CHAPTER 6

RESEARCH METHODOLOGY

6.1 INTRODUCTION

Research hypotheses for the study were formulated and discussed in Chapter 5. This chapter will provide a theoretical perspective on the research process as well as show how the theory will be applied practically in order to conduct the research component of the study.

Detail will be provided on the data collection methods, sampling process and questionnaire design. The chapter will be concluded with a discussion regarding the statistical procedures and techniques that will be adopted for the study.

6.2 INFORMATION SOURCES

A distinction is made between two sources of data, namely primary and secondary data (Sudman & Blair, 1998: 74). Secondary data refers to information that has already been collected for some other purpose, while primary data refers to information that has been gathered to address the research objectives at hand (Burns & Bush, 1998: 66).

Sources of secondary information include libraries, Internet searches, trade associations and Online services, while primary information is gathered by means of observation, survey interviews, group discussions or experiments (Sudman & Blair, 1998: 74).

Both primary and secondary data sources will be used for the study. Secondary sources were used to form the theoretical foundation of the study and will also be used to define and structure the methods as well as process that should be followed to conduct primary research. The remainder of this chapter will focus on the primary research-phase of the study.

6.3. DATA COLLECTION METHODS

It should be noted, before discussing various data collection methods, that the means of data collection during the research process can be classified into three broad categories: quantitative, qualitative and pluralistic (Burns & Bush, 1998: 209).

Quantitative research can be defined as research involving the use of structured questions with predetermined response options, with a large number of respondents involved. In contrast, **qualitative research** involves collecting of data and the analysis and interpretation thereof through observing what people do or say.

Pluralistic research can be defined as a combination of qualitative and quantitative research methods in an attempt to gain the advantages of the two methods. With pluralistic research, a researcher typically will begin with exploratory qualitative techniques (for example focus groups). The qualitative phase of the research project will then serve as a foundation for the quantitative phase of the project, since it provides first-hand knowledge of the research problem to the researcher.

For this study, the pluralistic approach will be followed by first using qualitative research techniques, followed by quantitative techniques.

Primary research data can be gathered in a number of ways, including observation, surveys, depth interviews, panels of research participants (focus groups), experiments and test markets (Sudman & Blair, 1998: 90). These methods will not be discussed in any detail, with the exception of focus groups (pre-test group) and surveys, since the primary research data for the study will be collected using these methods.

6.3.1 Pre-testing

As a qualitative research technique, it was decided to use a pre-test group, where the results would be used as input to the quantitative-phase of the research project.

It was decided to use a pre-test group to identify characteristics and underlying issues with regards to South African Internet users' Online behaviour. It was envisaged that by doing so higher quality input to the remainder of the research process would be ensured.

A single pre-test group session was held with 94 South African Internet users. The Internet users gathered at a movie theatre and were requested to complete the first draft of the proposed research questionnaire (compiled from viewing research reports and findings from South African research organisations as well as insights gained through the literature review) that would be used during the quantitative-component of the study. It should be mentioned that the respondents gathered at the movie theatre to participate in another research project (also Internet-related), where after they participated in the pre-test group session. Respondents were, following the pre-test session, treated to a free movie sponsored by ISP "X" (who conducted the first study).

The following objectives were set as required outputs from the pre-test session:

- to test the questionnaire to be used during the quantitative phase of the study (surveys);

- to ensure that all the product categories that respondents could have purchased or intend purchasing via the Internet in the future were covered in the questionnaire; and

- to refine the hypotheses set for the study.

The objectives set for the pre-test session were achieved when the results from the study were scrutinised. The major findings (and corrective actions taken as a result of the pre-test group) that were used to improve the quality of the remainder of the research project were:

- 34% of respondents who have purchased via the Internet previously indicated (when requested to select from which product or service categories they have purchased before) that they have purchased "other" products or services (other than the listed categories). This led to more research reports searched for to identify other possible product and services categories Internet users purchase from via the Internet. The researcher also "surfed" the Net and more specifically Web-sites of South African Internet sellers of products and services to identify more categories;

- based on the search, the listed product and service categories that will be used in the final questionnaire were increased to 38 (as opposed to the 21 categories used in the pre-test group);

- due to the importance of banking services, it was decided to list banking services as a separate category (in addition to the 38 listed categories);

- questions that could have been misunderstood or misinterpreted were refined and rephrased to provide commonly understood questions in the final questionnaire; and

- insights gained from the pre-test group assisted in refining the hypotheses set for the study (discussed in Chapter 5).

The main findings from the pre-test group are summarised in Appendix 6. The remainder of the section will focus on the quantitative phase of the research project, namely surveys.

6.3.2 Surveys

Three broad methods of data collection by means of surveys are distinguished, namely telephone interviews, personal interviews and mail interviews (Malhotra, 1996: 198). Cooper & Schindler (2001: 321) add to the classification by listing self-administered surveys as a final survey method as opposed to mail interviews. These authors do, however, include mail surveys as a component of self-administered surveys but also include other methods of questionnaire distribution (for example fax and the Internet). The three methods will be compared in Table 6.1 below by providing a short description of each method and showing the advantages and disadvantages thereof.

TABLE 6.1: SURVEY DATA COLLECTION METHODS

Personal interview	Telephone interview	Self-administered surveys
	Description	
People selected to be part of the sample are interviewed in person by a trained interviewer	People selected to be a part of the sample are interviewed on the telephone by a trained interviewer	Questionnaires are: a) Mailed, faxed, or couriered to be self-administered - with return mechanism generally included; b) Computer-delivered via the intranet, Internet and Online services – computer stores/forwards completed instruments automatically; c) People intercepted/studied via paper or computerised instrument in central location – without interviewer assistance.

Personal interview	Telephone interview	Self-administered surveys
Advantages		
• Good co-operation from respondents • Interviewer can answer questions about survey, probe for answers, use follow-up questions and gather information by observation • Special visual aids and scoring devices can be used • Illiterate and functionally illiterate respondents can be reached • Interviewer can pre-screen respondent to ensure he/she fits the population profile • CAPI – computer-assisted personal interviewing: Responses can be entered into a portable microcomputer to reduce error and cost	• Lower costs than personal interview • Expanded geographic coverage without drastic increase in costs • Uses fewer, more highly skilled interviewers • Reduced interviewer bias • Faster completion times • Better Access to hard-to-reach respondents through repeated call-backs • Can use computerised random-digit dialing • CATI – computer-assisted telephone interviewing: Responses can be entered directly into a computer file to reduce error and cost	• Allows contact with otherwise inaccessible respondents (e.g. CEOs) • Incentives may be used to increase response rate • Often low-cost option • Expanded geographic coverage without increase in cost (a) • Requires minimal staff (a) • Perceived as anonymous (a) • Allows respondents time to think about questions (a) • More complex instruments can be used (b) • Fast access to computer literate (b) • Rapid data collection (b,c) • Respondent who cannot be reached by phone (voice) may be accessible (b,c) • Sample frame lists viable locations rather than prospective respondents (b,c) • Visuals may be used (b,c)
Disadvantages		
• High costs • Need for highly trained interviewers • Long period needed in the field collecting data • May be wide geographical dispersion • Follow-up is labour intensive • Not all respondents are available or accessible • Some respondents are unwilling to talk to strangers in their homes • Some neighbourhoods are difficult to visit • Questions may be altered or respondent coached by interviewers	• Response rate is lower than for personal interview • Higher costs if interviewing geographically dispersed sample • Interview length must be limited • Many phone numbers are unlisted or not working, making directory listings unreliable • Some target groups are not available by phone • Responses may be less complete • Illustrations cannot be used	• Low response rates in some modes • No interviewer intervention available for probing or explanation (a) • Cannot be long or complex (a) • Accurate mailing list needed (a) • Often respondents returning survey represent extremes of the population – skewed responses (a) • Anxiety among some respondents (b) • Directions/software instruction for progression through the instrument (b) • Computer security (b) • Need for low-distraction environment for survey completion (c)

Source: Adapted from Cooper & Schindler (2001: 313)

It is important to briefly discuss the deciding factors when choosing between the various survey options before concluding the discussion on data collection methods. Sudman & Blair (1998: 155) provide a summary of the considerations (shown in Table 6.2) when choosing a survey method.

TABLE 6.2: CONSIDERATIONS WHEN CHOOSING A SURVEY METHOD

Consideration	Questions to be Answered
Questionnaire Design	What constraints does this method impose on the questionnaire? Does it allow us to do what we want?
Sampling	What constraints does this method impose on getting a sample of respondents? Will it allow us to get a sample of adequate quality?
Response Quality	Will any characteristic of this method discourage accurate answers?
Time and Cost	How long will it take to complete? How much will it cost?

