SURVEYS
Survey research in operations management: a process-based perspective

Cipriano Forza
Università di Padova, Vincenca, Italy

Keywords *Operations management, Methodology, Surveys, Research, Empirical study, Quality*

Abstract *This paper provides guidelines for the design and execution of survey research in operations management (OM). The specific requirements of survey research aimed at gathering and analysing data for theory testing are contrasted with other types of survey research. The focus is motivated by the need to tackle the various issues which arise in the process of survey research. The paper does not intend to be exhaustive: its aim is to guide the researcher, presenting a systematic picture which synthesises suitable survey practices for research in an OM context. The fundamental aim is to contribute to an increase in the quality of OM research and, as a consequence, to the status of the OM discipline among the scientific community.*

Introduction

If we compare contemporary research in operations management (OM) with that conducted in the early 1980s, we notice an increase in the use of empirical data (derived from field observation and taken from industry) to supplement mathematics, modelling, and simulation to develop and test theories. Many authors have called for this empirical research, since OM became a functional field of study (such as marketing, management information systems, etc.) within the management discipline (Meredith *et al.*, 1989; Flynn *et al.*, 1990; Filippini, 1997; Scudder and Hill, 1998). The rationale was to reduce the gap between management theory and practice, to increase the usefulness of OM research to practitioners and, more recently, to increase the scientific recognition of the OM field. Recognition of the value of empirical research in OM led to an increase in both the number and the percentage of studies based on empirical research and, especially, on survey research (Meredith, 1998; Amoako-Gyampah and Meredith, 1989; Scudder and Hill, 1998; Pannirselvam *et al.*, 1999; Rungtusanatham *et al.*, 2001). The number of survey research based articles increased steadily from the mid-1980s to the early 1990s, and increased sharply from 1993. By 1996, empirical research based articles accounted for approximately 30 per cent of the research published in the main OM outlets, and survey-based articles accounted for 60 per cent of this empirical subset. Furthermore, survey research was being used (sometimes in combination with other methods) to investigate phenomena in very different OM sub-fields (see Table I for details).

International Journal of Operations & Production Management,
Vol. 22 No. 2, 2002, pp. 152-194.

DOI 10.1108/01443570210414310

The author acknowledges Professors Paul Coughlan, Christer Karlsson, Fortunato Pesarin, Manus Rungtusanatham and Chris Voss, Dr Fabrizio Salvador and the anoymous reviewers for their suggestions durings the development of this paper.

	Survey	Modelling and survey	Theoretical conceptual and survey	Case study and survey	Simulation and survey	Total survey	Total topic	% survey
Strategy	77	3	6	2		88	213	41
Quality	51	2	5			58	222	26
Process design	33	3	2			38	221	17
Inventory control	16		1	1		18	317	6
Purchasing	15					15	39	38
Scheduling	13	1				14	500	3
Services	11	1		1		13	53	25
Distribution	7					7	61	11
Facility layout	2	3			1	6	149	4
Project management	3					3	34	9
Aggregate planning	3					3	13	23
Work measurement	3					3	10	30
Quality work life	3					3	4	75
Maintenance	2					2	40	5
Facility location		1				1	21	5
Forecasting	1					1	20	5
Capacity planning						0	41	0
Count total	240	14	14	4	1	273	1,958	14
Article total	206	11	10	3	1	231	1,754	13
Double count number	34	3	4	1	0	42	204	21

Note: Journals considered: *JOM, MS, IIE, DS, IJPR, IJOPM, POM*. Period considered 1992-1997

Source: Adapted from Pannirselvam *et al.* (1999)

Table I.
Survey research in OM sub-fields

In recent years "... remarkable progress has been demonstrated ... by the quality and the sophistication of the research endeavours ..." based on survey research (Rungtusanatham, 1998). Evidence of these improvements is to be found, for example, in Flynn *et al.* (1990) and, later, in a 1997 special issue of *IJOPM* (edited by Filippini and Voss) which included several applications of survey research in OM (Van Donselaar and Sharman, 1997; Collins and Cordon, 1997; Flynn *et al.*, 1997; Whybark, 1997).

There have been many calls for improved quality and more appropriate use of survey research in OM (Forza and Di Nuzzo, 1998; Malhotra and Grover, 1998; Hensley, 1999; Rungtusanatham *et al.*, 2001). These calls resonate throughout the OM research community. For example, Forza and Vinelli (1998) gathered the opinions and perceptions of 89 OM scholars and reported that there was:

- a need for greater clarity and explicitness in reporting information on the survey execution (these are basic requirements if critical use of results, comparison and replicability are to be possible);
- a clear demand for unambiguous, reliable methods in all phases of research;
- a need for common terminology in the field concerning the meaning of variables and their operationalisation;
- a need for the use of scientific (i.e. reliable and valid) measurement;
- a need for more careful sample selection and description;
- the need for an explicit, clear and strong theoretical background;
- a necessity for far greater discussion of the results in terms of generalisation.

A key objective of this paper is to provide suggestions to reduce the above shortcomings. In pursuing this objective, it focuses on theory testing survey research in the first section. Here, there is no intention to downplay the other types of survey as the penultimate section will highlight the main differences between theory testing and other types of survey research. However, the intention is to focus on the most demanding type of survey research in order to increase awareness both of possible shortcomings and also of useful preventative actions that can be taken. The paper, therefore, should help OM researchers, especially those engaging in survey research for the first time, with an overview of the survey research process. The paper is structured as follows:

(1) the first section provides insights into what survey research is and when it can be used;

(2) the following six sections provide a road map for conducting survey research;

(3) the final section provides some properties of well-conducted survey research.

What is survey research and when can it be used?

In OM, as in other fields of business, research can be undertaken to solve an existing problem in the work setting. This paper focuses on survey research conducted for a different reason – to contribute to the general body of knowledge in a particular area of interest. In general, a survey involves the collection of information from individuals (through mailed questionnaires, telephone calls, personal interview, etc.) about themselves or about the social units to which they belong (Rossi *et al.*, 1983). The survey sampling process determines information about large populations with a known level of accuracy (Rea and Parker, 1992).

Survey research, like the other types of field study, can contribute to the advance of scientific knowledge in different ways (Babbie, 1990; Kerlinger, 1986). Accordingly, researchers often distinguish between exploratory, confirmatory (theory testing) and descriptive survey research (Pinsonneault and Kraemer, 1993; Filippini, 1997; Malhotra and Grover, 1998):

- *Exploratory survey research* takes place during the early stages of research into a phenomenon, when the objective is to gain preliminary insight on a topic, and provides the basis for more in-depth survey. Usually there is no model, and concepts of interest need to be better understood and measured. In the preliminary stages, exploratory survey research can help to determine the concepts to be measured in relation to the phenomenon of interest, how best to measure them, and how to discover new facets of the phenomenon under study. Subsequently, it can help to uncover or provide preliminary evidence of association among concepts. Later again, it can help to explore the valid boundary of a theory. Sometimes this kind of survey is carried out using data collected in previous studies.
- *Confirmatory (or theory testing or explanatory) survey research* takes place when knowledge of a phenomenon has been articulated in a theoretical form using well-defined concepts, models and propositions. In this case, data collection is carried out with the specific aim of testing the adequacy of the concepts developed in relation to the phenomenon, of hypothesised linkages among the concepts, and of the validity boundary of the models. Correspondingly, all of the error sources have to be considered carefully.
- *Descriptive survey research* is aimed at understanding the relevance of a certain phenomenon and describing the distribution of the phenomenon in a population. Its primary aim is not theory development, even though through the facts described it can provide useful hints both for theory building and for theory refinement (Dubin, 1978; Malhotra and Grover, 1998; Wacker, 1998).

Some established OM sub-fields (such as manufacturing strategy and quality management) have been researched extensively, in part through survey

research, and the corresponding bodies of knowledge developed enough to allow researchers to embrace theory testing survey research (Handfield and Melnyk, 1998). In contrast, some emerging areas, such as e-commerce in operations, have scarcely been researched at all and require exploratory research. Finally, many issues, interesting both for practitioners and for academics – such as the level of adoption of software for statistical process control – can be researched through descriptive survey.

What is needed prior to survey research design?

Theory testing survey research is a long process which presupposes the pre-existence of a theoretical model (or a conceptual framework). It includes a number of related sub-processes: the process of translating the theoretical domain into the empirical domain; the design and pilot testing processes; the process of collecting data for theory testing; the data analysis process; and the process of interpreting the results and writing the report. This theory testing survey research process is illustrated in Figure 1.

The theoretical model

Before starting theory testing survey research, the researcher has to establish the conceptual model (Dubin, 1978; Sekaran, 1992; Wacker, 1998) by providing:

- *Construct names and nominal definitions*: clear identification, labels and definitions of all the constructs (i.e. "theoretical concepts" or, in a somewhat looser language, "variables") considered relevant.
- *Propositions*: presentation and discussion of the role of the constructs (independent, dependent, intervening, moderating), the important linkages between them, and an indication of the nature and direction of the relationships (especially if available from previous findings).
- *Explanation*: a clear explanation of why the researcher would expect to observe these relationships and, eventually, linkages with other theories (within or outside OM (Amundson, 1998)).
- *Boundary conditions*: definition of conditions under which the researcher might expect these relationships to hold; it includes the identification of the level of reference of the constructs and their statements of relationships (i.e. – where the researcher might expect the phenomenon to exist and manifest itself – individual, group, function, or organisation).

Very often the theoretical framework is depicted through a schematic diagram. While not a requirement, it may be useful to facilitate communication.

The researcher can find useful support for this task in methodological books in the social sciences (such as Dubin, 1978; Kerlinger, 1986; Emory and Cooper, 1991; Miller, 1991; Sekaran, 1992) or in OM (Anderson *et al.*, 1994; Flynn *et al.*, 1994), and in methodological articles in OM (Meredith, 1998; Wacker, 1998). By definition, survey research is not theory-testing survey research if, from the

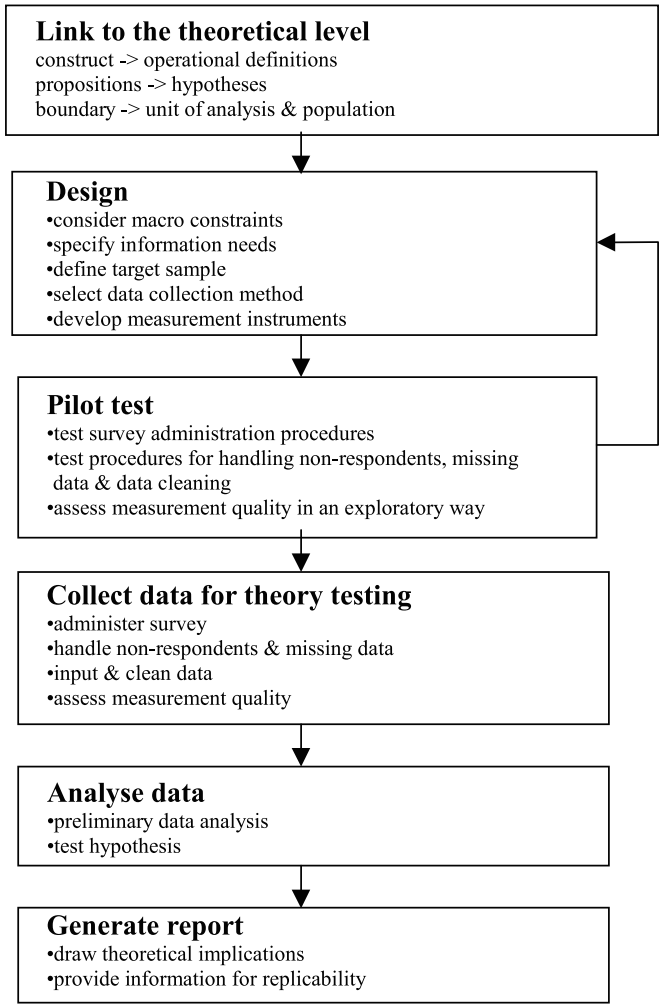


Figure 1.
The theory-testing survey research process

outset, it is not based on a theoretical model. Unfortunately, for many OM topics formal theory is underdeveloped. For many years OM has developed implicit theories but the lack of explicitness has prevented the testing of these theories. As a consequence, before embarking on theory-testing survey research, the OM researcher is often obliged to develop a theoretical framework. This development activity itself can be publishable (as for example Anderson *et al.*, 1994; Flynn *et al.*, 1994; Forza, 1995).

