´Reilly (eds.), *Computer Assisted Survey Information Collection*, Hoboken: Wiley, 167–185.

POULTON E. C. 1968. The new psychophysics: Six models for magnitude estimation judgments. *Psychological Bulletin*, 69, 1–19.

PRESSER S. 1984 . The use of survey data in basic research in the social sciences. In. C. F. Turner, and E. Martin (eds.), *Surveying Subjective Phenomena*: New York: Russell Sage Foundation, 110–132.

PRESSER S., AND J. BLAIR 1994. Survey pretesting: Do different methods produce different results? In P.V. Marsden (ed.), *Sociological Methodology*, Oxford: Basil Blackwell, 73–104.

PRUCHNO R. A., AND J. M. HAYDEN 2000. Interview modality: effects on costs and data quality in a sample of older women. Journal of Aging and Health, 12, 3–24.

PUTNAM R. D. 1993. *Making Democracy Work. Civic traditions in Modern Italy*. Princeton: Princeton University Press.

PUTNAM R. D. 2000. *Bowling Alone*. New York: Simon and Schuster.

QUIRK R., S. GREENBAUM, G. LEECH, AND J. SVARTVIK 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

RABINOWITZ G., S. E. MACDONALD, AND O. LISHUAG 1991. New players in an old game: party strategy in multiparty systems. *Comparative Political Studies*, 24, 147–185.

RASCH G. 1960. *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Education Research.

RAYKOV T. 1997. Scale reliability. Cronbach's coefficient alpha and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32, 329– 353.

RAYKOV T. 2001. Bias of Cronbach's coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, 25, 69–76.

RICHARDS J., C. J. PLATT, AND H. PLATT 1993. *Dictionary of Language Teaching and Applied Linguistics*. Harlow: Longman.

RODGERS W. L., F. M. ANDREWS, AND A. R. HERZOG 1992. Quality of survey measures: A structural modelling approach. *Journal of Official Statistics*, 8, 251–275.

ROKEACH M. 1973. *The Nature of Human Values*. New York: Free Press.

SARIS W. E., C. BRUINSMA, W. SCHOOTS, AND C. VERMEULEN 1977. The use of magnitude estimation in large scale survey research. *Mens en Maatschappij*, 52, 369–359.

SARIS W. E., W. M. DE PIJPER, AND J. MULDER 1978. Optimal procedures for estimation of factor scores. In: *Sociological Methods and Research*, 7, 85–105.

SARIS W. E. 1981. Different questions, different variables. In C. Fornell (ed.), *A Second Generation of Multivariate Analysis*, Vol. 2., New York: Praeger, 78–96.

SARIS W. E., M. DE PIJPER, AND P. NEIJENS 1982. Some notes on the computer steered interview. In C. P. Middendorp (ed,), *Handelingen van het Congres gehouden in Rotterdam*. Rotterdam: Dutch Sociometric Society, 95-104.

SARIS W. E., AND H. STRONKHORST 1984. *Causal Modeling in Nonexperimental Research: An Introduction to the LISREL Approach*. Amsterdam: Sociometric Research Foundation.

SARIS W. E., P. NEIJENS, AND J. A. DE RIDDER 1984. Resultaten van de keuze-

enquête in het kader van de B.M.D. In Stuurgroep Maatschappelijke Diskussie Engergiebeleid: *Het Eindrapport, Appendix*. Leiden: Stenfert Kroese.

SARIS W. E., AND W. M. DE PIJPER 1986. Computer assisted interviewing using home computers. *European Research*, 14, 144–152.

SARIS W. E, A. SATORRA, AND D. SÖRBOM 1987. Detection and correction of structural equation models. *Sociological Methodology*, 17, 105–131.

SARIS W. E. 1988. A measurement model for psychophysical scaling. Quality and Quantity, 22, 417–483.

SARIS W. E. (ed.), 1988. Variations in Response Functions: a Source of Measurement Error in Attitude Research. Amsterdam: Sociometric Research Foundation.

SARIS W. E., AND K. DE ROOY 1988. What kind of terms should be used for reference points. In W. E. Saris (ed.), Variations in Response Functions: A Source of Measurement Error in Attitude Research, Amsterdam: Sociometric Research Foundation, 199–219.

SARIS W. E., AND A. SATORRA 1988. Characteristics of structural equation models which affect the power of the likelihood ratio test. In W. E. Saris, and I. N. Galhofer (eds.), *Sociometric Research, Vol. 2,. Data Analysis*, London: MacMillan, 220–236.