Source: Sudman & Blair (1998: 155)

A number of additional criteria can be listed that should be considered when deciding between the different survey methods. These criteria include: diversity of the questions, use of physical stimuli, control of the data collection environment, control of the field force, quantity of the data and potential for interviewer bias (Malhotra, 1996: 205-211). In addition to these criteria, Burns & Bush (1998: 270) also list (amongst others): respondent characteristics - incidence rate, willingness to participate, ability to participate and diversity of respondents.

Although many other factors can also be listed as criteria when choosing between survey methods, Boyd, Westfall & Stasch (1989: 249) are of the opinion that if all the survey methods will produce the same results when data is collected, the decision can be made on the basis of speed or cost.

After pondering on the different methods of data collection and considerations when deciding which survey method to use, it was decided to use self-administered surveys by using the Internet. The decision was further supported by considering the following statement by Forrest (1999: 10), explaining why the Internet is regarded as a powerful research tool: "., the advantages of the Internet as a survey medium when compared to existing methods are many, especially in terms of cost, speed of distribution and the ability of researchers to gain access to real-time information".

Forrest (1999: 139) does, however, also list a major concern that should be kept in mind when considering to use Internet surveys, namely that the Internet population is not representative of the general population. Although this concern is noted, it will be ignored since the population (discussed when the sampling process is described) used in the study is Internet users and not the South African population as a whole.

Two self-administered survey options were viewed as most probably the best methods that could be used to conduct the research project by using the Internet, namely: distribution of questionnaires through e-mail for completion by the respondent and returned to the researcher by e-mail and Internet-based research conducted on a live Web-site.

After considering the two options, it was decided to use the live Web-based option for the following reasons: Firstly, the respondents will be familiar with the Web-environment since they will all be registered Internet users and the research project is specifically to determine buying behaviour of Internet users (to a great extent implying Internet-based purchases). Secondly, the researcher had a concern that respondents may have been reluctant to open an attachment to an e-mail from a person or organisation they did not know.

The e-mail with attachment option also may have proved too cumbersome for respondents since they would have had to save the questionnaire once completed to their computer's hard drive before returning it. If the completed questionnaire was not saved, "empty" questionnaires would have been received from respondents.

Based on this reasoning, it was decided to use the live Web-based self-administered survey option (attached in Appendix 9).

6.4 SAMPLING

A sample represents a limited number taken from a large group for testing and analysis and is based on the assumption that the sample can be treated as representative of the group (Crouch & Housden, 1996: 115). In basic terms, some elements in a population provide useful information of the entire population (Emory & Cooper, 1991: 242).

Before providing more detail on the terms associated with sampling (which should be discussed to provide clarity on the sampling process), it is important to firstly explain why samples are used in research and secondly to provide characteristics of "good" samples.

Two general reasons for the use of sampling can be distinguished, namely practical considerations such as cost and the size of the population and secondly, ability to analyse the data (Burns & Bush, 1998: 361 – 362). Cooper & Schindler (2001: 163) broaden this view by identifying four reasons for sampling, namely lower cost, greater speed of data collection, greater accuracy of results and availability of population elements.

The relation between **cost and speed** of data collection is supported by Crouch & Housden (1996: 116), indicating that the smaller the number of people from which data has to be collected, the cheaper and quicker it will be.

Greater **accuracy of results** is often assured, since fewer field-workers, who have to be trained and supervised, are needed. Sampling also allows for more thorough investigation of missing, wrong or suspicious information (Cooper & Schindler, 2001: 164 and Boyd et al., 1989: 358).

Finally, the **availability of population elements** is often a reason for sampling. This reason is typically listed when, for example, the breaking strength of materials has to be tested. The only way of ensuring that not all material is destroyed in testing, is by using a sample of the materials. This reason will also be prominent when the population is infinite (Cooper & Schindler, 2001: 164).

Considering the reasons stated above to explain why sampling is done, the question may arise as to what can be considered to be a good sample. Lohr (1999: 3) proposes that a perfect sample would be: "... a scaled-down version of the population, mirroring every characteristic of the whole population". Lohr (1999: 3) continues by stating that a perfect sample does, however, not exist for complicated populations and even if it did exist, there is no way of knowing that it is a perfect sample without measuring the whole population.

A good sample can therefore be regarded as one that reproduces the characteristics of interest in the population as closely as possible. It will be representative in the sense that each sample unit will represent the characteristics of a known number of units in the population (Lohr, 1999: 3).

Before continuing the discussion on sampling, the terminology associated with sampling will first be clarified. The following terms will be discussed: target population, sampled population, observation unit, sample unit, sample frame and the sample frame error.

The **target population** (also referred to as the **population**) is the collection of elements or objects (possessing the sought information) that the researcher intends to study and about which inferences are to be drawn (Malhotra, 1996: 360). Siegel (1994: 252) defines the population perhaps a bit more simply: "The **population** is the collection of units (people, objects, or whatever) that you are interested in knowing about". The **sampled population** represents the actual population from which the sample has been taken (Keller & Warrack, 2000: 153 and Lohr, 1999: 3)

The **observation unit** represents an object on which a measurement is taken and forms the basic unit of observation (also referred to as an **element**)(Lohr, 1999: 3). In simple terms, the **element** is the object from which or about which information is desired (Malhotra, 1996: 361).

A **sample unit** represents the unit (or a unit containing the element) that is actually sampled or is available for selection at some stage of the sampling process (Lohr, 1999: 3 and Malhotra, 1996: 361).

Finally, the **sample frame** refers to a list of all sample units in the population (Lohr, 1999: 3 and Burns & Bush, 1998: 360). For example, when conducting a telephone survey, the sampling frame could include all the telephone numbers in a particular city. It should be noted that a sample frame could contain a **sample frame error**, defined as the degree to which the sample frame fails to account for all of the population (Burns & Bush, 1998: 361).

As mentioned above, errors do occur with sampling. Two other errors associated with sampling that should briefly be discussed are sampling and non-sampling errors. **A sampling error** refers to any error in a survey that occurs because a sample is used. A sampling error is caused by two factors, namely the method of sample selection and the size of the sample (Burns & Bush, 1998: 360).

A **non-sampling error** is regarded as being more serious than a sampling error (Keller & Warrack, 2000: 160), since a larger sample will not diminish the size of the error or even the possibility of it occurring. Thompson (1997: 7) adds to this view by stating that the extent of sampling errors is much easier to estimate than the extent of non-sampling errors.

Non-sampling errors occur due to mistakes made in the acquisition of the data or through the improper selection of sample observations. Three types of non-sampling errors can be distinguished, namely errors in data acquisition (types of errors arising from the recording of incorrect responses), non-response errors or bias (occurs when responses are not obtained from some members of the sample) and selection bias (when the sampling plan is such that some members of the target population cannot possibly be selected for inclusion in the sample)(Keller & Warrack, 2000: 160).

The discussion above should provide sufficient information to form a general view of what sampling is and the terminology associated with sampling. The remainder of the section will provide details on the sampling process followed for the study. The discussion will be structured according to a sampling design process suggested by Malhotra (1996: 360). The process comprises 5 steps, namely:

Step 1: Define the population (Section 6.4.1)

Step 2: Determine the sampling frame (Section 6.4.2)

Step 3: Select the sampling techniques (methods)(Section 6.4.3)

Step 4: Determine the sample size (Section 6.4.3)

Step 5: Execute the sampling process (selection of the sample units)(Section 6.4.4)

Steps 3 and 4 will be combined for the purpose of the discussion.

6.4.1 Define the population

From the discussion in Section 1.2 (in Chapter 1) it could be seen that the population for this study was defined as **all Internet users in South Africa accessing the Internet from home.**

Table 1.4 (in Chapter 1) provided details regarding the number of South African Internet users by access method. From Table 1.4 it could be seen that the Internet users who access the Internet via dial-up modem amount to 782 000 (end of 2000).

For the purpose of this study, only the South African Internet users accessing the Net from home (782 000 users at the end of 2000) will be included to represent the population for the study. It is implied that users from all of the geographic regions (provinces) will be included in the sample.

Since this study will focus on consumers, it was decided to only focus on users accessing the Net from home (based on the assumption that consumers would access the Net from home, rather that from academic institutions or corporate networks).

6.4.2 Determine the sampling frame

South African Internet users accessing the Internet by means of dial-up modems gain access to the Internet through Internet Service Providers (ISPs) to which they subscribe. A number of different ISPs offer dial-up access in South Africa, including ABSA Internet, M-Web, Telkom Internet (incorporating Intekom) and World-Online. Detailed information regarding the number of subscribers to each Internet Service Provider could not be obtained. A possible reason for the lack of this information can be attributed to the competitive nature thereof to Internet Service Providers.