From the theoretical model to hypotheses

Once the constructs, their relationships and their boundary conditions have been articulated, then the propositions that specify the relationships among the constructs have to be translated into hypotheses, relating empirical indicators.

For example, the researcher might propose the following: "the adoption of TQM in organisations would have positive effects on organisational performance". Such a statement is at the conceptual level. At the empirical level (i.e. at the level of hypotheses), the following hypothesis might be tested: "ROI would be positively correlated with the degree of TQM adoption". In this hypothesis the "degree of TQM adoption" is an empirical and numerically based measure of how extensive is the adoption of TQM or how committed the organisation is to TQM.

In other words, before the researcher can talk about how to collect data it is necessary to:

- define the unit of analysis corresponding to the level of reference of the theory;
- provide and test the operational definitions for the various constructs; and
- translate the propositions into hypotheses.

Defining the unit of analysis. The empirical parallel of the level of reference of the theory is the "unit of analysis" issue. The unit of analysis refers to the level of data aggregation during subsequent analysis. The unit of analysis in OM studies may be individuals, dyads, groups, plants, divisions, companies, projects, systems, etc. (Flynn *et al.*, 1990).

It is necessary to determine the unit of analysis when formulating the research questions. Data collection methods, sample size and even the operationalization of constructs may sometimes be determined or guided by the level at which data will be aggregated at the time of analysis (Sekaran, 1992). Not having done so in advance may mean that later analyses, appropriate for the study, cannot be performed.

When the level of reference is different from the unit of analysis the researcher will encounter the cross-level inference problem, i.e. collecting data at one level and interpreting the result at a different level (Dansereau and Markham, 1997). If data are collected, or analysed, at group level (for example at plant level) and conclusions are drawn at individual level (for example at employee level), the researcher will encounter the ecological fallacy problem (Robinson, 1950; Babbie, 1990). The issue of cross-level inference becomes more important when more than one unit of analysis is involved in a study (Babbie, 1990). Discussion of methodological problems associated with the level of analysis (plant, SBU, company) can be found in Boyer and Pagell (2000), with reference to operations strategy and advanced manufacturing technology, and in O'Leary-Kelly and Vokurka (2000), with reference to manufacturing flexibility.

Develop and test the operational definitions. This section focuses mainly on the "what" part of an operational definition (the list of observable elements) while leaving the "how" part (exact questions, etc.) to the section "measurement instrument":

(1) *Develop the operational definitions*. The first problem that the researcher faces is in transforming the theoretical concepts into observable and measurable elements. If the theoretical concept is multidimensional, then all of its dimensions have to find corresponding elements in the operational definition. For example, the construct "learning" can be decomposed in its three dimensions (understanding, retention, application) and each dimension can be further decomposed in observable elements (Sekaran, 1992). The list of observable elements that constitutes the operational definition of learning are: answer questions correctly, give appropriate examples, recall material after some lapses of time, solve problems applying concepts understood and recalled, and integrate with other relevant material. Actually operational definitions of constructs "must specify both the [specific observable elements of a construct] and how they are to be observed" (Emory and Cooper, 1991).

This action of reducing abstract constructs so that they can be measured (i.e. construct operationalisation) presents several problems: alignment between the theoretical concepts and the empirical measures, the choice between objective and perceptual questions, or the selection of one or more questions for the same construct. These problems can be overcome by using operational definitions that have already been developed, used and tested. Unfortunately, the availability of such operational definitions is still very limited in OM, with some notable exceptions in sub-fields such as quality management (Handfield and Melnyk, 1998). Therefore the researcher is forced frequently to develop new measures: in this case works reporting previous experiences and providing suggestions on measure development may be useful (see for example Converse and Presser (1988), Hinkin (1995), Hensley (1999).

The translation from theoretical concepts to operational definitions can be very different from construct to construct. While some constructs lend themselves to objective and precise measurement, others are more nebulous and do not lend themselves to such precise measurement, especially when people's feelings, attitudes and perceptions are involved. When constructs, such as "customer satisfaction", have multiple facets or involve people's perceptions/feelings or are planned to be measured through people's perceptions it is highly recommended to use operational definitions which include multiple elements (Lazarsfeld, 1935; Payne, 1951; Malhotra and Grover, 1998; Hensley, 1999). When objective constructs are considered, a single direct question would be appropriate.

The process of identifying the elements to insert in the operational definition (as well as the items (questions) in the measure) may include both contacting those making up the population of interest to gain a practical knowledge of how the construct is viewed in actual organisations, and identifying important specifics of the industry being

studied. "The development of items using both academic and practical perspectives should help researchers develop good preliminary scales and keep questionnaire revision to a minimum" (Hensley, 1999).

(2) *Test the operational definitions for content validity*. When the operational definition has been developed, the researcher should test for content validity. The content validity of a construct measure can be defined as "the degree to which the measure spans the domain of the construct's theoretical definition" (Rungtusanatham, 1998). It is the extent to which the measure captures the different facets of a construct[1]. Evaluating face validity of a measure (i.e. the measure "on its face" seems like a good translation of the theoretical concept) can indirectly assess its content validity. Face validity is a matter of judgement and must be assessed before data collection (Rungtusanatham, 1998).

In addition to self-validating the measure – through an agreement on the content adequacy among the researchers who developed the measure – additional support should be sought from experts and/or the literature. While literature is important, it may not cover all aspects of the construct. Typically, OM researchers tend to overlook this issue but there are several approaches that can be used (Rungtusanatham, 1998). One approach used to quantify face validity involves a panel of subject-matter experts (SMEs) and the computation of Lawshe's (1975) content validity ratio for each candidate item in the measure (CVR_i). Mathematically, CVR_i is computed as follows (where n_e is the number of SMEs indicating the measurement item i as "essential", and N is the total number of SMEs in the panel):

$$CVR_i = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}.$$

Lawshe (1975) has further established minimum CVR_is for different panel sizes. For example, for a panel size of 25 the minimum CVR_i is 0.37).

Stating hypotheses. A hypothesis is a logically conjectured relationship between two or more variables (measures) expressed in the form of testable statements. A hypothesis can also test whether there are differences between two groups (or among several groups) with respect to any variable or variables. These relationships are conjectured on the basis of the network of associations established in the theoretical framework and formulated for the research study. Hypotheses can be set either in the propositional or the if-then statement form. If terms such as "positive", "negative", "more than", "less than" and "like" are used in stating the relationship between two variables or comparing two groups, these hypotheses are directional. When there is no indication of the direction of the difference or relationship they are called non-directional. Non-directional hypotheses can be formulated either when the relationships or

differences have never been previously explored, or when there are conflicting findings. It is better to indicate the direction when known.

The null hypothesis is a proposition that states a definitive, exact relationship between two variables. For example: the correlation between two variables is equal to zero; or, the difference between the means of two groups in the population is equal to zero.

In general the null statement is expressed as no (significant) relationship between two variables or no (significant) difference between two groups ... What we are implying through the null hypothesis is that any differences found between two sample groups (or any relationships found between two variables based on our sample) is simply due to random sampling fluctuations and not due to any "true" differences between the two population groups (or relationship between two variables). The null hypothesis is thus formulated so that it can be tested for possible rejection. If we reject the null hypothesis, then all permissible alternative hypotheses related to the tested relationship could be supported. It is the theory that allows us to trust the alternative hypothesis that is generated in the particular research investigation ... Having thus formulated the null H_0 and alternative H_a hypotheses, the appropriate statistical tests, which would indicate whether or not support has been found for the alternate, should be identified (Sekaran, 1992).

In formulating a hypothesis[2] on the linkage between two variables the OM researcher should be conscious of the form of relation being defined. For example, if the researcher hypothesises a correlation between two variables, a linear relationship is being assumed. However, if there is no subsequent evidence of a significant correlation between the two variables, the researcher cannot conclude that there is no association. It can only be stated that in the sample considered there is no evidence of a linear relationship between the variables. In sum, when stating the hypotheses, and later when choosing the appropriate test, the researcher should carefully think about the kind of linkage being assumed/tested.

How should a survey be designed?

Survey design includes all of the activities that precede data collection. In this stage the researcher should consider all of the possible shortcomings and difficulties and should find the right compromise between rigor and feasibility. Planning all of the future activities in a detailed way and defining documents to keep track of decisions made and activities completed are necessary to prevent subsequent problems.

Considering constraints and information needs at the macro level

Before embarking on a theory-testing survey, one should consider the suitability of the survey method and the overall feasibility of the research project. If a well-developed model is not available then the researcher should consider how much time and effort will be required to develop such a model. Time, costs and general resource requirements can constrain a survey project,

forcing a less expensive type of survey or, in the extreme, making it infeasible. Other possible constraints are the accessibility of the population and the feasibility of involving the right informants.

In survey research, there is a trade-off between time and cost constraints, on the one hand, and minimisation of four types of error, on the other hand:

(1) *Sampling error*. A sample with no (or unknown) capability of representing the population (because of inadequate sample selection or because of auto-selection effects) excludes the possibility of generalising the results beyond the original sample.

(2) *Measurement error*. Data derived from the use of measures which do not match the theoretical dimensions, or are not reliable, make any test meaningless.

(3) *Statistical conclusion error*. When performing statistical tests there is a probability of accepting a conclusion that the investigated relationship (or other effect) does not exist even when it does exist.

(4) *Internal validity error*. When the explanation given of what has been observed is less plausible than rival ones, then the conclusions can be considered erroneous.

While dissatisfaction with the above-mentioned constraints could halt the survey research, failure to minimise all of the above four errors "... can and will lead to erroneous conclusions and regression rather than progress in contribution to theory" (Malhotra and Grover, 1998).

To evaluate adequately the tightness of the constraints, the researcher should identify the main information needs (such as time horizon, information nature, etc.) which flow from the stated hypotheses and, ultimately, from the various purposes of the study. For example, if the study aims at a very rigorous investigation of causal relationships, or if the theoretical model implies some dynamics, longitudinal data may be required (i.e. data on the same unit at different points in time). Boyer and Pagell (2000) have called for such an extended time horizon when researching operations strategy research issues. Similarly, if the study requires information which is considered confidential in nature by the respondents, then the cost and time to get the information is probably high and a number of survey design alternatives are not viable. Finally, a study may aim not only to test a theory but also to perform additional exploratory analyses, while reducing the cost of the research and increasing the speed in generating knowledge. In this case, the problem is to satisfy questionnaire length constraints: classifing information items by priority can be of help later on in choosing what questions to eliminate (Alreck and Settle, 1985; Babbie, 1990).

Planning activities

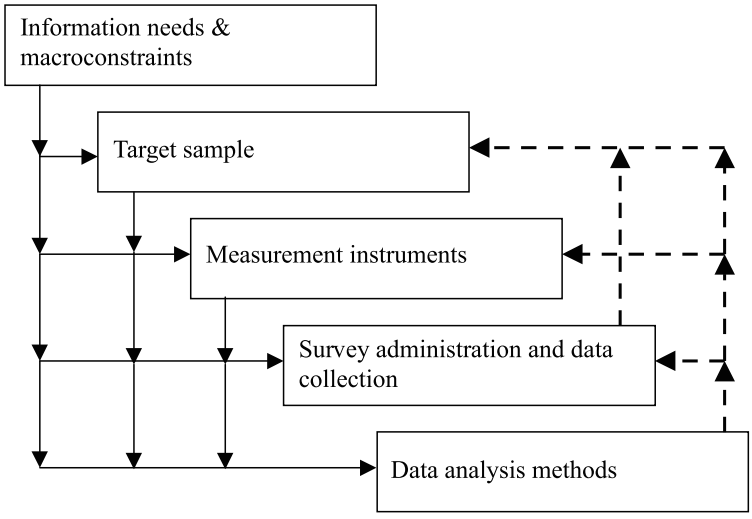
Theory-testing survey research is a process with a series of steps that are linked to each other (see Figure 1). Careful planning of this process is crucial to

prevent problems and to assure the quality of the research process. For this reason the design phase should be very detailed, and followed by a pilot testing phase aimed at assuring that the survey instrumentation and procedures are adequate.