SARIS W. E. 1990. The choice of a model for evaluation of measurement instruments. In W. E. Saris, and A. van Meurs (eds.), *Evaluation of Measurement Instruments by Meta-analysis of Multitrait-Multimethod studies*, Amsterdam: North Holland, 118–133.

SARIS W. E., AND F. M. ANDREWS 1991. Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, 575–599.

SARIS W. E. 1991. *Computer Assisted Interviewing*. Newbury Park CA: Sage.

SARIS W. E. 1996. A facet design to describe measures for ethnocentrism. *Proceedings of the IRMCS conference in Preddvor*, Slovenia 1996.

SARIS W. E. 1997. Comparability across mode and country. In W. E. Saris, and M. Kaase (eds.), *Eurobarometer Measurement for Opinions in Europe. ZUMA-Nachrichten Spezial 2.*, Mannheim: ZUMA, 125–139.

SARIS W. E., AND M. KAASE 1997 (eds.). Eurobarometer Measurement for Opinions in Europe. *ZUMA-Nachrichten Spezial 2*, Mannheim: ZUMA, 125–139.

SARIS W. E. 1998. Ten years of interviewing without interviewers: The telepanel. In M. P. Couper , R. P. Baker, J. Bethlehem, C. Clark, J. Martin, W. L. Nicholls II, and J. M. O'Reilly (eds.), *Computer-assisted Survey Information Collection*, New York: Wiley, 409–431.

SARIS W. E., AND I. N. GALLHOFER 1998. Classificatie van survey-vragen. *Tijdschrift voor Communicatie wetenschap*, 2, 96–122.

SARIS W. E. 1998. Words are sometimes not enough to express the existing information. In M. Fenema, C. van der Eyck, and H. Schijf (eds.), *In Search of Structure: Essays in Social Science and Methodology*, Amsterdam: Het Spinhuis, 98–115.

SARIS W. E., AND I. N. GALLHOFER 2001. *Report on the MTMM experiments in the pilot studies and proposals for Round 1 of the ESS. Report for the ESS*. London, ESS.

SARIS W. E., AND I. N. GALLHOFER 2002. Cross-cultural research comparability: The effects of random and systematic errors. Report for the ESS. London, ESS.

SARIS W. E. 2003. Multitrait-multimethods studies. In Harkness J. A., F. J. R. Van de Vijver, and P. Ph. Mohler (eds.), *Cross-Cultural Survey Methods*. Hoboken NJ: Wiley, 265–274.

SARIS W. E., AND C. AALBERTS 2003. Different explanantions for correlated errors in MTMM studies. *Structural Equation Modeling* , 10, 193–214.

SARIS W. E., AND I. N. GALLHOFER 2004. Operationalization of social science

concepts by intuition. *Quality and Quantity*, 38, 235–258.

SARIS W. E., W. VAN DER VELD, AND I. N. GALLHOFER 2004a. Development and improvement of questionnaires using predictions of reliability and validity. In S. Presser , J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken: Wiley, 275–299.

SARIS W. E., A. SATORRA, AND G. COENDERS 2004b. A new approach for evaluating quality of measurement instruments. *Sociological Methodology*, 3, 311–347.

W. E. SARIS, AND P. SNIDERMAN (eds.) 2004c. *The Issue of Belief: Essays in the Intersection of Nonattitudes and Attitude Change*. Princeton: University Press.

SARIS W. E., AND I. N. GALLHOFER 2006. The results of the MTMM experiments in round 2. Report for the ESS. London, ESS.

SARIS W. E., AND I. N. GALLHOFER 2007. Estimation of the effects of measurement characteristics on the quality of survey questions. Survey Research Methods, 1, 31–46.

SARIS W.E. *(forthcoming). Agree/disagree questions or forced choice questions?* Paper presented at a conference in in Preddvor, Slovenia.

SARIS W.E., AND J. KROSNICK (forthcoming) *Comparing questions with agree/disagree response options to questions with construct-specific response options*

SATORRA A. 1990. Robustness issues in structural equation modelling: a review of recent developments. *Quality and Quantity*, 24, 367–387.

SATORRA A. 1992. Asymptotic robust inferences in the analysis of mean and covariance structures. In P. V. Marsden (ed.), *Sociological Methodology 1992*. Oxford: Basil Blackwell, 249–278.

SATORRA, A. 1993. Asymptotic robust inferences in multi-sample analysis of augmented-moment matrices. In C. R. Rao, and C. M. Cuadras (eds.), Amsterdam: North Holland, 211–229.

SATORRA A. 2000. Goodness of fit testing of structural equation models with multiple group data and nonnormality. In R. Cudeck, S. du Toit, and D. Sörbom (eds.), *Structural Equation Modeling: Present and Future*, Lincolnwood: SSI, 231–257.