The researcher negotiated with a leading South African ISP to use their Internet user-database to conduct the study. **For confidentiality reasons (as requested by the ISP), the ISP will remain anonymous and will be referred to in the study as ISP "X".** Important to note is that ISP "X" provides two different dial-up access services, each trading under separate names. For the purpose of the study, the first service (and users/database thereof) will be referred to as ISP "X" (a) and the second (and users/database thereof) as ISP "X" (b). **The sampling frame for the study is, therefore, all Internet users who subscribe to ISP "X".**

6.4.3 Selecting the sampling method and size of the sample

Sampling methods can be broadly classified as non-probability and probability. Non-probability sampling relies on the personal judgement of the researcher rather than on chance to select elements, whereas with probability sampling methods the sample units are selected by chance (each element of the population has a fixed probabilistic chance of being selected for the sample (Cooper & Schindler, 1998: 218 and Malhotra, 1996: 365).

A number of different probability and non-probability sampling methods are distinguished. Although some authors differ as far as the actual terminology used for the various methods, there is general consensus regarding the major methods (Cooper & Schindler, 2001: 189 & 190; Lohr, 1999: 24; Burns & Bush, 1998: 363 & 376 and Sudman & Blair, 1998: 349 - 349). The most commonly used non-probability sampling methods are: convenience sampling, judgement sampling, referral sampling and quota sampling. The predominant probability sampling methods are: simple random sampling, systematic sampling, cluster sampling and stratified sampling.

For the purpose of the research study, probability sampling will be used. For this reason, only the four probability sampling methods identified above will receive further attention. These methods are shown in Table 6.3 below, providing a short description and the advantages and disadvantages of each method.

TABLE 6.3: PROBABILITY SAMPLING METHODS

Sampling method	Description	Advantages	Disadvantages
Simple Random	Each population element has an equal chance of being selected into the sample. Sample drawn using random number table/generator	* Easy to implement with automatic dialling (random digit dialling) and with computerised voice response systems (if telephone or fax is used)	* Requires a listing of population elements * Takes more time to implement * Uses larger sample sizes * Produces larger errors * Expensive
Systematic	Selects an element of the population at a beginning with a random start and following the sample fraction selects every k^{th} element	* Simple to design * Easier to use than simple random * Easy to determine sample distribution of mean or proportion * Less expensive than simple random	* Periodicity within the population may skew the sample and results * If the population list has a monotonic trend, a biased estimate will result based on the start point
Stratified	Divides population into sub-populations or strata and uses simple random on each strata. Results may be weighted and combined	* Researcher controls sample size and strata * Increased statistical efficiency * Provides data to represent and analyse subgroups * Enables use of different methods in strata	* Increased error will result if subgroups are selected at different rates * Expensive * Especially expensive if strata on the population have to be created
Cluster	Population is divided into internally heterogeneous sub-groups. Some are randomly selected for further study	* Provides an unbiased estimate of population parameters if properly done * Economically more efficient than simple random * Lowest cost per sample, especially with geographic clusters * Easy to do without a population list	* Often lower statistical efficiency (more error) due to subgroups being homogeneous rather than heterogeneous

Source: Adapted from Cooper & Schindler (2001: 190)

Couper (2000: 485) lists a specific (probability) Web-based sampling method, namely list-based samples. The basic approach of this type of Web survey is to begin with a frame of those with access to the Web (refer to Section 6.4.2 where it was noted that the frame for this study is specifically Internet users). Couper (2000: 485) explains that invitations are sent by e-mail to participate and access is controlled to prevent multiple completions by the same respondents (Section 6.7 details the interview procedure, commencing with an e-mail message sent to Internet users comprising the frame). A main concern when using this survey method is a non-response error. A non-response error arises through the fact that not all people included in the sample are willing or able to complete the survey (Couper, 2000: 473) and is, therefore, a function of the rate of non-response and the difference between respondents and non-respondents.

After careful consideration of the different probability sampling methods shown in Table 6.3 and paying attention to the advantages and disadvantages of each method (with specific consideration of the objectives and hypotheses formulated for the study), it was decided to use the **stratified method of sampling**. The reason why the objectives and hypotheses set for the study played such an important role in deciding which sampling method to use can be justified - the length of time being an Internet user is of critical importance to the success of the study since most of the hypotheses centre around identifying possible relationships with the period being an Internet user.

In an attempt to ensure that this critical element to the study is met, different periods of Internet usage for Internet users will first have to be identified (creating various strata – refer to Table 6.4). Once the periods have been identified, it will be possible to use simple random sampling to complete the sampling process. It can be derived from the discussion that the stratified method of sampling is best suited to meet the requirements to conduct the study.

Since ISP “X” provided the information of all their Internet users, it was decided to use 20 000 Internet users in the sample.

6.4.4 Selection of the sample units

The selection of the sample units was conducted with a specific requirement in mind, namely that provision had to be made for the period of Internet usage. With this requirement in mind, the sample was drawn from the ISP "X" database as follows:

Step 1: Setting the parameters

A) The following dates were specified as the periods wherein an Internet user had to have joined ISP "X" (a):

  a) between 1 January 1997 and 31 December 1999; and

  b) between 1 January 2001 and 25 November 2001 - the date on which the database was drawn. [It should be noted that in an attempt to ensure an even distribution, across all time periods, it was decided not to consider users who joined ISP "X" (a) in 2000 since ISP "X" (b) only became operational in 2000. Selecting more users from ISP "X" (a) who joined during 2000 could possibly skew the anticipated responses across the period of Internet usage when both user groups are considered.]

B) The following dates were specified as the periods wherein an Internet user had to have joined ISP "X" (b) as their ISP:

  a) between 1 January 2000 and 31 December 2000 [it should be noted that ISP "X" (b) was only launched in 2000 and, therefore, had no users prior to 1 January 2000]; and

  b) between 1 January 2001 and 25 November 2001 (date on which the database was drawn)

Step 2: Simple random sampling

Simple random sampling was used to draw 7 000 subscribers for each period specified above. The sample units were randomly selected, following a ratio between the number of Internet users in each time period and the total number required for each period (7 000). A total of 14 000 ISP “X” (a) and 14 000 ISP “X” (b) Internet subscribers were selected.

The reason for selecting 7 000 users per time period can be attributed to many records on the database lacking a primary e-mail address (possibly because subscribers use the ISP for Internet access only - if e-mail is required, subscribers can obtain free e-mail on the Net at, for example, HotMail or Yahoo Mail). The advantage of the latter is that even if the subscriber changes his/her ISP, the e-mail address will remain the same. The extra 2 000 users per period was thought to be sufficient to, once removed, leave 5 000 Internet users per specified period.

Step 3: Final selection

All data records of Internet users without a primary e-mail address on the ISP “X” (a) and ISP “X” (b) databases were removed, leaving 5 000 users per period (in cases where there were in excess of 5 000 users, all records from data entry 5 001 onwards were removed). The sample size that will be used in the study is summarised in Table 6.4 below:

TABLE 6.4: DETAILS OF STRATA FOR THE SAMPLE USED

ISP “X” (b): Period joined ISP		ISP “X” (a): Period joined ISP		Total
1 January 2000 to 31 December 2000	1 January 2001 to 25 November 2001	1 January 1997 to 31 December 1999	1 January 2001 to 25 November 2001	
5 000 Internet users	**5 000 Internet users**	**5 000 Internet users**	**5 000 Internet users**	**20 000 users**

6.5 MEASUREMENT AND MEASUREMENT SCALES

Measurement and measurement scales available to the researcher need to be explained before attention is focused on questionnaire design.

Measurement, in a research context, consists of assigning numbers to empirical events in compliance with a set of rules (Cooper & Schindler, 2001: 203). A question that may arise from the statement, is: “what is being measured?” Cooper & Schindler (2001: 204) and Burns & Bush (1998: 289 – 290) explain that the properties of objects are being measured. Objects include, for example, consumers, advertisements and books, whereas properties are the specific characteristics or features of an object that can be used to distinguish between objects.

Of importance when studying measurement and scaling, is to understand scale characteristics since the characteristics possessed by a scale determine the level of measurement thereof (Burns & Bush, 1998: 292). Cooper & Schindler (2001: 204 – 205) and Burns & Bush (1998: 291) identify four scale characteristics, namely description (Cooper & Schindler, 2001: 204 refer to classification), order, distance and origin. These four characteristics will briefly be discussed below.