However, in planning the activities the decisions made during the early steps affect the choices remaining at later steps (see Figure 2). It is not possible to proceed step by step: constraints and limitations in the later steps should be considered in the earlier steps. For these reasons, major decisions about data collection (telephone, interview and mail) and time horizon (cross-sectional or longitudinal) must always be made prior to designing and selecting a sample and constructing the questionnaire and the other material. It is important to match the capabilities and the limitations of the data-processing methods with the sampling and instrumentation. For more details on project planning see Alreck and Settle (1985).

The sample

Before discussing the sample we need to define the following terms: population, population element, population frame, sample, subject and sampling. Population refers to the entire group of people, firms, plants or things that the researcher wishes to investigate. An element is a single member of the population. The population frame is a listing of all the elements in the population from which the sample is to be drawn. A sample is a subset of the population: it comprises some members selected from the population. A subject is a single member of the sample. Finally, sampling is the process of selecting a sufficient number of elements from the population so that by studying the sample, and understanding the properties or the characteristics of the sample subjects, the researcher will be able to generalise the properties or characteristics to the



Source: Adapted from Alreck and Settle (1985)

Figure 2. Linkages between decisions in survey planning

population elements. Sampling overcomes the difficulties of collecting data from the entire population which can be impossible or prohibitive in terms of time, costs and other human resources.

Sample design is a step usually overlooked in OM surveys (Forza and Di Nuzzo, 1998; Rungtusanatham *et al.*, 2001). Many articles do not report adequately on how their sample was constructed, and do not provide sufficient information on the resulting sample. The majority of survey-based OM articles (approximately 88 per cent) do not rely on a probabilistic sampling approach (Rungtusanatham *et al.*, 2001). Poor sample design can constrain the application of more appropriate statistical techniques and the generalisability of the results. Two issues should be addressed: randomness and sample size. Randomness is associated with the ability of the sample to represent the population of interest. Sample size is associated with the requirements of the statistical procedures used for measurement quality assessment and hypothesis testing.

Population frame. The population frame should be drawn from widely available sources to facilitate the replicability of studies. The industry classification (usually specified through SIC codes) is an important aspect of framing the population. "SIC codes can provide a useful starting point, however their classifications may need to be modified, as appropriate to the needs of the POM researcher" since SIC codes "were not designed for POM research . . . for example process technology can vary considerably between two related SIC codes (e.g. computers are classified with machinery)" (Flynn *et al.*, 1990). To facilitate control of industry effects, a good practice is to consider four-digit SIC codes when building the frame and later on the research sample. "Controlling industry effects can compensate for variability between industries, in terms of processes, work force management, competitive forces, degree of unionisation, etc." (Flynn *et al.*, 1990).

There are other justifiable ways of choosing a sample, based on the specific aspects (for example common process technology, position in the supply chain, etc.) which should be controlled for the investigation of the phenomenon under study. For example, Dun's databases (e.g. "Dun's guide: the metalworking directory" in the USA, or "Duns's 25.000" in Italy) are useful sources since they provide such information (in some countries at plant level) as products made, number of employees, addresses, etc. (see http://www.dundb.co.il/). Other than industry, another important variable to be controlled is company size: number of employees and sales are easily available information which can be incorporated in the sample selection process.

Sample design. There are several sample designs, which can be grouped into two families: probabilistic and non-probabilistic sampling. In probabilistic sampling the population elements have some known probability of being selected, differently than non-probabilistic sampling. Probabilistic sampling is used to assure the representativeness of the sample when the researcher is interested in generalising the results. When time or other factors prevail on generalisability considerations then non-probabilistic sampling is usually

chosen. Table II shows some basic types of sampling approaches (for more details see Babbie (1990).

Stratified random sampling is a very useful type of sampling since it provides more information for a given sample size. Stratified random sampling involves the division of the population into strata and a random selection of subjects from each stratum. Strata are identified on the basis of meaningful criteria like industry type, size, performance, or others. This procedure ensures high homogeneity within each stratum and heterogeneity between strata. Stratified random sampling allows the comparison of population subgroups and allows control for factors like industry or size which very often affect results.

Sample size. Sample size is the second concern. It is a complex issue which is linked to the significance level and the statistical power of the test, and also to the size of the researched relationship (for example association strength or amount of difference).

When making statistical inference, the researcher can make either a Type I error (reject the null hypothesis H_0 when it is true) or a Type II error (H_0 is not rejected when the alternative hypothesis H_a is true). The probability of making a Type I error (α) is called significance level. Typically in most social sciences (OM included) α is taken to 0.05, however in several cases $\alpha = 0.01$ and $\alpha = 0.001$ are used. The null hypothesis is rejected if the observed significance level (p-value) is less than the chosen value of α (McClave and Benson, 1991). The probability of a Type II error is β, and the statistical power is equal to 1-β. A high statistical power is required to reduce the probability of failing to detect

Representativeness	Purpose is mainly	Type of sampling
Essential for the study = > probabilistic sampling	Generalisability	Simple random sampling. Systematic sampling
	Assessing differential parameters in subgroups of population	Proportionate stratified random sampling (for subgroups with an equal number of elements)
		Disproportionate stratified random sampling (for subgroups with a different number of elements)
	Collecting information in localised areas	Area sampling
	Gathering information from a subset of the sample	Double (or multistage) sampling
Not essential for the study = > non-probabilistic sampling	Obtain quick, even if unreliable, information	Convenience sampling
	Obtain information relevant to and available only from certain groups	Judgement sampling (when looking for information that only a few experts can provide)
		Quota sampling (when the responses of special interest minority groups are needed)

Table II.
Sampling approaches

an effect when it is present. A balance between the two types of errors is needed because reducing any one type of error raises the likelihood of increasing the probability of the other type of error. Low power leads to a study which is not able to detect large size effects, while high power leads to committing unnecessary resources only in order to be able to detect trivial effects. Methodologists are only now beginning to agree that a power of about 0.8 represents a reasonable and realistic value for research in social/behavioural sciences (Verma and Goodale, 1995). This means that only 20 per cent of the repeated studies will not yield a significant result, even when the phenomenon exists.

Even though the power of a statistical test depends on three factors (α, effect size and sample size), from a practical point of view only the sample size is used to control the power. This is because the α level is effectively fixed at 0.05 (or some other value) and the effect size (for example the size of the difference in the means between two samples or the correlation between two variables) can also be assumed to be fixed at some unknown value (the researcher may wish not to change the effect but only detect it). The required sample sizes, with desired statistical powers of 0.8 and 0.6, are shown in Table III as a function of effect size (and significance levels). One can see that the required sample size increases while increasing the statistical power, and/or decreasing the significance level, and/or decreasing the size of the effect researched. Verma and Goodale (1995) provide more detail (and selected bibliography) on this issue. They also provide some figures of the statistical power evident in OM articles published in *JOM* and *DS* in the period 1990-1995.

Data collection method

Data can be collected in a variety of ways, in different settings, and from different sources. In survey research, the main methods used to collect data are interviews and questionnaires. Interviews may be structured or unstructured. They can be conducted either face to face or over the telephone. Questionnaires can be administered personally, by telephone or mailed to the respondents. The researcher can also use the telephone to improve the response rate of mail surveys by making prior notification calls.

Each data collection method has merits as well shortcomings. Decisions on which method is best cannot be made in the abstract; rather, they must be based on the needs of the specific survey as well as time, cost and resource constraints.

Table III.
Effect size and statistical power and sample size

	Stat. power = 0.6		Stat. power = 0.8	
	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Large effect (e.g. strong association)	12	18	17	24
Medium effect (e.g. medium association)	30	45	44	62
Small effect (e.g. small association)	179	274	271	385

In a mail survey, questionnaires are printed and sent by mail. The respondents are asked to complete the questionnaire on their own and to send it back. Mailed questionnaires have the following advantages: cost savings; they can be completed at the respondent's convenience; there are no time constraints; they can be prepared to give an authoritative impression; they can ensure anonymity; and they can reduce interviewer bias. On the other hand, mailed questionnaires have a lower response rate than other methods, involve longer time periods, and are more affected by self-selection, lack of interviewer involvement and lack of open-ended questions.

In a face-to-face survey, the interviewer solicits information directly from a respondent in personal interviews. The advantages are: flexibility in sequencing the questions, details and explanation; an opportunity to administer highly complex questionnaires; improved ability to contact hard-to-reach populations; higher response rates; and increased confidence that data collection instructions are followed. There are some disadvantages, including: higher cost; interviewer bias; the respondent's reluctance to co-operate; greater stress for both respondents and interviewer; and less anonymity.

Telephone surveys involve collecting information through the use of telephone interviews. The advantages are: rapid data collection; lower cost; anonymity; large-scale accessibility; and ensuring that instructions are followed. The disadvantages are: less control over the interview situation; less credibility; and lack of visual materials.

Table IV summarises the relative strengths of the different methods. Here, "1" indicates that the method that has the maximum strength, and "3" the minimum, in the factor noted. Dillman (1978, pp. 74-6) and Rea and Parker (1992) provide a more detailed comparison.

Recently a new way to approach companies and administer questionnaires has appeared. The researcher can send a questionnaire through e-mail or ask respondents to visit a Web site where the questionnaire can be filled in and returned electronically. One advantage is the minimal cost compared with other means of distribution (Pitkow and Recker, 1995). However, potential problems lie in sampling and controlling of the research environment (Birnbaum, 1999).

Factors influencing coverage and secured information	Mailed questionnaires	Personal interview	Telephone survey
Lowest relative cost	1	3	2
Highest response rate	3	1	2
Highest accuracy of information	2	1	3
Largest sample coverage	3	1	3
Completeness, including sensitive materials	3	1	2
Overall reliability and validity	2	1	3
Time required to secure information	3	2	1
Ease of securing information	1	3	2

Source: Adapted from Miller (1991, p. 168)

Table IV.
Comparison of data collection methods

The measurement instrument

One of the main characteristics of the survey is that it relies on structured instruments to collect information. Once the researcher has decided on the content of a measure (the specific empirical aspects that have to be observed), several tasks remain to develop the measurement instruments, namely:

- define the way questions are asked to collect the information on a specific concept (see subsection "wording");
- for each question decide the scale on which the answers are placed (see subsection "scaling");
- identify the appropriate respondent(s) to each question (see subsection "respondent identification");
- put together the questions in questionnaires that facilitate and motivate the respondent(s) to respond (see subsection "rules of questionnaire design").

The main issues related to each task are discussed in the following subsections. It should be noted, however, that the actual design of the survey questionnaire depends on whether the questionnaire is to be administered by telephone interview, on site through interview, on site using pen and paper, or by mail using pen and paper.

Wording. In formulating the questions the researcher should ensure that the language of the questionnaire is consistent with the respondent's level of understanding. If a question is not understood or interpreted differently by respondents, the researcher will get unreliable responses to the question, and these responses will be biased. The researcher also has to choose between an open-ended (allowing respondents to answer in any way they choose) or closed question (limiting respondents to a choice among alternatives given by the researcher). Closed questions facilitate quick decisions and easy information coding, but the researcher has to ensure that the alternatives are mutually exclusive and collectively exhaustive. Another choice in formulating the questions is the mix of positively and negatively worded questions in order to minimise the tendency in respondents to mechanically circle the points toward one end of the scale.