SCHERPENZEEL, A. C., AND W. E. SARIS 1993. The quality of indicators of satisfaction across Europe. A meta-analysis of multitrait-multimethod studies. *Bulletin de Methodologie Sociologique*, 39, 3–19.

SCHERPENZEEL A. C. 1995. *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies*. KPN Research: Leidschendam.

SCHERPENZEEL A. C., AND W. E. SARIS 1997. The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*. 25, 341–383.

SCHERPENZEEL A. C., AND W. E. SARIS 2006. Multitrait-Multimethod models for longitudinal research. In K.van Montford, H. Oud and A. Satorra (eds.), *Longitudinal Models in Behavioral and Related Sciences*, London: Lawrence Erlbaum, 381–403.

SCHOBER M. F., AND F. G. CONRAD 1997. Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly*, 60, 576–602.

SCHUMAN H., AND S. PRESSER 1981. *Questions and Answers in Attitude Survey: Experiments on Question Form, Wording and Context*. New York: Academic Press.

SCHWARZ N., AND H. J. HIPPLER 1987. What response scales may tell your respondents: Information functions of response alternatives. In H. J. Hippler, N. Schwarz, and S. Sudman (eds.), *Social Information Processing and Survey Methodology*, New York: Springer, 163–178.

SCHWARZ N., AND S. SUDMAN (eds.), 1996. *Answering Questions: Methodology for Determining Cognitive and Communica-*

tive Processes in Survey Research. San Francisco: Jossey-Bass.

SCHWARTZ S. H. 1997. Values and culture. In D. Muno, S. Carr, and J. Schumaker (eds.), *Motivation and Culture*, New York: Routledge, 69–84.

SCHWARTZ, S. H., AND A. BARDI 2001. Value hierarchies across cultures: Taking similarities perspective. *Journal of Cross Cultural Psychology*, 32, 268–290.

SILBERSTEIN A. S., AND S. SCOTT 1991. Expenditure diary surveys and their associated errors. In. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*. New York: Wiley, 303–327.

SKINNER B. F. 1953. *Science and Human Behavior*. New York: Macmillan.

SMITH T. W. 1987. The art of asking questions 1936-1985. *Public Opinion Quarterly*, 51, 95–108.

SNIDERMAN P. M., AND PH. E. TETLOCK 1986. Reflections on American racism. *Journal of Social Issues*, 42, 173–187.

SNIDERMAN P. M., R. A. BRODY AND P. E. TETLOCK 1991. *Reasoning and Choice. Explorations in Political Psychology*. Cambridge MA: Cambridge University Press.

SNIDERMAN P. M., AND S. THERIAULT 2004. The structure of political argument and the logic of issue framing. In W. E. Saris, and P. M. Sniderman (eds.), Studies in Public Opinion: Gauging Attitudes, Nonattitudes, Measurement Error and Change, Princeton: Princeton University Press, 133–166.

SNIJKERS G. 2002. *Cognitive Laboratory Experiences: On Pretesting, Computerized Questionnaires and Data Quality*. PhD thesis, University of Utrecht.

SOROKIN P. 1928. *Contemporary Sociological Theories*. New York: Harper.

STEENKAMP J., AND H. BAUMGARTNER 1998. Assessing measurement invariance in Cross-national consumer research. *Journal of Consumer Research*, 25, 78–90.

STEVENS S.S. 1975. *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects*. New York: Wiley.

STOKES D. E. 1963. Spatial models of party competition. *American Political Science Review*, 57, 368–377.

STOOP I. 2005. *The Hunt for the Last Respondent: Nonresponse in Sample Surveys*. The Hague: SCP.

SUDMAN S., AND N. M. BRADBURN 1983. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey Bass.

SUDMAN S., AND N. M. BRADBURN, AND N. SCHWARZ 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

SWAN M. 1995. *Practical English Usage*. Oxford: Oxford University Press.

TESSER A., AND L. MARTIN 1996. The psychology of evaluation. In E. T. Higgins, and A. W. Kruglinski (eds.), *Social Psychology. Handbook of Basic Principles*, New York: Guilford Press, 400–432.

TÖNNIES F. 1887. Gemeinschaft in der Gesellschaft. Grundbegriffe der reinen Sociologie. Berlin: Curtius.

TORCAL LORIENTE M. L., M. DIEZ DE ULZURRUN, AND S. PEREZ-NIEVAS MONTIEL (eds.), 2005. España; Sociedad y Politica en Perspectiva Comparada. Valencia: Tirant Lo Blanch.