Description refers to the use of a label or unique descriptor to stand for each designation in the scale, for example "yes" and "no" or "agree" and "disagree". Burns & Bush (1998: 291) note that all scales include description in the form of characteristic labels that identify what is being measured.

The relative sizes of descriptors are referred to as **order**. Key to the explanation is the term "*relative*" and includes descriptors such as "greater than", "less than" or "equal to". For example, a consumer's least-preferred brand is "less than" the most-preferred brand. Not all scales possess order characteristics, as opposed to description. For example, is a "buyer" greater than a "non-buyer"? One cannot establish a *relative* size distinction.

The third characteristic, **distance**, is when the absolute difference between descriptors is known (and may be expressed in units). Order is also given to a scale if distance exists. For example, a consumer purchasing three bottles of beer purchases two bottles more than the consumer only purchasing one bottle. Order can be identified since it is clear that the three-bottle-buyer purchases "more than" the one-bottle-buyer and distance can be determined between the two buyers (two bottles).

The final characteristic is **origin**, implying that the scale has a unique beginning or true zero point. A zero is therefore the origin, for example, for the number of bottles of beer consumed. Not all scales used in research are characterised by origin, for example "do you agree or disagree with the following statement". The researcher can therefore not say, for that specific scale, that the respondent to the question has a true zero level of agreement.

The discussion above should provide sufficient information to form an understanding of the importance of measurement and characteristics of scales in the research process. The characteristics of scales will be considered when deciding which scales to use when designing the questionnaire. The four levels of measurement (measurement scale types) are discussed below.

6.5.1 Measurement scale types

The four measurement scale types that will now be briefly discussed are nominal scales, ordinal scales, interval scales and ratio scales. Showing the characteristics of measurement that each type of scale possesses will conclude the section.

Nominal measurement scales are used when objects are assigned to mutually exclusive, labelled categories, with no necessary relationship between the categories (Aaker & Day, 1990: 273). Nominal scales therefore only use labels and contain only the characteristic of description, for example race, gender and answers that involve "yes/no" or "agree/disagree" options.

Ordinal measurement scales are obtained by ranking objects in order with regard to some common variable (Aaker & Day, 1990: 273). Burns & Bush (1998: 293) add that ordinal scales indicate only relative size differences among objects. This type of scale has description and order characteristics but does not show how far apart the descriptors are since it does not show distance or origin.

For example, a respondent indicating "purchase less than once a week" when asked to indicate how frequently a product is purchased.

Scales where the distance between each descriptor is known are referred to as **interval scales** (Burns & Bush, 1998: 293). The distance characteristic is usually defined as one scale unit. For example, rating the taste of coffee as a "3" value is one unit away from a "4"-rating.

The final measurement scaling method is **ratio scales**. Aaker & Day (1990: 274) explain that ratio scales are a special kind of scale that has a meaningful zero point. For example, when asking a respondent: "how many bottles of beer have you purchased?" One respondent may have purchased twice as many bottles as another. The ratio explained in the example above could not have applied to interval scales, since it is not possible to conclude that one coffee brand is ¼ better than another brand (example used above to illustrate interval scales).

Table 6.5 provides a summary of the four different measurement scale types together with the characteristics of measurement associated with each type.

TABLE 6.5: CHARACTERISTICS OF DIFFERENT MEASUREMENT SCALES

	Scale characteristics possessed			
Level of measure	Description	Order	Distance	Origin
Nominal scale	Yes	No	No	No
Ordinal scale	Yes	Yes	No	No
Interval scale	Yes	Yes	Yes	No
Ratio scale	Yes	Yes	Yes	Yes

Source: Burns & Bush (1998: 292)

The classification and characteristics of different measurement scales were used when deciding which scale types to consider when designing the research questionnaire. Although various different scale types can be used when designing a questionnaire, only the rating scales used to develop the questionnaire used in the study will be briefly discussed. The discussion is based on the guidance provided by Cooper & Schindler (2001: 231 – 235).

The **simple category scale,** also called a dichotomous scale, offers two mutually exclusive response choices and is particularly useful where a dichotomous response is adequate. For example, the answer to a research question is either "yes" or "no". This scale type produces nominal data.

The **multiple choice, single-response scale** is used where multiple options are provided but only a single answer is sought. For example, respondents are requested to indicate their population group when the categories are provided: "white", "black" and "coloured". Only a single answer is sought from multiple options. This type of scale produces nominal data.

A variation to the above is the **multiple choice, multiple-response scale**, also referred to as a checklist, where the respondent can either choose one or several options provided. For example, a respondent is required to indicate which magazines are read at home, where multiple magazine titles are listed. This type of scale produces nominal data.

The **Likert scale** is a variation of the summated rating scale, consisting of statements that express either a favourable or unfavourable attitude towards an object. With the Likert scale, a respondent is asked to agree or disagree with the statements provided. Each response is given a numerical score, reflecting its degree of attitudinal favourableness. For example, respondents are requested to indicate the degree to which they agree or disagree with a specific statement (or a number of statements), where the options are "totally agree"; "agree"; "neither agree nor disagree"; "disagree" and "totally disagree". The Likert scale produces interval data.

The questionnaire design will be discussed in the following section, considering the different measurement scales available to the researcher.

6.6 QUESTIONNAIRE AND WEB-SITE DESIGN

The questionnaire used in the study was compiled after considering a number of factors that specifically influenced the design thereof. The major considerations were: the objectives and hypotheses set for the study, the survey method used, measurement scales selected for data collection, actual data capturing method and results obtained from the pre-test group. A copy of the questionnaire is attached (Appendix 9) together with a copy of the e-mail message that was sent to the sample population (Appendix 8), inviting them to respond to the questionnaire.

The section will provide details regarding the structure of the questionnaire, Web-site design, mapping of questions to address hypotheses set for the study, measurement scales used and rationale for questions used.

6.6.1 Questionnaire structure

The questionnaire was divided into four different sections, namely:

Section A: Classification questions;

Section B: Non-internet shoppers: factors considered when deciding whether or not to purchase via the Internet; consideration to purchasing via the Internet in the future; and tendency to search for information on the Internet prior to purchasing from "traditional", non-Internet-based sellers;

Section C: Internet shoppers: factors considered when deciding whether or not to purchase via the Internet; current product and services categories purchasing from via the Internet (and future intentions to purchase); and tendency to search for information on the Internet prior to purchasing from "traditional", non-Internet-based sellers;

Section D: Demographic information

The four sections with the relevant questions for each section will be discussed briefly in tabular form below, indicating the question (without the options that the respondents could choose), the variable (V) number applicable to each question and the scale type used.

It is important to note that the Web-site was designed in such a way that respondents would only complete questions according to their answers provided for the preceding question. For example, if the respondent answered "Yes" to question 7, the Web-site will automatically route the respondent to question 9. If the respondent answered "No" to question 7, the respondent will automatically be routed to question 8. The design of the questionnaire branching (attached in Appendix 7) is discussed in more detail in Section 6.6.2 (Web-site design).

Section A: Classification questions

Section A will attempt to classify the respondents according to the length of Internet access, determine from where they access the Internet, consider their Internet banking activities and determine whether or not they have purchased via the Internet before.

TABLE 6.6: QUESTIONS APPLICABLE TO SECTION A

Question	Variable	Scale type
1. From where do you access the Internet? (Multiple Answers)	V1 – V6	Multiple choice, multiple-response scale
2. From where do you most frequently access the Internet? (One answer only)	V7	Multiple choice, single-response scale
3. For how long have you been an Internet User? (Considering all the Internet Service Providers you have subscribed to)	V8	Multiple choice, single-response scale
4. How many Internet Service Providers have you subscribed to in the past? 5a. For how long are you subscribed to your current Internet Service Provider?	V9 V10	Multiple choice, single-response scale Multiple choice, single-response scale
5b. Do you subscribe to more than one Internet Service Provider?	V11	Multiple choice, single-response scale
6. Please indicate the extent to which you agree or disagree with each of the statements listed below: [5 statements listed]	V12 – V16	7-point Likert scale
7. Do you use Internet banking?	V17	Simple category scale

Question	Variable	Scale type
8. Are you considering using Internet banking facilities in the future?	V18	Simple category scale
9. For how long have you been using Internet banking?	V19	Multiple choice, single-response scale
10. How frequently do you/do you think you will use Internet banking?	V20	Multiple choice, single-response scale
11. Have you ever purchased products or services via the Internet before? (excluding Banking Services)	V21	Simple category scale

It is important to note that, as can be seen from Question 11, banking services has been excluded as a product or service category option when classifying whether or not a respondent has purchased via the Internet before. The reason for keeping banking services as a separate category is based on the results from the pre-test group and the concern that by including banking services the findings of the research project may be skewed.