The researcher should replace double-barrelled questions (i.e. questions that have different answers to its subparts) with several separate questions. Ambiguity in questions should be eliminated as much as possible. Leading questions (i.e. questions phrased in a way that lead the respondent to give responses that the researcher would like to, or may come across as wanting to, elicit) should be avoided as well. In the same way loaded questions (i.e. questions phrased in an emotionally charged manner) should be eliminated. Questions should not be worded to elicit socially desirable responses. Finally, a question or a statement should not exceed 20 words of full line in print. For further details on wording, see for example Horst (1968), Converse and Presser (1986) and Oppenheim (1992).

Scaling. A second task in developing the measurement instrument concerns the scale to be used to measure the answers. The scale choice depends on the ease with which both the respondent can answer and the subsequent analyses will be done. There are four basic types of scale: nominal, ordinal, interval and ratio (see Table V). The sophistication of the application for which the scales are suited increases with the progression from nominal to ratio. As the sophistication increases, so also does the information detail, the ability to differentiate the individuals, and the flexibility in using more powerful tests. For a more detailed treatment of the use of scales in OM, see Flynn *et al.* (1990).

When addressing data analysis later in this paper, we will note the importance of considering two basic kinds of data – non-metric (qualitative) and metric (quantitative):

> Nonmetric data includes attributes, characteristics, or categorical properties that can be used to identify or describe a subject. Nonmetric data differs in kind. Metric data measurement is made so that subjects may be identified as differing in amount or degree. Metrically measured variables reflect relative quantity or distance, whereas nonmetrically measured variables do not. Nonmetric data is measured with nominal or ordinal scales and metric variables with interval or ratio scales (Hair *et al.*, 1992).

Respondent identification. Very often the unit of analysis in OM research is the plant or the company. However the plant (company) cannot give the answers: it is the people who work in the plant (company) that provide information on that plant (company).

Due to the functional specialisation and hierarchical level in the organization, some people are knowledgeable about some facts while others know only about others. The researcher should therefore identify the appropriate informants for each set of information required. Increasing the number of respondents, however, increases the probability of receiving only some completed questionnaires, leading to incomplete information, which can impact on the results of relational studies. On the other hand, answers from respondents who are not knowledgeable cannot be trusted and increase random or even bias error.

Further, if perceptual questions are asked, one can gather a perception which is very personal. In order to enhance confidence in findings, the researcher can

Basic scale type	What highlights	Scaling technique
Nominal	Difference	Multiple choice items, adjective checklist, stapel scale
Ordinal	Difference, order	Forced ranking scale, paired comparison scale
Interval	Difference, order, distance	Likert scale, verbal frequency scale, comparative scale, semantic differential scale
Ratio	Difference, order, distance with 0 as meaningful natural origin	Fixed sum scale

Table V.
Scales and scaling techniques

use some form of triangulation such as the use of multiple respondents for the same question or the use of multiple measurement methods (for example qualitative and quantitative). These actions reduce the common method/source variance, i.e. potentially inflated empirical relationships which can occur when the data have been collected using the same method or have been provided by the same single source (Rungtusanatham *et al.*, 2001). O'Leary-Kelly and Vokurka (1998) and Boyer and Pagel (2000) discuss this issue in relation to research on manufacturing flexibility, operations strategy and manufacturing technology.

Rules of questionnaire design. Once the questions have been developed and their associations to respondent(s) have been established the researcher can put together the questionnaire (Converse and Presser, 1986). There are some simple things that the researcher should keep in mind. Some basic rules of courtesy, presentability, readability are key for successful data collection. An attractive and neat questionnaire with an appropriate introduction, instructions, and a well-arrayed set of questions with good alignment and response alternatives will make it easier for the respondents to answer the questions. Coloured questionnaires (especially bright ones) remind the respondent about the request to complete the questionnaire.

For both the researcher and the respondent, related questions (for example "what is the percentage of customer orders received by EDI?" and "What is the percentage of of customer orders value received by EDI?") closely placed facilitate cross checks on the responses. Mixing items belonging to different measures contributes to avoiding stereotype answering. The presence of reversal questions keeps attention high. The length of the questionnaire affects the response rate and attention in filling in the questionnaire. Finally, codes can facilitate subsequent data input.

Approach companies and respondents

To increase the probability of the success of data collection the researcher should carefully plan the execution of survey research and provide detailed instruction on the following: how sampling units are going to be approached; and how questionnaires are going to be administered. In other words, the protocol to be followed in administering the developed questionnaire has to be developed.

Increasingly, companies and respondents are being asked to complete questionnaires, and are becoming more reluctant to collaborate. Researchers, therefore, must find ways of obtaining the collaboration of companies and specific respondents. Dillman (1978) underlines that the response to a questionnaire should be viewed as a social exchange, suggesting that the researcher should:

- reward the respondent by showing positive regard, giving verbal appreciation, using a consulting approach, supporting his or her values, offering tangible rewards, and making the questionnaire interesting;

- reduce costs to the respondent by making the task appear brief, reducing the physical and mental efforts that are required, eliminating chances for embarrassment, eliminating any implication of subordination, and eliminating any direct monetary cost;
- establish trust by providing a token of appreciation in advance, identifying with a known organisation that has legitimacy, or building on other exchange relationships.

An additional problem in OM survey research lies in the difficulty of reaching the right respondent. Very often researchers send a questionnaire to a company without the name of the respondent. In this case there is a high probability that the questionnaire will be lost or delivered to a person which is not interested (or knowledgeable) on the subject. The contact strategy should take this problem into account and vary the approach based on such influencing variables as, for example, company size which can influence the presence of certain professional/managerial positions.

Dillman (1978) provides detailed advice on achieving very high response rates. In OM Flynn *et al.* (1990, 1997) suggest – and also successfully implemented – a contact strategy based on contacting potential respondents and obtaining their commitment to questionnaire completion, prior to distribution. When respondents understand the purpose of a study, lack of anonymity may not be so problematic. This facilitates the provision of feedback to respondents, which may serve as an incentive to participation. This also establishes personal contacts, which facilitates the acquisition of missed data.

Pilot testing the questionnaire

Purpose and modality of pilot testing

Once the questionnaires, the protocol to follow in administering these questionnaires, and the identity of sampling units are defined, the researcher has to examine the measurement properties of the survey questionnaires and examine the viability of the administration of these surveys. In other words, the researcher has to test what has been designed. It is remarkable the number of problems that testing can highlight even when all the previous steps have been followed with maximum attention.

Pre-testing a questionnaire should be done by submitting the "final" questionnaire to three types of people: colleagues, industry experts and target respondents. The role of colleagues is to test whether the questionnaire accomplishes the study objectives (Dillmann, 1978). The role of industry experts is to prevent the inclusion of some obvious questions that might reveal avoidable ignorance of the investigator in some specific area. The role of target respondents is to provide feedback on everything that can affect answering by and the answer of the targeted respondents. The target respondents can pre-test the questionnaire separately or in a group. If the questionnaire is mailed it can be sent to a small pre-testing sample. Telephone questionnaires must be

tested by telephone as some aspects cannot be tested in a face-to-face situation (Dillmann, 1978). This type of questionnaire is easy to test and the researcher can modify and use the revised questionnaire the same day.

From experience, I propose that the best way to pre-test a self-administered questionnaire is to proceed in two phases, each with completely different but complementary objectives.

In the first phase the researcher fills in the questionnaire with a group of potential respondents (Fowler, 1993) or when visiting three to four potential respondents. The respondents should complete the questionnaire as they would if they were part of the planned survey. Meanwhile the researcher should be present, observing how respondents fill in the questionnaire and recording the feedback. Subsequently the researcher can ask whether:

- the instructions were clear;
- the questions were clear;
- there were any problems in understanding what kind of answers were expected, or in providing answers to the questions posed; and
- the planned administration procedure would be effective.

In the second phase (not always performed in OM surveys) the researcher carries out a small pre-test sample (for example 15 units) to test the contact-administration protocol, to gather data to perform an exploratory assessment of measurement quality, and to obtain information to define better the sample and the adequacy of measures in relation to the sample. In this phase the researcher can also carry out a preliminary analysis of the data to investigate:

- whether the answers to certain questions are too concentrated due to the choice of scale;
- whether the content of answers differs from what was expected; and
- whether the context modifies the appropriateness of questions (for example, a question can be meaningful for B2B companies but not for B2C companies, or can be appropriate for medium-size companies but not for very small or large companies).

Furthermore, it may be possible to see the effects of missing data and non-response bias in order to define appropriate countermeasures. This pilot study can help to define the sample better and to plan for a "controlled sample" instead of the "observational" one which is generally more problematic but unfortunately more common in OM. In sum, this pilot test should resemble as closely as possible the actual survey that will be conducted for theory testing.

Handling non-respondents and non-response bias

Non-respondents alter the sample frame and can lead therefore to a sample that does not represent the population even when the sample was adequately designed for that purpose. Non-respondents, as such, can limit the generalisability of results. In the pilot testing phase the researcher should

identify a way to address this problem. For the OM discipline, it is important to reach a response rate that is greater than 50 per cent (Flynn *et al.*, 1990), as is found in the other social sciences. Other researchers set the limit at 20 per cent (Malhotra and Grover, 1998). This point is much debated since many researchers find it hard to agree on the response rate percentages. However, especially for theory-testing survey research, the example provided by Fowler (1993, p. 43) – and reported in Table VI – is instructive.

Fowler estimates the presence of blond-haired persons in a population of 100 persons with 25 blonde-haired individuals. If the response rate is 70 per cent and 75 per cent of non-respondents have blond hair, it means that out of the 30 non-respondents 0,75*30~22 have blond hair and therefore only 25-22=3 blond-haired individuals respond. Therefore, the estimate is three blond-haired persons in the population while, in reality, there are 25 such individuals. Table VI shows that when there are major biases (such that non-respondents have characteristics – e.g. blond hair – systematically different from the respondents) even studies with response rates of approximately 70 per cent produce considerable errors in estimates. When response rates are lower, estimates are not very good even when bias is modest. The problem is that "one usually does not know how biased non-response is, but [it] is seldom a good assumption that non-response is unbiased" (Fowler, 1993).

OM researchers could consider articles from other disciplines in order to increase awareness on non-respondent causes (see Roth and BeVier, 1998; Greer *et al.*, 2000) and effects (see Wilson, 1999) which underpin the resulting lack of external validity). To calculate the response rate the researcher can refer to Dillman (1978, pp. 49-52), who provides some details on how to do this.

The non-respondent problem can be addressed in two ways:

(1) by trying to increase response rate; and

(2) by trying to identify the non-respondents to control whether they are different from the respondents.

Response rates can be increased considerably when a subsequent follow-up programme is applied:

- after one week a postcard is sent to everyone (it serves as a reminder and as a thank you);

Response rate (%)	Bias level (percentage of non-respondents with characteristics (blond hair)) (10)	(20)	(25)	(30)	(40)	(50)	(75)
90	27	26	25	24	23	22	19
70	31	27	25	23	19	14	3
50	40	30	25	20	10		
30	60	37	25	13			

Source: Fowler (1993, p. 43)

Table VI.
Effect of biased non-response on survey estimates

- after three weeks a letter and a replacement questionnaire are sent only to non-respondents; and
- final mailing similar to the previous one (or even a telephone call).

Dillman (1978) provides detailed information on follow-up strategies. From experience, I propose that a phone call is more useful, since it makes it possible to:

- ensure that the target respondent has received the questionnaire;
- establish a personal contact;
- have some captive minutes to explain the research;
- help the respondent; and
- gather some information on non-respondents.

Researchers should at least keep track of the non-respondents. They should survey some of them (even using a condensed questionnaire or using a telephone call) to understand whether and how much bias has been introduced (see for example Ward *et al.* (1994). An alternative method is to check for differences between the first wave of respondents and later returns (Lambert and Harrington, 1990).