TORGERSON W. S. 1958. *Theory and Methods of Scaling*. New York: Wiley.

TORTORA R. D. 1985. Cati in an agricultural statistical agency. *Journal of Official Statistics* 1, 301–314.

TOURANGEAU R., AND T. W. SMITH 1996. Asking sensitive questions: The impact of data collection mode, question format and question context. *Public Opinion Quarterly*, 60, 275–304.

TOURANGEAU R., K. RASINSKI, J. B. JOBE, T. W. SMITH, AND W. PRATT 1997. Sources of error in a survey of sexual behavior. Journal of Official Statistics, 13, 342–365.

TOURANGEAU R., L. J. RIPS, AND K. RASINSKI 2000. *The Psychology of Survey Response*. Cambridge MA: Cambridge University Press.

TRABASSO T., H. ROLLINS, AND E. SHAUGHNESSEY 1971. Storage and verification stages is processing concepts. *Cognitive Psychology*, 2, 239–289.

TROLDAHL V. C., AND R. E. CARTER 1964. Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71–76.

TURNER C. F., L. KU, S. M. ROGERS, L. D. LINDBERG, J. H. PLECK, AND F. L. SONENSTEIN, 1998. Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, 280, 867–873.

VAN MEURS A., AND W. E. SARIS 1990. Memory effects in MTMM studies. In W. E. Saris and A. van Meurs (eds.), *Evaluations of Measurement Instruments by Metaanalysis of Multitrait-Multimethod Studies*, Amsterdam: North Holland, 134–146.

VAN DER PLIGT J., AND N. K. DE VRIES 1995. *Opinies en Attitudes. Meting, Modellen en Theorie*. Amsterdam-Meppel: Boom.

VAN DER VELD W., W. E. SARIS, AND I.N. GALLHOFER 2000. *Survey Quality Prediction: SQP*. Paper presented at the ISA Methodology Conference in Cologne, Germany.

VAN DER VELD W., AND W.E. SARIS 2003. Separation of error , method effects, instability and attitude strength. In W. E. Saris, and P. Sniderman (eds.), *The Issue of Belief: Essays in the Intersection of Nonattitudes and Attitude Change*. Princeton: University Press, 37–63.

VAN DER VELD W. 2006. *The Survey Response Dissected: A New Theory about the Survey Response Process. PhD thesis*, University of Amsterdam.

VAN DER ZOUWEN J., W. DIJKSTRA, AND J. H. SMIT 1991. Studying respondent-interviewer interaction: The relationship between interviewing style, interviewer behavior and response behavior. In P. Biemer, R. M. Groves, L. E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: Wiley, 419–437.

VAN DER ZOUWEN J., AND W. DIJKSTRA 1996. Trivial and non-trivial question-answer sequences, types determinants and effects on data quality. *In Proceedings of the International Conference on Survey Measurement and Process Quality*, Bristol April1-4, 1995, American Statistical Association (ASA): Alexandria, 81–86.

VAN DER ZOUWEN J. 2000. An assessment of the difficulty of questions used in the ISSP questionnaires, the clarity of their wording and the comparability of the responses. *ZA-information* 45, 96–114.

VAN SCHUUR W. H. 1988. Stochastic unfolding. In W. E. Saris, and I. Gallhofer (eds.), *Sociometric Research, Vol. I, Data collection and scaling*, London: McMillan Press, 137–159.

VETTER A. 1997. Political Efficacy: Alte und neue Meßmodelle im Vergleich. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 49, 53–73.

VON WINTERFELDT D., AND W. EDWARDS 1986. *Decision Analysis and Behavioral Research*. Cambridge MA: Cambridge University Press.

VOOGT R. 2003. *I am not interested. Nonresponse bias, response bias and stimulus effects in election research. PhD thesis*, University of Amsterdam.

VOOGT R., AND W. E. SARIS 2003. To participate or not participate: The link between survey participation, electoral participation and political interest. *Political Analysis*, 11, 164–179.

WEBER E. G. 1993. *Varieties of Questions in English Conversation*. Amsterdam: J. Benjamin.

WEGENER B. (ed.), 1982. *Social Attitudes and Psychophysical Measurement of Opinions*. Hillsdale: Erlbaum.

WERTS C. E., AND R. L. LINN 1970. Path analysis. Psychological examples. *Psychological Bulletin*, 74, 193–212.

WILEY D. E., AND J. A. WILEY 1970. The

estimation of measurement error in panel data. *American Sociological Review*, 35, 112–117.

WILSON T. D., AND D. DUNN 1986. Effects of introspection on attitude-behavior consistency: Analyzing resaons versus focusing on feelings. *Journal of Experimental Social Psychology*, 22, 249–263.