The pre-test group showed that 34% of respondents indicated that "other" products or services were purchased, which possibly could have included a large percentage of Internet banking users. The researcher also considered a respondent who used Internet banking as not necessarily an Internet shopper, since the respondent could perhaps use Internet banking purely for convenience purposes.

In an attempt to determine which Internet users "truly" purchase products and services via the Internet, banking services were separated and will be analysed separately due to the importance of this category.

Section B: Non-internet shoppers

Respondents who indicate that they have not purchased products or services via the Internet before (question 11), will automatically be routed to Section B, applicable to non-Internet shoppers only.

Section B will consider which factors are considered by non-Internet shoppers when deciding whether or not to purchase via the Internet, determine whether or not they consider purchasing via the Internet in the future, ascertain which product and service categories they consider purchasing from and determine whether or not they use the Internet to search for information before they purchase from "traditional", non-Internet based sellers (so-called brick and mortar organisations).

TABLE 6.7: QUESTIONS APPLICABLE TO NON-INTERNET SHOPPERS (SECTION B)

Question	Variable	Scale type
12. Please indicate how important the factors listed below are to you when deciding whether or not to purchase via the Internet: [24 factors listed]	V22 – V45	7-point Likert scale
13. Do you consider purchasing products and/or services via the Internet in the future?	V46	Simple category scale
14. Would you consider to purchase via the Internet if more established, non-Internet based, South African businesses also offer products and services on the Internet? (e.g. Game Stores, OUTsurance, Musica)	V47	Simple category scale
15. From which of the following product and services categories will you seriously consider purchasing via the Internet in the future? (Multiple answers) [38 product and services categories listed]	V48 – V85	Multiple choice, multiple-response scale
16. Have you ever searched for or do you consider searching for product or service information on the Internet prior to purchasing from a non-Internet based seller? (e.g. A physical store or telephone shopping)	V86	Simple category scale
17. From which of the following product and service categories have you searched for or do you consider searching for information on the Internet prior to purchasing from a non-Internet based seller? (e.g. Physical store or telephone shopping) (Multiple answers) [38 product and service categories listed]	V87 – V124	Multiple choice, multiple-response scale

As can be seen from Table 6.7, the questions asked to *non-Internet shoppers* focused on the factors they consider to be of importance when considering whether or not to purchase via the Internet. It also tries to establish if they consider purchasing via the Internet in the future and from which product and service categories they consider purchasing.

Section B finally tries to establish whether or not non-Internet shoppers use the Internet to search for information before purchasing from "traditional", non-Internet-based sellers. Upon completing Section B, respondents will automatically be routed to Section D.

Section C: Internet shoppers

Respondents indicating that they have purchased products or services via the Internet before (a "Yes" answer to Question 11 in Section A), will automatically be routed to Section C.

Section C will try to determine which factors Internet shoppers consider as being important when deciding whether or not to purchase via the Internet. The same statements as listed for non-Internet shoppers will be presented (question 12) for comparison purposes. Section C will also establish which product and service categories Internet shoppers have purchased from , will purchase from again and which they have not purchased from in the past they will consider purchasing from in the future.

Finally, as with Section B for non-Internet shoppers, Section C will determine whether or not Internet shoppers search for information on the Internet prior to purchasing from non-Internet based sellers. Section C will also try to establish from which categories they search on the Web. Table 6.8 summarises the questions asked to Internet shoppers.

TABLE 6.8: QUESTIONS APPLICABLE TO INTERNET SHOPPERS (SECTION C)

Question	Variable	Scale type
18. Please indicate how important the factors listed below are to you when deciding whether or not to purchase via the Internet: [24 factors listed]	V125 – V148	7-point Likert scale
19. From which of the following product and service categories have you purchased before and do you seriously consider purchasing via the Internet in the future? (Multiple answers) [38 product and service categories listed]	V149 – V186	Multiple choice, multiple-response scale
20. Have you ever searched for or do you consider searching for product or service information on the Internet prior to purchasing from a non-Internet based seller? (e.g. A physical store or telephone shopping)	V187	Simple category scale
21. From which of the following product and services categories have you searched for or do you consider searching for information on the Internet prior to purchasing from a non-Internet based seller? (e.g. Physical store or telephone shopping) (Multiple answers) [38 product and services categories listed]	V188 – V225	Multiple choice, multiple-response scale

Question 18 will try to determine which factors Internet shoppers consider when deciding whether to purchase via the Web. The results for this question will be compared to that of the non-Internet shoppers to determine whether or not any pertinent differences are noted between Internet shoppers and non-Internet shoppers. More detail on the manner in which the comparison will be made is provided in Section 6.9 when the statistical procedures and techniques adopted for the study are discussed.

Section C will also determine which product and service categories Internet shoppers have purchased before. Respondents will also have to indicate which they will purchase from again ("yes", "no" or "uncertain" responses have to be provided). Respondents will also be requested to indicate from which product and service categories they have not purchased before, they consider purchasing from in the future.

Section C will, as with Section B, conclude by determining whether or not Internet shoppers search for product and service information on the Internet prior to purchasing from non-Internet based sellers. Those respondents who indicated that they search for information will also be requested to indicate for which product and service categories they search on the Web.

Once Internet shoppers completed Section C, they will automatically be routed to the final section of the questionnaire, Section D.

Section D: Demographic information

The aim of the final section of the questionnaire, Section D, was to derive demographic information from the respondents who completed the questionnaire.

Table 6.9 below provides a summary of the questions that were used to compile a demographic profile of the respondents to the study.

TABLE 6.9: QUESTIONS APPLICABLE TO DEMOGRAPHIC INFORMATION

Question	Variable	Scale type
Please provide the following information about yourself:		
22. Gender	V226	Multiple choice, single-response scale
23. Age	V227	Multiple choice, single-response scale
24. Household Language	V228	Multiple choice, single-response scale
25. Gross Monthly Income	V229	Multiple choice, single-response scale
26. Highest Qualification	V230	Multiple choice, single-response scale
27. In which area do you live or which area is closest to you?	V231	Multiple choice, single-response scale
28. Population Group	V232	Multiple choice, single-response scale
29. Marital Status	V233	Multiple choice, single-response scale
30. Number of people actively using the Internet (more than once a week) in your household	V234	Multiple choice, single-response scale
31. Number of people in your household	V235	Multiple choice, single-response scale

Due to possible sensitivity in providing demographic information, it was explicitly stated, before respondents provided the required information, that the information provided will be treated as highly confidential and totally anonymous. It was envisaged that by ensuring confidentiality, respondents would be more willing to provide personal information.

In conclusion to the questionnaire structure discussion, Table 6.10 below indicates which questions included in the questionnaire (with relevant variables) will provide data that can be used to evaluate the hypotheses formulated for the study.

TABLE 6.10: RESEARCH HYPOTHESES AND QUESTIONNAIRE MATRIX

Hypotheses	Questions	Variables
H_1 : The decision to purchase via the Internet is significantly influenced by factors consumers consider prior to purchase	Q11; Q12: Q18	V21; V22–V45; V125 – V148
H_2 : The factors Internet shoppers consider prior to purchasing via the Internet are significantly influenced by the period of Internet usage	Q3; Q18	V8; V125-148
H_3 : The period of Internet usage significantly influenced the decision to have purchased via the Internet	Q3; Q11	V8; V21
H_4 : The period of Internet usage significantly influences the decision of non-shoppers to purchase via the Internet in the future	Q3; Q11; Q13; Q14	V8; V21; V46; V47
H_5 : The period of Internet usage significantly influences the decision to search for product or service information on the Net prior to purchasing from non-Internet based sellers	Q3; Q16; Q20	V8; V86; V187
H_6 : There is a significant difference between Internet shoppers and non-shoppers in their decision to search for product and service information on the Internet prior to Offline purchases	Q11; Q16; Q20	V21; V86; V187
H_7 : The period of Internet usage significantly influenced the product and service categories Internet shoppers have purchased via the Internet	Q3; Q19	V8; V149-V186
H_8 : The period of Internet usage significantly influences the product and service categories Internet shoppers and non-shoppers consider purchasing via the Internet in the future	Q3; Q15; Q19	V8; V48 – V85; V149 – V186
H_9 : Demographic variables of Internet users significantly influence whether Internet users purchased products or services via the Internet	Q11; Q22-Q29	V21; V226-V235
H_{10} : Demographic variables of Internet users significantly influence the product and service categories Internet users purchased via the Internet	Q19; Q22-Q29	V149 – V186; V226 – V235
H_{11} : Demographic variables of Internet users significantly influence the product and service categories Internet shoppers and non-shoppers consider purchasing via the Internet in the future	Q15; Q19; Q22-Q29	V48 –V85; V149 – V186; V226 – V235

6.6.2 Web-site design

The Web-site was specifically designed with a number of " rules" for successful execution of the study in mind and to be compatible with both Microsoft Internet Explorer and Netscape Navigator Internet Browsers. These "rules" will briefly be discussed to create an understanding for the rationale for doing so.