Since OM tends to rely on small sample sizes it would be useful at this point to check the credibility of the available sample. Sudman (1983, p. 154-63) provides a scale (see Table VII) to evaluate the credibility of a small sample. In the range [– 8...5] the survey credibility is very poor, [6...15] limited credibility, [16...25] good credibility and [26...35] very good credibility. These scores are qualitative judgements and not quantitative evaluations, and as such they have

Characteristics	Score
Generalisability	
Geographic spread	Single location (0), several combined or compared locations [(4) if limited geography, (6) if widespread geography], total universe (10)
Discussion of limitation	No discussion (0), brief discussion (3), detailed discussion (5)
Use of special populations	Obvious biases in the sample that could affect results (– 5), used for convenience with no obvious bias (0), necessary to test theory (5), general population (5)
Sample size	Too small for meaningful analysis (0), adequate for some but not all major analyses (3), adequate for purpose of study (5)
Sample execution	Haphazard sample (– 3), poor response rate (0), some evidence of careless field work (3), reasonable response rate and controlled field operations (5)
Use of resources	Poor use of resources (0), fair use of resources (3), optimum use of resources (5)
Range of points	[– 5 ... 35]

Table VII.
Credibility scale for small samples

Source: Adapted from Sudman (1983, p. 154)

some degree of arbitrary but are able to discriminate in a consistent way between different levels of sample credibility.

Behind these scores are some characteristics. Usually a sample taken from a limited geographic area represents the population less than a sample taken from multiple locations. Articles which discuss possible sample bias are more credible than those that do not. The use of a special population in some cases is a powerful tool to test a theory but if used for convenience it can introduce obvious biases. It is possible that sample sizes are satisfactory when the total sample is considered but, after breakdowns, the resulting sub-samples may not be adequate in size for more detailed analyses. When the response rate is poor it is very likely that some bias has been introduced by self-selection of respondents. Sometimes the researcher is pressed by lack of time or cost or resources; even in this case some sample designs are more effective in using the available resources than others.

To give an example of the application of the credibility scale, consider a sample drawn from plants located in a town of 100,000 inhabitants (0 points), with no discussion of biases (0 points), taken from the list of companies associated with the local industrial association (0 points), with a size adequate for the purpose of the study (5 points), with a reasonable response rate and care in controlling data collection (5 points), which performed a telephone questionnaire with a limited budget and little available time (5 points). This sample tots up 15 points and, therefore, its credibility is limited.

Inputting and cleaning data

The first step in processing data usually entails transcribing the data from the original documents to a computer database. In this process, about 2-4 per cent of the data can be incorrectly transcribed (Swab and Sitter, 1974, p. 13). The errors arise from two situations: the transcriber misreads the source document but correctly transcribes the misinterpreted data (86 per cent of transcription errors are of this type); and the transcriber reads the source document correctly but incorrectly transcribes the data (Karweit and Meyers, 1983). Independent verification of any transcription involving the reading and interpretation of hand-written material is therefore advisable.

When an error is detected the researcher may use the following options, singly or in combination, to resolve the error (Karweit and Meyers, 1983):

- consult the original interview or questionnaire to determine if the error is due to incorrect transcription;
- contact the respondent again to clarify the response or obtain missing data;
- estimate or impute a response to resolve the error using various imputation techniques;
- discard the response or designate it as bad or missing data;
- discard the entire case.

In the last 20-30 years, progress have been made in the way data are collected and cleaned. Computers with screens and keyboards made obsolete keypunch operators. Optical scanning and Web based questionnaires allow automatic inputting of data thus reducing errors. Computer, assisted personal (CAPI) or telephone (CATI) interviewing allow interviews to be completed with answers entered directly in database, thus reducing intermediate steps and errors. The data input programs can perform checks on the data (ensuring, for example, that the values are within a certain range, or that other logical constraints are satisfied). New techniques are available not only for inputting data but also for distributing and even developing questionnaires. "Integrated" software, such as SPSS Data Entry Survey Software or Sphinx Survey, assist in questionnaire development, questionnaire distribution (on www for example), building the database and analysis of the collected data.

Assessing the measurement quality

Importance of ensuring and assessing measurement quality. The section entitled "How should a survey be designed?" highlighted that when researchers move from the theoretical level to the empirical level they must operationalise the constructs present in the theoretical framework. Carmines and Zeller (1990) note that "if the theoretical constructs have no empirical referents, then the empirical tenability of the theory must remain unknown". When measurements are unreliable and/or invalid, analysis can possibly lead to incorrect inferences and misleading conclusions. Without assessing reliability and validity of measurement it would be impossible to "disentangle the distorting influences of [measurement] errors on theoretical relationships that are being tested" (Bagozzi *et al*., 1991).

Measurement error represents one of the major sources of error in survey research (Biemer *et al*., 1991; Malhotra and Grover, 1998) and should be kept at the lowest possible level. Furthermore, recognising how much it affects the results, it should be known to the researchers as well as to the readers.

When we address the issue of measurement quality, we think of the quality of the survey instruments and procedures used to measure the constructs of interest. However, the most crucial aspect concerns the measurement of complex constructs by multi-item measures, the focus of the remaining part of this section.

Measure quality criteria. The goodness of measures is mainly evaluated in terms of validity and reliability. Validity is concerned with whether we are measuring the right concept, while reliability is concerned with stability and consistency in measurement. Lack of validity introduces a systematic error (bias), while lack of reliability introduces random error (Carmines and Zeller, 1979). There are discussed below:

(1) *Reliability*. Reliability indicates dependability, stability, predictability, consistency and accuracy, and refers to the extent to which a measuring procedure yields the same results on repeated trials (Kerlinger, 1986;

Carmines and Zeller, 1979). Reliability is assessed after data collection. The four most common methods used to estimate reliability are:

- test-retest method;
- alternative form method;
- split halves method; and
- internal consistency method.

Core readings on this issue are Nunnally (1978) and Carmines and Zeller (1979).

The test-retest method calculates the correlation between responses obtained through the same measure applied to the same respondents at different points of time (e.g. separated by two weeks). It estimates the ability of the measure to maintain stability over time. This aspect is indicative of the measure stability and low vulnerability to change in uncontrollable testing conditions and in the state of the respondents.

The alternative form method calculates the correlation between responses obtained through different measures applied to the same respondents in different points of time (e.g. separated by two weeks). It assesses the equivalence of different forms for measuring the same construct.

The split halves method subdivides the items of a measure into two subsets and statistically correlates the answers obtained at the same time to them. It assesses the equivalence of different sets of items for measuring the same construct.

The internal consistency method uses various algorithms to estimate the reliability of a measure from measure administration at one point in time. It assesses the equivalence, homogeneity and inter-correlation of the items used in a measure. This means that the items of a measure should hang together as a set and should be capable of independently measuring the same construct. The most popular test within the internal consistency method is the Cronbach coefficient alpha (Cronbach, 1951). Cronbach's alpha is also the most used reliability indicator in OM survey research. Cronbach's α can be expressed in terms of $\bar{\rho}$, the average inter-item correlation among the n measurement items in the instrument under consideration, in the following way:

$$\alpha = \frac{n\bar{\rho}}{1 + (n-1)\bar{\rho}}.$$

Cronbach's α is therefore related to the number of items, n, as well as to the average inter-item correlation $\bar{\rho}$. Nunnally (1978) states that new developed measures can be accepted with $\alpha \geq 0.6$, otherwise $\alpha \geq 0.7$ should be the threshold. With $\alpha \geq 0.8$ the measure is very reliable. These criteria are well accepted in OM. Computation of Cronbach's α coefficient is well supported by statistical packages.

(2) *Construct validity*. Of the different properties that can be assessed about a measure, construct validity is the most complex and, yet, the most critical to substantive theory testing (Bagozzi *et al.*, 1991). For details and examples of application in OM see Rungtusanatham and Choi (2000) and O'Leary-Kelly and Vokurda (1998). However, the concept of construct validity deserves further consideration by OM researchers in the context of recent developments in other social sciences disciplines, such as the notion of validity as an unified concept proposed by Messick (1995).

A measure has construct validity if the set of items constituting a measure faithfully represents the set of aspects of the theoretical construct measured, and does not contain items which represent aspects not included in the theoretical construct. "Since the construct cannot be directly addressed empirically, only indirect inference about construct validity can be made by empirical investigation" (Flynn *et al.*, 1990). Indeed, "in attempting to evaluate construct validity we must consider both the theory of which the construct is part and the measurement instrument being used" (Emory and Cooper, 1991).

The empirical assessment of construct validity basically focuses on convergence between measures (or items of a measure) of the same construct (convergent validity) and separation between measures (or items of a measure) of different constructs (discriminant validity). When a test, conducted to assess an aspect of construct validity, does not support the expected result, either the measurement instrument or the theory could be invalid. It is a matter of researcher judgement to interpret adequately the obtained results. For details see Bagozzi *et al.* (1991) and O'Leary-Kelly and Vokurda (1998).

Testing for consistency across measurement items for the same construct is well established in OM. This form of convergent validity is called construct unidimensionality. Saraph *et al.* (1989) and Flynn *et al.* (1994) use exploratory factor analysis to check unidimensionality, while Ahire *et al.* (1996) use confirmatory factor analysis. Factor analysis can be performed on items belonging to a single summated scale or items of several summated scales (Flynn *et al.*, 1990; Birnbaum *et al.*, 1986). Factor analysis procedures are well supported by statistical packages (see Hatcher, 1994).

Testing for separation across measures of different constructs (discriminant validity) is not common practice in OM. It can be assessed through confirmatory factor analysis on items belonging to measures of different constructs (see for example Koufteros (1999)). The number of factors and the list of factors which load on each dimension should be specified *a priori*. Comparing the results of factor analysis with the pre-specified factors and loadings, the researcher can obtain an indication of the construct validity.

(3) *Criterion-related validity*. "When an instrument is intended to perform a prediction function, validity depends entirely on how well the instrument correlates with what it is intended to predict (a criterion)" (Nunnally, 1978, p. 111).

Criterion-related validity is established when the measure differentiates individuals on a criterion it is expected to predict. Establishing concurrent validity or predictive validity can do this. Concurrent validity is established when the scale discriminates individuals who are known to be different. Predictive validity is the ability of the measure to differentiate among individuals as to a future criterion.

In OM criterion-related validity has been supported using multiple correlations (see Saraph *et al.*, 1989), canonical correlations (see Flynn *et al.*, 1994), and LISREL (see Ahire *et al.*, 1996) Rungtusanatham and Choi (2000).

Steps in assessing validity and reliability. Developing valid and reliable measures is a process parallel to that aimed at building and testing a theory. Here, measures go through a process of developing and testing (see for example the framework for developing multi-item measures provided by Malhotra and Grover (1998)). The aim is not only to build an instrument to allow theory testing but also to have an instrument reusable for other theories as well as for application purposes.

When developing measures (in a pilot-testing phase or in an exploratory research), cut-off levels (for Cronbach alpha) are less stringent and, due to small sample sizes, assessments (of unidimensionality) are of an exploratory nature (Nunnally, 1978). The number of different types of validity and reliability assessment is limited.

When testing measures (after data collection for hypothesis testing) cut-off levels are set at higher values, confirmatory methods should be used and all the various relevant aspects of validity and reliability should be considered. If an already-developed measure is used in a modified form then the measure quality should be re-assessed and contrasted with one from the previous version.

Assessing measure quality therefore takes place at various stages of survey research: before data collection, within pilot testing and after data collection for hypothesis testing. However, conducting reliability and validity assessments can be organised as a three-step, iterative process: face validity assessment, reliability assessment and construct validity assessment (Rungtusanatham and Choi, 2000). The elimination of items in the second and third steps requires the researcher to return to the first step and redo the analyses for the modified measure. Examples of application are Parasuraman *et al.* (1988) and Saraph *et al.* (1989).

Survey execution

Redo activities to a larger sample

At the end of pilot testing, either the researcher can proceed with theory testing or the survey questionnaires, the survey administration process, and/or both

would have to be revised. In the latter case, the researcher would have go back to look at the issues raised in the sections entitled "How should a survey be designed?" and "Pilot testing the questionaire". Therefore the researcher should move to the survey execution phase only when all relevant issues have been addressed. Ideally, data collection problems and measurement problems should have been reduced to the minimum level. Therefore, at survey execution the researcher has an opportunity to direct attention elsewhere until the data have been returned.