WOTHKE W. 1996. Models for multitrait-multimethod matrix analysis. In G. C. Marcoulides, and R. E. Schumacker (eds.), *Advanced Structural Equation Modeling: Issues and Techniques*, Mahwah NJ: L. Erlbaum, 7–56.

WOUTERS M. 2001. *A design for the evaluation of the quality of open-ended questions*. Paper presented at the IRMCS meeting in Gent 25–27 May.

YULE G. 1998. *Explaining English Grammar*. Oxford: Oxford University Press.

ZALLER J. R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

ZANNA M. P., AND J. K. REMPEL 1988. Attitudes: A new look at an old concept. In D. Bar-Tal, and A. Kruglanski (eds.), *The Social Psychology of Knowledge*, Cambridge UK: Cambridge University Press, 210–245.

Subject Index

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *Robert M. Groves, Graham Kalton, J. N. K. Rao, Norbert Schwarz, Christopher Skinner*

The ***Wiley Series in Survey Methodology*** covers topics of current research and practical interests in survey methodology and sampling. While the emphasis is on application, theoretical discussion is encouraged when it supports a broader understanding of the subject matter.

The authors are leading academics and researchers in survey methodology and sampling. The readership includes professionals in, and students of, the fields of applied statistics, biostatistics, public policy, and government and corporate enterprises.

ALWIN · Margins of Error: A Study of Reliability in Survey Measurement
*BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
BIEMER and LYBERG · Introduction to Survey Quality
BRADBURN, SUDMAN, and WANSINK ·Asking Questions: The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social Health Questionnaires, *Revised Edition*
BRAVERMAN and SLATER · Advances in Survey Research: New Directions for Evaluation, No. 70
CHAMBERS and SKINNER (editors · Analysis of Survey Data
COCHRAN · Sampling Techniques, *Third Edition*
COUPER, BAKER, BETHLEHEM, CLARK, MARTIN, NICHOLLS, and O'REILLY (editors) · Computer Assisted Survey Information Collection
COX, BINDER, CHINNAPPA, CHRISTIANSON, COLLEDGE, and KOTT (editors) · Business Survey Methods
*DEMING · Sample Design in Business Research
DILLMAN · Mail and Internet Surveys: The Tailored Design Method
GROVES and COUPER · Nonresponse in Household Interview Surveys
GROVES · Survey Errors and Survey Costs
GROVES, DILLMAN, ELTINGE, and LITTLE · Survey Nonresponse
GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, and WAKSBERG · Telephone Survey Methodology
GROVES, FOWLER, COUPER, LEPKOWSKI, SINGER, and TOURANGEAU · Survey Methodology
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume 1: Methods and Applications
*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume II: Theory
HARKNESS, VAN DE VIJVER, and MOHLER · Cross-Cultural Survey Methods
KALTON and HEERINGA · Leslie Kish Selected Papers
KISH · Statistical Design for Research
*KISH · Survey Sampling
KORN and GRAUBARD · Analysis of Health Surveys
LESSLER and KALSBEEK · Nonsampling Error in Surveys
LEVY and LEMESHOW · Sampling of Populations: Methods and Applications, *Third Edition*
LYBERG, BIEMER, COLLINS, de LEEUW, DIPPO, SCHWARZ, TREWIN (editors) · Survey Measurement and Process Quality

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MAYNARD, HOUTKOOP-STEENSTRA, SCHAEFFER, VAN DER ZOUWEN · Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview

PORTER (editor) · Overcoming Survey Research Problems: New Directions for Institutional Research, No. 121

PRESSER, ROTHGEB, COUPER, LESSLER, MARTIN, MARTIN, and SINGER (editors) · Methods for Testing and Evaluating Survey Questionnaires

RAO · Small Area Estimation

REA and PARKER · Designing and Conducting Survey Research: A Comprehensive Guide, *Third Edition*

SARIS and GALLHOFER · Design, Evaluation, and Analysis of Questionnaires for Survey Research

SÄRNDAL and LUNDSTRÖM · Estimation in Surveys with Nonresponse

SCHWARZ and SUDMAN (editors) · Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research

SIRKEN, HERRMANN, SCHECHTER, SCHWARZ, TANUR, and TOURANGEAU (editors) · Cognition and Survey Research

SUDMAN, BRADBURN, and SCHWARZ · Thinking about Answers: The Application of Cognitive Processes to Survey Methodology

UMBACH (editor) · Survey Research Emerging Issues: New Directions for Institutional Research No. 127

VALLIANT, DORFMAN, and ROYALL · Finite Population Sampling and Inference: A Prediction Approach