The first requirement when designing the questionnaire was that **"Internet speed" or "download speed"** while on the Web-site should be a priority. The reason for focusing on the speed element is that respondents might have been discouraged to complete the questionnaire, or abandon the questionnaire after starting, if the interaction time with the Web-site took too long. The end-result from this requirement was that the background design had to be kept simple with no elements (for instance graphics or pictures) that would require additional downloading time.

A second requirement was that the responses to questions should be directly **captured in a database once the questionnaire had been completed in full**. There were two reasons for this requirement. Firstly, by writing responses to the database only once the entire questionnaire had been completed, would ensure completeness of all responses. There would, therefore, not be half-completed questionnaires (and database entries). The second reason is linked to the first, namely that the respondent would not be able to complete a questionnaire partly on a number of occasions (Web-sessions), possibly leading to duplications for the same respondent.

The second requirement led to subsequent requirements that the Web-site had to meet. Firstly, to ensure that a respondent who had to complete a questionnaire through a number of Web-linked sessions, the Web-server generated a unique reference number each time an Internet user visited the Web-site. The respondent could at any point, while completing the questionnaire, decide to save the responses to questions already provided and leave the Web-site or terminate the Internet session. The reference number could then be used when the responded re-connected to the Web-site to continue with the questionnaire at the point at which the previous session was terminated.

Two important considerations for this requirement should be mentioned. Firstly, the reference number "protected" the database in the sense that data would only be captured once the entire questionnaire had been completed. Secondly, the reference tool also acted as encouragement to the respondent to complete the questionnaire, since the respondent could simply continue with the questionnaire instead of starting again from the beginning. It is important to note why a respondent would possibly not complete the questionnaire in one session. One possible reason is that the respondent's connection to the Web could have been terminated (e.g. a poor connection, a power failure, visitors arriving or call-waiting on their telephone line – which would most probably terminate the connection).

The second requirement also led to the issue of database integrity due to capturing only taking place once the questionnaire had been completed. In short, capturing in this manner was made possible by writing responses to temporary files on the Internet server with a command that once the final question had been completed, the system would write the files associated with the respondent's responses to the database. The problem that had to be addressed, was that respondents could move back to questions (pages on the Web-site) and complete other options not selected before, resulting in a corrupt database entry. For example, a respondent could select "have not purchased before" and be automatically routed to Section B of the questionnaire.

After completing the section, the respondent could move back to the main "routing question" ("have not purchased before") and select a different option, i.e. "have purchased before". The Web-site would automatically route the respondent to Section C to be completed. Since the data, at this point of the process, is captured in a temporary file, a final entry in the database once the questionnaire had been completed would have shown that the respondent indicated a "has purchased before" as well as a "has not purchased before" response. This imposed a serious threat to the integrity of the data and the decision was made to de-activate the "back" ("previous") option on the entire Web-site.

A third requirement, aimed at ensuring complete database entries, was to ensure a question (and all sub-sections thereof) was completed before the system would allow the respondent to proceed to the next question. The requirement was met by blocking each entry-field for each possible response for each question (with the exception of question 19) and writing a programming rule that each question and sub-section had to contain an answer before the respondent would be routed to the next question. If an answer was missing, the system would notify the respondent by means of a pop-up screen that an answer or sub-section thereof had to be completed.

As mentioned above, question 19 was the exception to the rule. The reasoning for excluding this question was that respondents (only applicable to current Internet shoppers) had to indicate whether or not they have purchased from a category before and from which categories they considered purchasing from in the future. If the rule above was used for this question, respondents would have had to indicate that they purchased from all categories, which would make the findings from this question totally invalid.

Question 19 did however contain a set of rules applicable to this question only. It was considered important, for the purpose of the study, to determine whether or not current Internet shoppers who have purchased from a specific category before would purchase from the same category again. Respondents, therefore, had to indicate for each category from which they have purchased before, whether or not they would purchase from the same category again, will not purchase from it again, or were uncertain whether or not they will purchase from it again. If the future consideration to the selected category was not indicated, the system would notify the respondent by means of a pop-up instruction to complete the specific omitted category. The system would route the respondent to the next question only once the rule/requirement was adhered to.

Once the final question had been answered, the Web-server automatically wrote a respondent's response to the questionnaire directly to the database.

6.7 INTERVIEW PROCEDURE

As indicated earlier, the sample population was reached by means of an e-mail that was distributed from a server hosted by ISP "X". The e-mail invited ISP "X" (a) and ISP "X" (b) Internet users to participate in the research project and mentioned that, by completing the questionnaire in full, they could be eligible to win a prize. The e-mail letter (attached in Appendix 8) contained an automatic link that, if clicked on, would route the respondent directly to the Web-site (due to confidentiality reasons, the Web-site address can not be divulged) where the questionnaire was hosted (on the ISP "X" Internet Server).

Once logged onto the Web-site, the respondent was greeted with an introductory message that provided more information than was contained in the e-mail message. The central message that a potential respondent should understand from reading the introductory page was that participation to the research project was totally voluntary and that confidentiality was assured. The need for the selected Internet users to understand the conditions of participating in the research project was a critical condition set by the ISP "X" database owner (to ensure that the rights and privacy of the Internet subscriber is protected).

The introductory page also indicated which prizes could be won by participating in the research project. It was envisaged that by offering potential respondents the opportunity to be rewarded for participating, a greater percentage of the users would participate. The prizes, all of which were sponsored by ISP “X”, that participants to the study could win in a lucky draw were: 10 golf shirts and caps; 10 LCD telephones and 4 DIVA internal ISDN modems. The prizes were to be delivered to the physical addresses of the winners.

Respondents had a choice after reading the introductory page of the questionnaire whether or not they wanted to participate in the research project. If they chose to continue (by clicking on the continue button), they would be routed to Section A of the questionnaire. If they were not interested and clicked on the exit button, they would be routed to the ISP “X” Internet Web-site.

Once respondents started completing the questionnaire, they would automatically be routed through the Web-site according to their answers to the questions. Appendix 7 shows how the automatic branching navigated respondents through the Web-site.

6.8 CODING, EDITING AND TRANSFERRING OF DATA

Coding involves the process whereby numbers or symbols are assigned to answers for analysis purposes (Cooper & Schindler, 2001: 424; Burns & Bush, 1998: 453 and Sudman & Blair, 1998: 415). Two categories of coding are distinguished (Cooper & Schindler, 2001: 424), namely alphanumeric coding (when letters are used in combination with numbers and symbols for coding) and numeric coding (exclusive use of numbers). Numeric coding was used for the study.

The questionnaire used in the study did not contain any open-ended questions. Pre-coding of the questionnaire was therefore done when the questionnaire was developed, ensuring that the answers provided by respondents to questions were directly captured in a coded format onto a database.

The purpose of **editing** is to detect errors and omissions and to correct them if possible (Cooper & Schindler, 2001: 423). The data will be checked to identify any possible errors and it will, if possible, be corrected. Any errors detected and corrections done will be reported.

After discussions with the Statistics Department at the University of Pretoria (who will analyse the data), it was found that the manner in which the data was recorded could not be used as input to the SAS computer statistical package. The data was captured in the database, residing on the Web-server, in a "string format", i.e. q1_3,q2_3,q3_1,q4_3,q5_4 etc. For the database to be imported to the SAS computer program, the required format should have been as follows:

Q1	Q2	Q3	Q4	Q5	etc.
3	3	1	3	4	etc.

where only the actual **coded value** is reported in the relevant question column.

Two options could be followed to overcome this obstacle. Firstly, a manual process could have been followed where the "string format" data could be captured manually in a Microsoft Excel Workbook. Secondly, a computer programmer could be approached to write a computer software programme that would automatically re-write the "string format" database into a new Microsoft Excel Workbook database. From both a time and cost perspective, as well as an accuracy perspective (possible errors due to manual intervention), a computer software programme was written especially to meet the requirements for further analysis of the data.

The final data file (approximately 5 Meg.) was transferred (through e-mail) from the researcher to the Statistics Department at the University of Pretoria.