Fundamentally the researcher in this phase has to repeat the pilot-testing activities with a large sample:

- approach companies/respondents and collect data;
- control and reduce the problems caused by the non-respondents;
- perform data input and cleaning;
- if possible, recall companies to reduce problematic/missing data; and
- assess measurement quality.

A further activity is providing feedback to companies/respondents in order to motivate their present and future involvement. This feedback could be a standard summary report, personalised feedback, invitation to meetings where results are communicated, or something else that could be useful to the respondents.

Handling missing data

Handling missing data should be a key concern during data collection. "When statistical models and procedures are used to analyse a random sample, it is usually assumed that no sample data is missing. In practice, however, this is rarely the case for survey data" (Anderson *et al.*, 1983). A review of the literature regarding how to handle randomly missing data is provided by Anderson *et al.* (1983). Sometimes data can be estimated or reconstructed due to redundancies in the data themselves. However, the best approach is to prevent the presence of missing data by increasing respondent involvement, giving clear instructions, a well-designed questionnaire, support and recall to ensure completeness. Despite all efforts some data will be missed. Two broad strategies can be adopted: deletion and estimation.

> When data is missed randomly the estimates resulting from deletion strategy are generally unbiased (but may have to be adjusted by correction terms) but less efficient than when no data is missed . . . The second broad strategy first estimates the missing observation in some way and then proceeds with a statistical analysis of the data set as if it had been completed . . . The most common procedure for estimating randomly missing values in socio-economic data is, however, by regression, principal components, or factor analysis performed on the variables (Anderson *et al.*, 1983).

Link measure quality assessment to hypothesis testing

This section highlighted that measurement quality assessment can be done in an exploratory way when pilot testing. Further, it deserves confirmatory analyses

when doing the analyses with the data which will be used to test hypotheses. However this is not enough to be very accurate in the analysis. Traditionally, in fact, procedures to assess measure validity-reliability are "applied independently of statistical procedures to test causal hypotheses ... [The consequence is that] whereas construct validation procedures typically establish the presence of significant amounts of measurement and/or method error, contemporary hypothesis-testing procedures assume it away entirely" (Bagozzi *et al.*, 1991). Measurement and method error can cause "spurious confirmation of inadequate theories, tentative rejection of adequate theories, and/or distorted estimates of the magnitude and relevance of actual relationships" (Bagozzi *et al.*, 1991). Structural equation modelling (also known as LISREL) provides an instrument to test measurement quality and to consider it while testing the hypotheses. An exemplary application in OM can be found in Koufteros (1999).

Now that you have good data, what statistical methods can you use?

Data analysis can be schematically divided into two phases: preliminary data analysis and hypothesis testing. These phases are described below and the most commonly used data analysis methods are presented briefly. The objective is to provide some information to complete the overview of the theory-testing survey research process. However, this issue deserves far more discussion and the reader is encouraged to pursue this issue further in statistical manuals and with statisticians.

Before getting into the details of the analysis we should briefly look at the kind of data analyses that have been used in OM. Scudder and Hill (1998) analysed the method used in 477 OM empirical research articles published during the period 1986-1995 in the 13 main journal outlets for OM research. They found that 28 per cent of articles did not use any statistical data analysis method (almost all of these articles were based on case studies), while some articles used more than one data analysis method. Furthermore they found that 72 per cent of articles used descriptive statistics, 17 per cent regression/correlation, 9 per cent means testing, 7 per cent data reduction (principal component analysis, etc.), 4 per cent ANOVA and MANOVA, and 3 per cent cluster analysis.

Preliminary data analysis

To acquire knowledge of the characteristics and properties of the collected data some preliminary data analyses are usually performed before performing measurement quality assessment or conducting tests of hypotheses. Carrying out such analyses before assessing measurement quality gives preliminary indications of how well the coding and entering of data have been done, how good the scales are, and whether there is a suspicion of poor content validity or systematic bias. Before testing hypotheses, it is useful to check the assumptions underlying the tests, and to get a feeling for the data in order to interpret the results of the tests better.

Preliminary data analysis is performed by checking central tendencies, dispersions, frequency distributions, correlations. It is good practice to calculate:

- the frequency distribution of the demographic variables;
- the mean, standard deviation, range and variance of the other dependent and independent variables; and
- an inter-correlation matrix of the variables.

Table VIII gives some of the most frequently used descriptive statistics used within preliminary data analysis. Some statistical packages (for example SAS) provide tools for exploratory or interactive data analysis which facilitate preliminary data analysis activities through emphasis on visual representation and graphical techniques.

For suggestions on distribution displaying and examination techniques in business research see Emory and Cooper (1991). They note that:

> . . . frequency tables array data from highest to lowest values with counts and percentages . . . are most useful for inspecting the range of responses and their repeated occurrence. Bar-charts and pie-charts are appropriate for relative comparisons of nominal data, while histograms are optimally used with continuous variables where intervals group the responses (Emory and Cooper, 1991, p. 509).

Emory and Cooper suggest also using stem-and-leaf displays and boxplots since they are:

> . . . exploratory data analysis techniques that provide visual representations of distributions. The former present actual data values using a histogram-type device that allows inspection of spread and shape. Boxplots use five-number summary to convey a detailed picture of the main body, tails, and outliers of the distribution. Both rely on resistant statistics to overcome the limitations of descriptive measures that are subject to extreme scores. Transformation

Type of analysis	Explanation	Relevance
Frequencies	Refers to the number of times various subcategories of certain phenomenon occur	Generally obtained for nominal variables
Measures of central tendencies	Mean (the average value), median (half of the observation fall above and the other half fall below the median) and mode (the most frequently occurring value) characterise the central tendency (or location or centre) of a set of observations	To characterise the central value of a set of observations parsimoniously in a meaningful way
Measures of dispersion	Measures of dispersion (or spread or variability) include the range, the standard deviation, the variance, and the interquartile range	To concisely indicate the variability that exists in a set of observations
Measures of shape	The measures of shape, skewness and kurtosis describe departures from the symmetry of a distribution and its relative flatness (or peakedness), respectively	To indicate the kind of departures from a normal distribution

Table VIII.
Descriptive statistics

> may be necessary to re-express metric data in order to reduce or remove problems of asymmetry, inequality of variance, or other abnormalities.

Finally they highlight the possibility of using cross-tabulations to perform preliminary evaluation of relationships involving nominally scaled variables. "The tables used for this purpose consist of cells and marginals. The cells contain combination of count, row, column, and total percentages. The tabular structure is the framework for later statistical testing".

Analyse data for hypothesis testing

Significance tests can be grouped into two general classes: parametric and non-parametric. Parametric tests are generally considered more powerful because their data are typically derived from interval and ratio measurements when the likelihood model (i.e. the distribution) is known, except for some parameters. Non-parametric tests are also used, with nominal and ordinal data. Experts on non-parametric tests claim that non-parametric tests are comparable in terms of power (Hollander and Wolfe, 1999). However, in social science at the moment:

> ... parametric techniques are [considered] the tests of choice if their assumptions are met. Some of the assumptions for parametric tests include:
> (1) the observations must be independent (that is, the selection of any one case should not affect the chances for any other case to be selected in the sample);
> (2) the observation should be drawn from normally distributed populations;
> (3) these populations should have equal variance;
> (4) the measurement scales should be at least interval so that arithmetic operations can be used with them.
>
> The researcher is responsible for reviewing the assumptions pertinent to the chosen test and performing diagnostic checks on the data to ensure the selection's appropriateness ... Parametric tests place different emphases on the importance of assumptions. Some tests are quite robust and hold up well despite violations. With others, a departure from linearity or equality of variance may threaten result validity. Nonparametric tests have fewer and less stringent assumptions. They do not specify normally distributed populations or homogeneity of variance. Some tests require independent cases while others are expressly designed for situations with related cases (Emory and Cooper, 1991).

Therefore, when the population distribution is undefined, or violates assumption of parametric tests, non-parametric tests must be used.

In attempting to choose a particular significance test, at least three questions should be considered (Emory and Cooper, 1991):

(1) does the test involve one sample, two sample or k samples?

(2) If two samples or k samples are involved, are the individual cases independent or related?

(3) Is the measurement scale nominal, ordinal, interval or ratio?

Additional questions may arise once answers to these are known. For example, what is the sample size? If there are several samples, are they of equal size? Have the data been weighed? Have the data been transformed? The answers

can complicate the selection, but once a tentative choice is made, most standard statistic textbooks will provide further details. Decision trees provide a more systematic means of selecting techniques. One widely used guide from the Institute for Social Research (Andrews *et al.*, 1976) starts with a question about the number of variables, nature of variables and level of measurement and continues with more detailed ones, so providing indications to over 130 solutions.

Tables IX and X give examples of some parametric (Table IX) and non-parametric tests (Table X).

In any applied field, such as OM, most tools are, or should be, multivariate. Unless a problem is treated as a multivariate problem in these fields, it is treated superficially. Therefore multivariate analysis (simultaneous analysis of more than two variables) is, and will continue to be, very important in OM. Table XI presents some of the more established techniques as well as some of the emerging ones (for more details see Hair *et al.* (1992)).

Interpret results

The choice and the application of an appropriate statistical test is only one step in data analysis for theory testing. In addition, the results of the statistical tests must be interpreted. When interpreting results the researcher moves from the empirical to the theoretical domain. This process implies considerations of inference and generalisation (Meredith, 1998).

In making an inference on relations between variables, the researcher could incur a statistical error or an internal validity error. The statistical error (see type I and type II errors discussed earlier) can be taken into account by considering the issue of statistical power, significance level, sample size, effect size. The internal validity error erroneously attributes the cause of variation to a dependent variable. For example, the researcher can say that variable A

Test	When used	Function
Pearson correlation	With interval and ratio data	Test hypothesis which postulates significant positive (negative) relationships between two variables
t-test	With interval and ratio data	To see whether there is any significant difference in the means for two groups in the variable of interest. Groups can be either two different groups or the same group before and after the treatment
Analysis of variance (ANOVA)	With interval and ratio data	To see whether there are significant mean differences among more than two groups. To see where the difference lies, tests like Sheffe's test, Duncan Multiple Range test, Tukey's test, and student-Newman-Keul's test are available

Table IX.
Example of parametric tests

Test	When used	Function
Chi-squared (χ^2)	With nominal data for one sample or two or more independent samples	Test for equality of distributions
Cochran Q	With more than two related samples measured on a nominal scale	Similar function as χ^2, it helps when data fall into two natural categories
Fisher exact probability	With two independent samples measured on a nominal scale	More useful than χ^2 when expected frequencies are small
Sign test	With two related samples measured on an ordinal scale	Test for equality of two group distributions
Median test	With one sample	To test the equality in distribution under the assumption of homoschedasticicity
Mann-Witney U test	With two independent samples on ordinal data	Analogue to the two independent sample t-tests with ordinal data
Kruskall-Wallis one-way ANOVA	With more than two independent samples on an ordinal scale	An alternative to one-way ANOVA with ordinal data
Friedman two-way ANOVA	With more than two related samples on ordinal data	Analogue to two way ANOVA with ranked data when interactions are assumed absent
Kolmogorov-Smirnov	With one sample or two independent samples measured on an ordinal scale	Test for equality of distribution with ordinal scale

Source: Adapted from Sekaran, 1992 p. 279

Table X.
Example of non-parametric tests

causes variable B, while there is an un-acknowledged variable C which causes both A and B. The link that the researcher observes between A and B is therefore spurious. "POM researchers, in the absence of experimental designs, should try to justify internal validity. This can be done informally through a discussion of why causality exists or why alternate explanations are unlikely" (Malhotra and Grover, 1998).

Even in the situation when data analysis results are consistent with the theory at the sample level, the researcher should take care in inferring that the same consistency holds at the population level, because of previous discussed issues of response rate and response bias. A further facet of result interpretation relates to the discussion of potential extension of the theory to other populations.

What information should be in written reports?

In the written report the researcher should provide, in a concise but complete manner, all of the information which allows reviewers and readers to:

- understand what has been done;
- evaluate critically what the work has achieved; and
- reproduce the work or compare the results with similar studies.

Multivariate technique	When used	Function
Multiple regression	With a single metric dependent variable presumed to be related to one or more metric independent variables	To predict the changes in the dependent variable in response to changes in the several independent variables
Multiple discriminant analysis	When the single dependent variable is dichotomous (e.g. male-female) or multidichotomous (e.g. high-medium-low) and therefore nonmetric	To understand group differences and predict the likelihood that an entity (individual or object) will belong to a particular class or group based on several metric independent variables
Multivariate analysis of variance (MANOVA) Multivariate analysis of covariance (MANCOVA)	Useful when the researcher designs an experimental situation (manipulation of several non-metric treatment variables) to test hypotheses concerning the variance in group response on two or more metric dependent variables	To simultaneously explore the relationship between several categorical independent variables (usually referred to as treatments) and two or more dependent metric variables
Canonical correlation	An extension of multiple regression analysis	To simultaneously correlate several metric independent variables and several dependent metric variables
Structural equation modelling	When multiple separate regression equations have to be estimated simultaneously	To simultaneously test the measurement model (which specifies one or more indicator to measure each variable) and the structural model (the model which relates independent and dependent variables)
Factor analysis	When several metric variables are under analysis and the researcher wishes to reduce the number of variables to manage or to find out the underlying factors	To analyse interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (factors)
Cluster analysis	When metric variables are present and the researcher wishes to group entities	To classify a sample of entities (individuals or objects) into a smaller number of mutually exclusive subgroups based on the similarities among the entities

Table XI.
Main multivariate analysis methods

To understand which information is to be included one can refer to Verma and Goodale (1995), Malhotra and Grover (1998), Forza and Di Nuzzo (1998), Hensley (1999), Rungtusanatham *et al.* (2001). The main points to consider are summarised in Table XII.

All the information listed in Table XII is necessary if the article has a theory-testing purpose and should satisfy the requirements that were discussed throughout this paper.

Main issues	Detailed points
Theoretical base	Name and definitions of constructs, relations between variables, validity boundary of the relations, unit of analysis, previous literature on each of these points
Expected contribution	Purpose of the study (whether it is exploration, description, or hypothesis testing), research questions/hypotheses, types of investigation (causal relationships, correlations, group differences, ranks, etc.)
Sample and data collection approach	Sampling process, source of population frame, justification of sample frame, *a-priori* sample, resulting sample, response rate, bias analysis Time horizon (cross-sectional or longitudinal), when and where data have been collected, type of data collection (mail, telephone, personal visit), pilot testing, contact approach, kind of recall
Measurement	Description of measure construction process, reference/comparison to similar/identical measures, description of respondents, list of respondents for each measure, measure pre-testing, adequacy to the unit of analysis, adequacy to the respondents, face validity, construct validity, reliability, appendix with the measurement instrument, description of the measurement refinement process including information on techniques used, description of the data aggregation process (from informants to unit of analysis)
Data analysis	Description of the techniques used, evidence that the technique assumptions are satisfied, statistical power, results of the tests including level of significance, interpretation of the results in the context of the hypotheses
Discussion	Discusses what the substantiation of the hypotheses means in terms of the present research and why some of the hypotheses (if any) may not have been supported Consider through intuitive but appropriate and logical speculations how inadequacies in the sampling design, the measures, the data collection methods, control of critical variables, respondent bias, questionnaire design and so on effect the results, their trustability and their generalisability

Table XII.
Information to include in the report

Descriptive and exploratory survey research are important and widely used in OM. Therefore, in concluding this paper it is useful to outline the different requirements of the various types of survey. Obviously if a particular requirement is relaxed then the necessary information detail regarding this requirement diminishes. Table XIII summarises the differences in requirements among different survey types.

Final considerations and conclusions

This paper has focused on theory-testing survey research in OM, since it is the most demanding type of survey research, and has showed how the requirements can be shaped if the researcher is to consider descriptive or exploratory survey research.

The paper has presented and discussed the various steps in a theory-testing survey research process. For each step the paper has provided responses to the following questions:

Survey type element/dimension	Exploratory	Descriptive	Theory testing
Unit(s) of analysis	Clearly defined	Clearly defined and appropriate for the questions/hypotheses	Clearly defined and appropriate for the research hypotheses
Respondents	Representative of the unit of analysis	Representative of the unit of analysis	Representative of the unit of analysis
Research hypotheses	Not necessary	Questions clearly stated	Hypotheses clearly stated and theoretically motivated
Representativeness of sample frame	Approximation	Explicit, logical argument; reasonable choice among alternatives	Explicit, logical argument; reasonable choice among alternatives
Representativeness of the sample	Not a criterion	Systematic, purposive, random selection	Systematic, purposive, random selection
Sample size	Sufficient to include the range of the interest phenomena	Sufficient to represent the population of interest and perform statistical tests	Sufficient to test categories in the theoretical framework with statistical power
Pre-test of questionnaires	With subsample of sample	With subsample of sample	With subsample of sample
Response rate	No minimum	Greater than 50 per cent of targeted population and study of bias	Greater than 50 per cent of targeted population and study of bias
Mix of data collection methods	Multiple methods	Not necessary	Multiple methods

Source: Adapted from Pindonneault and Kramer (1993)

Table XIII. Requirements difference among surveys

(1) What is this step?

(2) Why should it be done?

(3) What is suggested to be done?

Throughout, the paper has provided references to examples of applications in OM and to a more general reference literature. Table XIV summarises the questions that the researcher should ask at the various steps of survey research as a quality control instrument.

By following the guidelines provided in this paper, the researcher should be able to execute survey research that will satisfy the main characteristics of a scientific research project as outlined by Sherakan (1992):

(1) *Purposiveness*: the researcher has started with a definite aim or purpose for the research.

(2) *Rigor*: a good theoretical base and a sound methodological plan are necessary to collect the right kind of information and to interpret it appropriately.

Survey phase	Check questions to assure survey research quality
Prior to survey research design	(1) Is the unit of analysis clearly defined for the study?
	(2) Are the construct operational definitions clearly stated?
	(3) Are research hypotheses clearly stated?
Defining the sample	(4) Is the sample frame defined and justified?
	(5) What is the required level of randomness needed for the purposes of the study?
	(6) What is the minimum sample size required for the planned statistical analyses?
	(7) Can the sampling procedure be reproduced by other researchers?
Developing measurement instruments	(8) Are already-developed (and preferably validated) measures available?
	(9) Are objective or perceptual questions needed?
	(10) Is the wording appropriate?
	(11) In the case of perceptual measures, are all the aspects of the concept equally present as items?
	(12) Does the instrumentation consistently reflect that unit of analysis?
	(13) Is the chosen scale compatible with the analyses which will be performed?
	(14) Can the respondent place the answers easily and reliably in this scale?
	(15) Is the chosen respondent(s) appropriate for the information sought?
	(16) Is any form of triangulation used to ensure that the gathered information is not biased by the respondent(s) or by method?
	(17) Are multi-item measures used (in the case of perceptual questions)?
	(18) Are the various rules of questionnaire design (see above) followed or not?
Collecting data	(19) What is the response rate and is it satisfactory?
	(20) How much is the response bias?
Assessing measure quality	(21) Is face validity assessed?
	(22) Is field-based measure pre-testing performed?
	(23) Is reliability assessed?
	(24) Is construct validity assessed?
	(25) Are pilot data used for purifying measures or are existing validated measures adapted?
	(26) Is it possible to use confirmatory methods?
Analysing data	(27) Is the statistical test appropriate for the hypothesis being tested?
	(28) Is the statistical test adequate for the available data?
	(29) Are the test assumptions satisfied?
	(30) Do outliers or influencing factors affect results?
	(31) Is the statistical power sufficient to reduce statistical conclusion error?
Interpretation of results	(32) Do the findings have internal validity?
	(33) Is the inference (both relational and representational) acceptable?
	(34) For what other populations results could still be valid?

Table XIV.
Questions to check quality of ongoing survey research

(3) *Testability*: at the end the researcher can see whether or not the data supports his conjectures or hypothesis developed after careful study of the problem situation.

(4) *Replicability*: it should be possible to repeat the study exactly. If the results are the same again and again the conjectures will not be supported (or discarded) merely by chance.

(5) *Precision and confidence*. refers to how close the findings are to "reality" and to the probability that our estimations are correct. This issue derives from our inability to observe the entire universe of aspects, events or population in which we are interested, facts which imply that the conclusions based on the data analysis results are rarely "definitive".

(f) *Objectivity*. the conclusion drawn through the interpretation of the data analysis results should be based on facts resulting from the actual data and not on our own subjective or emotional values.

(g) *Generalisability*. refers to the applicability scope of the research findings in one organisational setting to another setting.

(h) *Parsimony*. simplicity in explaining the phenomena or problems that occur, and in the application of solutions to problems, is always preferred to complex research frameworks that consider an unmanageable number of factors.

Notes

1. The concept of "content validity" has been controversial in social indicators research. This kind of validity deserves futher consideration by OM researchers in the context of recent developments in its conceptualisation (Sireci, 1998).
2. It should be noted that hypothesis generation and testing can be done both through the process of deduction (i.e. develop the model, formulate testable hypotheses, collect data, then test hypotheses) and the process of induction (i.e. collect the data, formulate new hypotheses based on what is known from the data collected and test them). This paper follows a traditional positivistic perspective and therefore refers to the first approach. However a researcher who follows a different epistemological approach can disagree. Bagozzi *et al.* (1991), for example, state that the two approaches can be applied in the same research. They propose a new methodological paradigm for organisational research called holistic construal. This approach "is neither rigidly deductive (or formalistic) nor purely exploratory. Rather it subsumes a process by which theories and hypotheses are tentatively formulated deductively and then are tested on data, and later are reformulated and retested until a meaningful outcome emerges". This approach "is intended to encompass aspects of both the theory-construction and theory-testing phases". Therefore in a paper which follow this approach we can typically observe a starting model and a refined model.

References

Ahire, S.L., Goldhar, D.Y. and Waller, M.A. (1996), "Development and validation of TQM implementation constructs", *Decision Sciences*, Vol. 27 No. 1, pp. 23-56.

Alreck, P.L. and Settle, R.B. (1985), *The Survey Research Handbook*, Irwin, Homewood, IL.

Amoako-Gyampah, K. and Meredith, J.R. (1989), "The operations management research agenda: an update", *Journal of Operations Management*, Vol. 8 No. 3, pp. 250-62.

Amundson, S.D. (1998), "Relationships between theory-driven empirical research in operations management and other disciplines", *Journal of Operations Management*, Vol. 16 No. 4, pp. 341-59.

Anderson, A.B., Basilevsky, A. and Hum, D.P.J. (1983), "Missing data", in Rossi, P.H., Wright, J.D. and Anderson, A.B., *Handbook of Survey Research*, Academic Press, New York, NY, pp. 415-94.

Anderson, J.C., Rungtusanatham, M. and Schroeder, R.G. (1994), "A theory of quality management underlying the Deming management method", *Academy of Management Review*, Vol. 19 No. 3, pp. 472-509.

Andrews, F.M., Klem, L., Davidson, T.N., O'Malley, P.M. and Rodgers, W.L. (1976), *A Guide for Selecting Statistical Techniques for Analysing Social Science Data*, Institute for Social Research, Ann Arbor, MI.

Babbie, E. (1990), *Survey Research Methods*, Wadsworth, Belmont, CA.

Bagozzi, R.P., Yi, Y. and Phillips, L.W. (1991), "Assessing construct validity in organizational research", *Administrative Science Quarterly*, Vol. 36 No. 4, pp. 421-34.

Biemer, P.P., Groves, R.M., Lyber, L.E., Mathiowetz, N.A. and Sudman, S. (1991), *Measurement Errors in Surveys*, Wiley, New York, NY.