6.9 STATISTICAL PROCEDURES AND TECHNIQUES ADOPTED FOR THE STUDY

This section will provide details regarding the manner in which missing responses will be dealt with. A detailed discussion will be provided regarding the descriptive statistics that will be reported as well as the statistical techniques that will be utilised. The SAS computer statistical software package will be used for data processing.

6.9.1 Missing responses

There are many ways of dealing with missing responses/data. Sudman & Blair (1998: 455 – 456) suggest two ways: Firstly, if the volume of missing data is small enough, it is very unlikely to affect the conclusions from an analysis. The best way of handling small volumes of missing data is, therefore, to exclude it from the analysis. Secondly, if the volume of missing data is large enough to affect the conclusions, the best way of dealing with it is to include it in the results. This can be accommodated in two ways. Firstly, the missing data can be retained as a separate category and should be reported in the results. Secondly, values can be estimated for the missing data by using values from the reported responses.

As discussed in Section 6.6.2, the questionnaire had been designed and posted on the Web in such a way that it was very unlikely that any missing responses would be received. The predicted lack of missing responses can be attributed to the rules written for the Web-site, whereby a response had to be registered for each question before a respondent could continue to the next question and because the data was only captured once the entire questionnaire had been completed.

Based on the reason stated above, all data records that contain missing values to any question of the questionnaire will be excluded. The treatment of missing response will only be adhered to if there is a small number of data records containing missing responses.

6.9.2 Descriptive statistics

Descriptive statistics involves arranging, summarising and presenting a set of data in such a way that the meaningful essentials of the data can be extracted and easily interpreted (Keller & Warrack, 2000: 18). A descriptive analysis is typically used early in the analysis process and becomes the foundation for subsequent analysis (Burns & Bush, 1998: 456). Two types of measures are distinguished, namely that of central tendency and measures of dispersion (Sudman & Blair, 1998: 456).

The two groups will briefly be defined, since some of the measures will be used when reporting the findings of the study in Chapter 7.

(a) Measures of central tendency

The objective of using measures of central tendency is to report a single piece of information that describes the most typical response to a question (Burns & Busch, 1998: 459 and Sudman & Blair, 1998: 456). Three principal measures of central tendency (also called measures of location - Cooper & Schindler, 2001: 442 and Keller & Warrack, 2000: 90) will be discussed, namely the mean, mode and median.

(i) The mean

The mean represents the arithmetic average and is the most common method for finding a typical value for a set of numbers (Cooper & Schindler, 2001: 442; Burns & Bush, 1998: 461 and Siegel, 1994: 71). The mean represents the sum of the observed values divided by the number of observations. The mean is the measure of central tendency most frequently used for interval-ratio data (Cooper & Schindler, 2001: 442).

(ii) The mode

The mode is the value that is observed more frequently than any other value (Sudman & Blair, 1998: 456).

(iii) The median

The median (or halfway point) of a set of observations represents the value that falls in the middle when the observations are arranged in order of magnitude (Keller & Warrack, 2000: 90 and Siegel, 1997: 65). This value has an equal number of observations above and below it (Sudman & Blair, 1998: 456).

(b) Measures of dispersion

Measures of dispersion (also called measures of spread or variability) describe how scores cluster or scatter in a distribution (Cooper & Schindler, 2001: 443), or more simplisticly explained, they depict the "typical" difference between values in a set of values (Burns & Bush, 1998: 462). Only three measures of dispersion will be discussed, namely the variance, standard deviation and the range.

(i) The variance

The variance represents the average squared distance between the values of individual observations on some variable and the mean of that variable (Sudman & Blair, 1998: 459).

(ii) The standard deviation

The standard deviation indicates how far away the average is from the data values (Cooper & Schindler, 2001: 443 and Siegel, 1997: 106).

(iii) The range

The range represents the difference (distance) between the largest (maximum) and smallest (minimum) score (value) in the distribution (set of values) (Cooper & Schindler, 2001: 443 and Burns & Bush, 1998: 463).

6.9.3 Statistical techniques applicable to the study

To achieve the objectives set for the study and to test the hypotheses stated in Chapter 5, a number of statistical techniques will be used. This section will be devoted to only those relevant statistical techniques that will be applied.

A) Factor analysis

Factor analysis is a procedure that groups variables together in an attempt to discover if an underlying combination of the original variables (called a factor) can summarise the original set (Cooper & Schindler, 2001: 575 and Sudman & Blair, 1998: 547). The objective is therefore to reduce many variables that belong together and have overlapping measurement characteristics, to a manageable number (Cooper & Schindler, 2001: 591).

Two approaches to factor analysis can be used, namely principal factor analysis and common factor analysis. **Principal factor analysis** transforms a set of variables into a new set of composite variables or principal components that are not correlated with each other (Cooper & Schindler, 2001: 591), aiming to develop factors that explain the maximum amount out of the total variance in the variables being analysed (Sudman & Blair, 1998: 551). The objective when using **common factor analysis** is to explain the maximum amount out of the variance shared in common by the variables in the analysis. Principal factor analysis, the most frequently used approach (Cooper & Schindler, 2001: 592), will be used in the study.

Two key descriptive results obtained from factor analysis need clarification, namely factor loadings and eigenvalues (Sudman & Blair, 1998: 548).

Factor loadings represent the correlations between a factor and the individual variables being analysed (Aaker & Day, 1990: 547). Each factor will have loadings for all the variables being analysed. Variables with loadings with absolute values larger than 0.50 are said to "load highly" on a factor and are considered to be members of a group of variables identified by the factor.

The **eigenvalue** for a factor equals the sum of the squared loadings for all variables on that factor. The first factor has the largest eigenvalue since it is chosen so as to maximise the sum of squared correlations without any constraints. The second factor has the second highest eigenvalue and so on.

The eigenvalues supply a measure of the percentage of variance in the contributing variables that is explained by the factor, while the sum of the eigenvalues represents the total amount of variance to be explained in the analysis.

All factors with eigenvalues values larger than 1.0 will be retained, since an eigenvalue of 1.0 is regarded as the amount of variance attributable to a single variable, whereas factors with eigenvalues of less than 1.0 are viewed as "explaining" less than one variable's worth of variance (Sudman & Blair, 1998: 549). Factors with eigenvalues less than 1.0 will therefore be dropped from further consideration since they will be regarded as non-significant.

Factor analysis is often characterised by **rotation**, used when unrotated factors are not enlightened. Researchers hope to find, when rotation is used, some pattern in which factors are more heavily loaded on some variables (Cooper & Schindler, 2001: 593). Rotation is performed when a rotation scheme (for example Varimax) literally rotates the factors so that they become closer to some variables and further away from others (Sudman & Blair, 1998: 555).

Two methods of rotation are distinguished as orthogonal and oblique rotation. When using **orthogonal rotation**, repositioning of factors is subject to a constraint that they remain orthogonal (at right angles) to each other, implying that the factors have to remain uncorrelated. **Oblique rotation**, as opposed to orthogonal rotation, is not constrained to remain at the right angles, meaning that factors can become correlated.

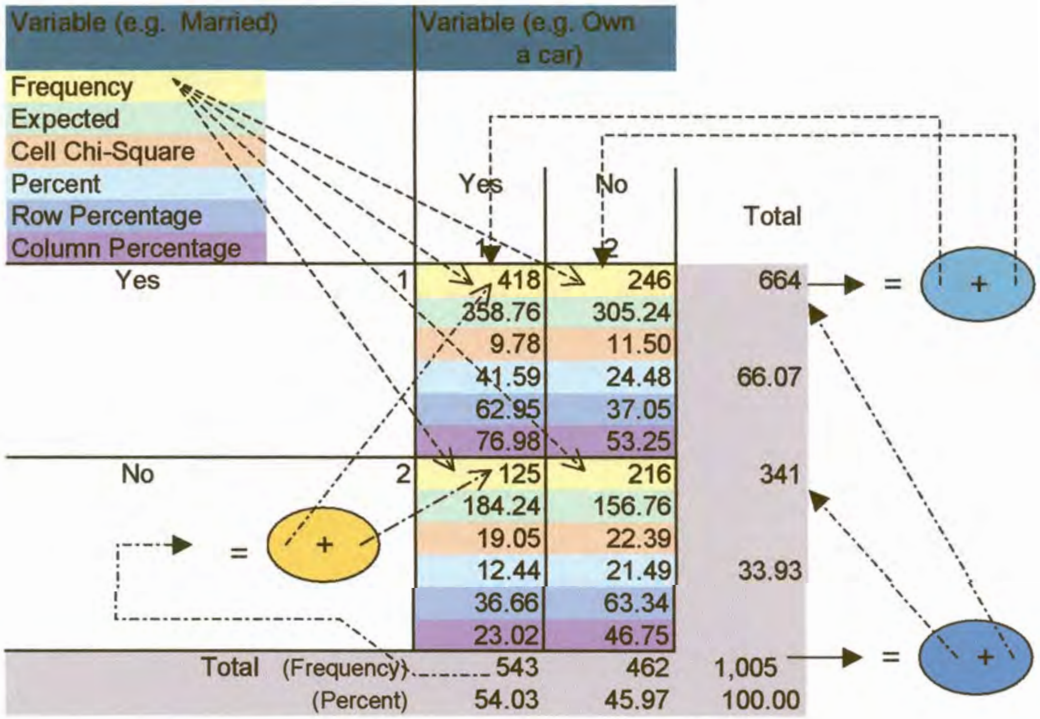
Varimax as a rotation scheme (orthogonal rotation) is used by the researcher since it searches for a set of factor loadings so that each factor has some loadings close to zero and some loadings close to –1 or +1. The reason for using Varimax is that, due to its application, it makes interpretation easier when the variable-factor correlations are close to +1 or –1, indicating a clear association between the variable and the factor (Aaker & Day, 1990: 550).

B) Cross-tabulation

Cross-tabulation is a technique used to compare two classification variables using a row and column format. The basic descriptive result from cross-tabulation is a frequency count for each cell in the analysis (Cooper & Schindler, 2001: 470; Burns & Bush, 1998: 541 and Sudman & Blair, 1998: 475).

Cross-tabulation will frequently be used in Chapter 7 when the results from the research study are discussed. An example, provided below in Figure 6.1, will clarify how the cross-tabulation technique will be used and interpreted in the study. Cross-tabulation will also be used as a statistical technique (inferential analysis) to either accept or reject a number of hypotheses that was set for the study. This will be done by means of using chi-square analysis, discussed later in the section.

FIGURE 6.1: CROSS-TABULATION EXAMPLE



Variable (e.g. Married)		Variable (e.g. Own a car)		
Frequency Expected Cell Chi-Square Percent Row Percentage Column Percentage		Yes 1	No 2	Total
Yes	1	418 358.76 9.78 41.59 62.95 76.98	246 305.24 11.50 24.48 37.05 53.25	664 66.07
No	2	125 184.24 19.05 12.44 36.66 23.02	216 156.76 22.39 21.49 63.34 46.75	341 33.93
Total (Frequency) (Percent)		543 54.03	462 45.97	1,005 100.00

The cross-tabulation example shown in Figure 6.1 has been colour-coded for explanation purposes. The data shown in different colours correspond to the descriptions with the same colours. This is graphically shown for frequency (light yellow), represented by four sets of data (also in light yellow). A number of terms used in the cross-tabulation example shown above have to be explained so that the proposed presentation of cross-tabulation to be used is understood.

Frequency represents the actual frequency recorded in the initial data analysis (that is, the number of responses to the question posed in the questionnaire) (Burns & Bush, 1998: 543). In the example provided above, 125 respondents indicated that they were not married and owned a car. The "Total" figure provided in the same row as the frequency data represents the sum of all data figures in the frequency row (across all columns). That is, 125 ["Yes" option (own a car)] **+** 216 ["No" option (don't own a car)] **=** 341 ("Total" column).

Expected frequencies are defined by Bums & Bush (1998: 543) as: "...the theoretical frequencies that are derived from this hypothesis of no association between the two variables". Bums & Bush (1998: 543) continue by stating that the degree to which the frequency (observed data) depart from the expected frequencies is expressed in a single number called the chi-square statistic. The expected frequency is calculated by multiplying the column total (for all frequencies in all columns) by the row total (for all frequencies in the applicable rows), divided by all the respondents who completed the question. From the example above it can be derived that the expected frequency for unmarried car owners is 184.24, calculated as follow: [543 (all car owners) **x** 341 (all unmarried respondents)] **/** 1,005 (all respondents who completed the question) **=** 184.24. Cell ChiSquare (chi-square for the specific cell under discussion) is 19.05. Chi-square will be discussed in more detail later in the section.

Percent is calculated by dividing the frequency by the **total number** of respondents that completed the questionnaire. Using the same example, it is clear that 12.44% of all the respondents who completed the question are unmarried and own a car. The "Total" column indicates (in the same row as the percentage data) the sum of all the percentages indicated in that row. That is, 12.44% ["Yes" option (own a car)] **+** 21.49% ["No" option (don't own a car)] **=** 33.93% ("Total" column).

Row percent is calculated by dividing the frequency by the **"row specific total"** (indicated in the "Total" column – in the frequency row). That is, using the same example: 125 [frequency (unmarried and own a car)] **/** 341 [frequency (all unmarried respondents)] **x** 100 **=** 36.66%.

Column percent is calculated by dividing the frequency by the **"column specific total"** (indicated in the "Total" row – in the "Yes/No" own a car columns). That is, using the same example: 125 [frequency (unmarried and own a car)] **/** 543 [frequency (all respondents who own a car)] **x** 100 **=** 23.02%.

Finally, the "Total (Percent)" figures are calculated by adding all the Percent figures indicated for each column. From the example it can be seen that 54.03% of the respondents own a car and 45.97% don't own a car (the cumulative percentage – 100.00% - is also shown).

The discussion above should provide sufficient insight into the cross-tabulation process that will follow in Chapter 7 when the results and interpretation for the study is provided.

The final statistical technique that will be discussed, and that is closely related to cross-tabulation, is the chi-square test.

C) The chi-square test

Diamantopoulos & Schlegelmilch (2000: 175) explain that the chi-square test should be used when two groups are compared on a variable which is measured on a nominal scale. Cooper & Schindler (2001: 499) add by stating that the chi-square (X^2) test is used to test for significant differences between observed distribution of data among categories and the expected distribution based on the null hypothesis. Burns & Bush (1998: 543) provide a more detailed explanation by stating that chi-square analysis examines the frequencies for two nominal-scaled variables in a cross-tabulation table to determine whether the variables have a non-monotonic relationship.

The chi-square procedure always starts with the formulation of a statistical null hypothesis that the two variables under investigation are **not** associated (Burns & Bush, 1998: 543). Stated simplistically: chi-square analysis always begins with the assumption that no association exists between the two nominal-scaled variables being analysed.

Of importance to note is that chi-square tests apply to the overall relationship between the variables and not for individual differences (Sudman & Blair, 1998: 479). Using the example in Figure 6.1, a chi-square value calculated would be applicable to the relationship between the two variables (marital status and car ownership) and not for individual categories, for example married car owners.

Burns & Bush (1998: 548) provide a cautionary statement when using chi-square values: "... chi-square value says nothing by itself – you must consider the number of degrees of freedom in the cross-tabulation table because more degrees of freedom are indicative of higher critical chi-square table values for the same level of significance". Burns & Bush (1998: 548) continue by explaining why this statement is made: "The logic of this situation stems from the number of cells. With more cells, there is more opportunity for departure from the expected values". (The degree of freedom (DF) is calculated as (*R*-1)(*C*-1), where *R* is the number of rows and *C* is the number of columns - Sudman & Blair, 1998: 479).

The discussion above is of extreme importance, since the computer programme used when the chi-square values are calculated will also provide the probability of the null hypothesis by taking the nurnber of degrees of freedom into account. The probability percentage will be key to either accepting or rejecting the hypotheses set for the study.

The discussion on the chi-square test can be closed by stating that a chi-square analysis yields the probability that the researcher will find evidence in support of the null hypotheses if the study is repeated numerous times with independent samples (Burns & Bush, 1998: 548). For example, if a probability of 0.02 is found for the null hypothesis, the researcher would be able to conclude that evidence to support the null hypothesis will be found only two percent of the time. This implies that there **is a significant association** between the variables, resulting in the researcher being able to **reject** the null hypothesis (keep in mind that the null hypothesis states that there is **not** an association between variables).

6.10 SUMMARY

Chapter 6 provided a theoretical perspective on research methodology. The researcher indicated how the theory will be applied to conduct the primary research component of the study. The chapter indicated how the sampling process was followed, provided a discussion on the questionnaire that will be used and indicated how the Web-site was designed to accommodate the survey process that will be followed together with the methods on how the data will be collected. Providing a discussion on descriptive statistics and statistical techniques that will be used to interpret the data that will be obtained from respondents concluded the chapter.

The results obtained from the primary research component of the study and the interpretation thereof will be discussed in Chapter 7.