Birnbaum, M.H. (1999), "Testing critical properties of decision making on the Internet", *American Psychological Society*, Vol. 10 No. 5, pp. 399-407.

Birnbaum, P.H., Farh, J.-L. and Wong, G.Y.Y. (1986), "The job characteristics model in Hong Kong", *Journal of Applied Psychology*, Vol. 71 No. 4, pp. 598-605.

Boyer, K.K. and Pagel, M. (2000), "Measurement issues in empirical research: improving measures of operations strategy and advanced manufacturing technology", *Journal of Operations Management*, Vol. 18 No. 3, pp. 361-74.

Carmines, E.G. and Zeller, R.A. (1990), *Reliability and Validity Assessment*, Sage, New York, NY.

Collins, R.S. and Cordon, C. (1997), "Survey methodology issues in manufacturing strategy and practice research", *International Journal of Operations & Production Management*, Vol. 17 No. 7, pp. 697-706.

Cronbach, L.J. (1951),"Coefficient Alpha and the internal structure of tests", *Psychometrika*, Vol. 16 No. 4, pp. 297-334.

Converse, J.M. and Presser, S. (1988), *Survey Questions. Handcrafting the Standardized Questionnaire*, Sage, New York, NY.

Dansereau, F. and Markham, S.E. (1997), "Level of analysis in personnel and human resources management", in Rowland, K. and Ferris, G. (Eds), *Research in Personnel and Human Resources Management*, Vol. 5, JAI Press, Greenwich, CT.

Dillman, D.A. (1978), *Mail and Telephone Survey: The Design Method*, John Wiley & Sons, New York, NY.

Dubin, R. (1978), *Theory Building*, The Free Press, New York, NY.

Emory, C.W. and Cooper, D.R. (1991), *Business Research Methods*, Irwin, Homewood, IL.

Filippini, R. (1997), "Operations management research: some reflections on evolution, models and empirical studies in OM", *International Journal of Operations & Production Management*, Vol. 17 No. 7, pp. 655-70.

Flynn, B.B., Schroeder, R.G. and Sakakibara, S. (1994), "A framework for quality management research and an associated measurement instrument", *Journal of Operations Management*, Vol. 11 No. 4, pp. 339-66.

Flynn, B.B., Sakakibara, S., Schroeder, R.G., Bates, K.A. and Flynn, E.J. (1990), "Empirical research methods in operations management", *Journal of Operations Management*, Vol. 9 No. 2, pp. 250-84.

Flynn, B.B., Schroeder, R.G., Flynn, E.J., Sakakibara, S. and Bates, K.A. (1997), "World-class manufacturing project: overview and selected results", *International Journal of Operations & Production Management*, Vol. 17 No. 7, pp. 671-85.

Forza, C. (1995), "Quality information systems and quality management: a reference model and associated measures for empirical research", *Industrial Management and Data Systems*, Vol. 95 No. 2, pp. 6-14.

Forza, C. and Di Nuzzo, F. (1998), "Meta-analysis applied to operations management: summarizing the results of empirical research", *International Journal of Production Research*, Vol. 36 No. 3, pp. 837-61.

Forza, C. and Vinelli, A. (1998), "On the contribution of survey research to the development of operations management theories", in Coughlan, P., Dromgoole, T. and Peppard, J. (Eds), *Operations Management: Future Issues and Competitive Responses*, School of Business Studies, Dublin, pp. 183-8.

Fowler, F.J. Jr (1993), *Survey Research Methods*, Sage Publications, New York, NY.

Greer, T.V., Chuchinprakarn, N. and Seshadri, S. (2000), "Likelihood of participating in mail survey research – business respondents' perspectives", *Industrial Marketing Management*, Vol. 29 No. 2, pp. 97-109.

Hair, J.F. Jr, Anderson, R.E., Tatham, R.L. and Black, W.C. (1992), *Multivariate Data Analysis*, Maxwell Macmillan, New York, NY.

Handfield, R.B. and Melnyk, S.A. (1998), "The scientific theory-building process: a primer using the case of TQM", *Journal of Operations Management*, Vol. 16 No. 4, pp. 321-39.

Hatcher, L. (1994), *A Step-by-step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, SAS-Institute incorporated, Cary, NC.

Hensley, R.L. (1999), "A review of operations management studies using scale development techniques", *Journal of Operations Management*, Vol. 17 No. 2, pp. 343-58.

Hinkin T.R. (1995), "A review of scale development practices in the study of organisations", *Journal of Management*, Vol. 21 No. 5, pp. 967-88.

Hollander, M. and Wolfe, D.A. (1999), *Nonparametric Statistical Methods*, 2nd ed., Wiley, New York, NY.

Horst, P. (1968), *Personality: Measurement of Dimensions*, Jossey-Bass, San Francisco, CA.

Karweit, N. and Meyers, E.D. Jr (1983), "Computers in survey research", in Rossi, P.H., Wright, J.D. and Anderson, A.B., *Handbook of Survey Research*, Academic Press, New York, NY, pp. 379-414.

Kerlinger, F.N. (1986), *Foundations of Behavioral Research*, 3rd ed., Harcourt Brace Jovanovich College Publishers, New York, NY.

Koufteros, X.A. (1999), "Testing a model of pull production: a paradigm for manufacturing research using structural equation modelling", *Journal of Operations Management*, Vol. 17 No. 4, pp. 467-88.

Lambert, D.M. and Harrington, T.C. (1990), "Measuring nonresponse bias in customer service mail surveys", *Journal of Business Logistics*, Vol. 11 No. 2, pp. 5-25.

Lawshe C.H. (1975), "A quantitative approach to content validity", *Personnel Psychology*, Vol. 28 No. 4, pp. 563-75.

Lazarsfeld, P.F. (1935), "The art of asking why", *National Marketing Research*, Vol. 1, pp. 26-38.

McClave, J.T. and Benson, P.G. (1991), *Statistics for Business and Economics*, Macmillan, New York, NY.

Malhotra, M.K. and Grover, V. (1998), "An assessment of survey research in POM: from constructs to theory", *Journal of Operations Management*, Vol. 16 No. 17, pp. 407-25.

Meredith, J.R. (1998), "Building operations management theory through case and field research", *Journal of Operations Management*, Vol. 16 No. 4, pp. 441-54.

Meredith, J.R., Raturi, A., Amoako-Jampah, K. and Kaplan, B. (1989), "Alternative research paradigms in operations", *Journal of Operations Management*, Vol. 8 No. 4, pp. 297-326.

Messick, S. (1995), "Validity of psychological assessment", *American Psychologist*, Vol. 50 No. 9, pp. 741-9.

Miller, D.C. (1991), *Handbook of Research Design and Social Measurement*, Sage Publications, London.

Nunnally, J.C. (1978), *Psychometric Theory*, 2nd ed., McGraw-Hill, New York, NY.

O'Leary-Kelly, S.W. and Vokurka, R.J. (1998), "The empirical assessment of construct validity", *Journal of Operations Management*, Vol. 16 No. 4, pp. 387-405.

Oppenheim, A.N. (1992), *Questionnaire Design, Interviewing and Attitude Measurement*, Pinter, New York, NY.

Pannirselvam, G.P., Ferguson, L.A., Ash, R.C. and Siferd, S.P. (1999), "Operations management research: an update for the 1990s", *Journal of Operations Management*, Vol. 18 No. 1, pp. 95-112.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1988), "SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality", *Journal of Retailing*, Vol. 64 No. 1, pp. 12-40.

Payne, S.L. (1951), *The Art of Asking Questions*, Princeton University Press, Princeton, NJ.

Pinsonneault, A. and Kraemer, K.L. (1993), "Survey research methodology in management information systems: an assessment", *Journal of Management Information Systems*, Vol. 10 No. 2, pp. 75-106.

Pitkow, J.E. and Recker, M.M. (1995), "Using the Web as a survey tool: results from the second WWW user survey", GVU User Surveys (On-line), Available at: http://www.cc.gatech.edu/gvu/user_surveys/User_Survey_Home.html

Rea, L.M. and Parker, R.A. (1992), *Designing and Conducting Survey Research*, Jossey-Bass, San Francisco, CA.

Robinson, W.S. (1950), "Ecological correlations and the behaviours of individuals", *American Sociological Review*, Vol. 15, June.

Rossi, P.H., Wright, J.D. and Anderson, A.B. (1983), *Handbook of Survey Research*, Academic Press, New York, NY.

Roth, P.L. and BeVier, C.A. (1998), "Response rates in HRM/OB survey research: norms and correlates, 1990-1994", *Journal of Management*, Vol. 24 No. 1, pp. 97-118.

Rungtusanatham, M.J. (1998), "Let's not overlook content validity", *Decision Line*, July, pp. 10-13.

Rungtusanatham, M.J. and Choi, T.Y. (2000), "The reliability and validity of measurement instrument employed in empirical OM research: concepts and definitions", Working Paper, Department of Management, Arizona State University.

Rungtusanatham, M.J., Choi, T.Y., Hollingworth, D.G. and Wu, Z. (2001), "Survey research in production/operations management: historical analysis and opportunities for improvement", Working Paper of Department of Management, Arizona State University, Tampa, AZ.

Saraph, J.V., Benson, P.G. and Schroeder, R.G. (1989), "An instrument for measuring the critical factors of quality management", *Decision Sciences*, Vol. 20 No. 4, pp. 810-29.

Scudder, G.D. and Hill, C.A. (1998), "A review and classification of empirical research in operations management", *Journal of Operations Management*, Vol. 16 No. 1, pp. 91-101.

Sekaran, U. (1992), *Research Methods for Business*, John Wiley & Sons, New York, NY.

Simon, H. (1980), "The behavioral and social sciences", *Science*, Vol. 209, pp. 72-8.

Sireci, S.G. (1998), "The construct of content validity", *Social Indicators Research*, Vol. 45, pp. 83-117.

Sudman, S. (1983), "Applied sampling", in Rossi, P.H., Wright, J.D. and Anderson, A.B., *Handbook of Survey Research*, Academic Press, New York, NY, pp. 144-94.

Swab, B. and Sitter, R. (1974), "Economic aspects of computer input-output equipment", in House, W.C. (Ed.), *Data Base Management*, Petrocelli Books, New York, NY.

Van Donselaar, K. and Sharman, G. (1997), "An innovative survey in the transportation and distribution sector", *International Journal of Operations & Production Management*, Vol. 17 No. 7, pp. 707-20.

Verma, R. and Goodale, J.C. (1995), "Statistical power in operations management research", *Journal of Operations Management*, Vol. 13 No. 2, pp. 139-52.

Wacker, J.G. (1998), "A definition of theory: research guidelines for different theory-building research methods in operations management", *Journal of Operations Management*, Vol. 16 No. 4, pp. 361-85.

Ward, P.T., Leong, G.K. and Boyer, K.K. (1994), "Manufacturing proactiveness and performance", *Decision Sciences*, Vol. 25 No. 3, pp. 337-58.

Wilson, E.J. (1999), "Research practice in business marketing – a comment on response rate and response bias", *Industrial Marketing Management*, Vol. 28 No. 3, pp. 257-60.

Whybark, D.C. (1997), "GMRG survey research in operations management", *International Journal of Operations & Production Management*, Vol. 17 No. 7, pp. 686-96.

Further reading

Baroudi, J.J. and Orlikowski, W.J. (1989), "The problem of statistical power in MIS research", *MIS Quarterly*, Vol. 13 No. 1, pp. 87-106.

Peter, J.P. (1979), "Reliability: a review of psychometric basics and recent marketing practices", *Journal of Marketing Research*, Vol. 16 No. 1, pp. 6-17.

Peter, J.P. (1981), "Construct validity: a review of basic issues and marketing practices", *Journal of Marketing Research*, Vol. 18 No. 2, pp. 133-45.

Sharma, S. (1996), *Applied Multivariate Techniques*, Wiley, New York, NY.

Straub, D.W. (1989), "Validating instruments in MIS research", *MIS Quarterly*, Vol. 13 No. 2, pp. 147-69.