CHAPTER FOUR:

RESEARCH AND STATISTICAL METHODOLOGY

4.1 INTRODUCTION

> *If we knew what it was we were doing, it would not be called research, would it?*
>
> (Albert Einstein, cited in Calaprice, 2005:15)

The discussion thus far has dealt with gaining a deeper understanding of the formulation of a hypothetical model of the measurement construct, namely *perceptions of the advanced automated aircraft training climate.* The literature review built a foundation for the model and the subsequent framework, which demonstrated *why* it is important to develop such a measurement. This chapter contains a discussion of *how* the research was conducted. The discussion includes the type of research design employed, the philosophy that underpinned the construction of the measuring instrument and a rationale for the methodological approach adopted in this research.

According to Babbie (2010), possibly the most effective methodology that can be used to gather information about a large population, is the survey method. Surveying an appropriately selected sampling frame of the population is particularly useful when a researcher wishes to measure attitudes and orientations towards hypothesised phenomena (Schreiner, 2010). Furthermore, it is suggested that the survey method is an ideal vehicle for the purposes of conducting descriptive, exploratory and explanatory studies (Babbie, 2010; Cooper & Schindler, 2003).

When designing the present study, heed was taken of Pepper's (1970:71) eloquent proposition that "to the positivist a hypothesis is a human convention for the purpose of keeping data in order; it has no cognitive value in itself. He is therefore often cynical, or gently indulgent with the wonder and admiration of the common man for scientific predictions". What Pepper suggests is that casually stating hypotheses for the sake of conducting an investigation is not necessarily the correct approach in a positivist paradigm. Good scientific research should therefore start from the basis of established prior theory, which may then provide a logical basis for any stated hypothesis (Creswell, 2002). Since this study was highly empirical in nature, it was appropriate to heed Pepper's *caveat*. By composing specific objectives, while not exclusively relying on pre-stated or contingent hypotheses when insufficient prior theory existed. Thus, overall research goals were ultimately achieved with a combination of validated (when prior theory existed) hypothesised models, statements and specific propositions.

4.2 RESEARCH DESIGN

De Vaus (2003:9) defines the function of a research design as "ensuring that the evidence obtained enables us to answer the initial question as unambiguously as possible". In designing social sciences research, two fundamental approaches are usually followed, namely, positivism or interpretivism. A positivist understands society by using tools typically found in the natural sciences, thereby obtaining rigorous (precise) definitions of phenomena in a contemporary manner by using observable logic to discover causal rules which can be used to predict general patterns of behaviour (Byrne, 1998). This is an important method in the sciences, because testing results' level of replication can challenge or falsify any new theory, model or conclusion. When new and substantive evidence is brought forward, which shows that a prior conclusion may likely be false, the scientific method calls for an adjustment of said theory (Feyerabend, 1985). These requirements were maintained in order to make the present research a highly scientific endeavour. By contrast, the interpretive method involves interpreting society and the behaviour of people in their natural settings, thereby obtaining more fluid definitions of a situation, rather than constructing empirically falsifiable theory (Byrne, 1998). Conclusions drawn in this manner can become highly subjective, and therefore rendered anecdotal. For this reason, the method was not considered for the present research.

This study was dominated by positivism, as it was assumed that the sample group consisted of rational individuals whose behaviour is shaped by their environment. Because the research is a study of a cross-section in the dynamic operational behaviour of airline pilots within a training setting, the approach was eclectic. In other words, various combinations of logical positivism were used in the creation of a hypothesised model of the situation, its operationalization, final analysis and interpretation of the results (for example triangulating empirical data from more than one source).

4.2.1 The research paradigm

Prior to describing the strategy and design of a study, the work must be placed in the context of a particular paradigm that guides the research. According to Creswell (2002), a research paradigm is a philosophical approach to the general nature of the world (ontology) and how we understand it (an epistemology). Similarly, Denzin and Lincoln (2005a; 2005b) explain that an epistemology and ontology are the assumed worldviews adopted by scientists for a particular field of inquiry.

The overall approach taken in this study involved empirical data collection, coupled to a quantitative analytical methodology. Therefore, the specific method used to acquire knowledge from this research was found in *empiricism*, where a structured questionnaire was used as the observation tool of choice. Rationalism was propagated through this scientific method to build any new theory based on the results of the research. Possibly the most effective strategy to gather objective information about the real world, independent of our perceptual knowledge-gathering activities, is the use of the rational scientific method (Feyerabend, 1985). This technique was deemed highly effective for the current research, because human behaviour can be most easily classified, categorised and quantified using statistical methods (Creswell, 2002).

In this study, it was possible to maintain a high level of objectivity because a scientific and methodical quantitative analysis of observed empirical data was undertaken. This entailed a systematic scientific (postmodernist) research design grounded in a theoretical base. As Babbie (2010:10) points out, science “is sometimes characterized as logico-empirical”.

For research to be deemed of high quality and useful, it must possess a number of specific properties. The following criteria were followed in terms of Schreiner's (2010) requirements for good scientific research:

- *reference to seminal findings* – by examining the outcomes of previous findings in the field, good research was built on already discovered sound principles;
- *replication in the chain of reasoning* – systematic logic in the reasoning of results improved the quality of the scientific research. When other researchers are able to follow through with the methodology adopted in a particular study, it shows that the findings are plausible, adding to the quality of a study (Muijs, 2004);
- *objective data collection and sampling methodology* – to make inferences regarding the population, mathematical reasoning for selecting the sample were clearly outlined. In addition, the extraction of the necessary data from the sample was conducted systematically and ethically; and
- *concise explanation of phenomena* – to be termed scientific, the final analysis was based on the gathering of observable, empirical, measurable and replicable evidence.

4.2.2 A classification of the overall research design

The following descriptors were selected as best describing the overall design of this study:

- *Empirical research*: In developing a psychological scale to measure perceptions, the study collected and analysed primary data founded in the principles of sociological positivism.
- *Fundamental research*: Findings from the study were not intended to address a specific management dilemma *per se*. The basic aim was to add to the current academic body of knowledge related to the topic. Although fundamental research is born from curiosity, in many instances, it can also provide commercial benefits in the long term (Nelson, 1959).
- *Exploratory and descriptive research*: The design of the study was based on an exploratory premise, because a preliminary literature review revealed that very

little is currently known about perceptions on the advanced automated aircraft training climate (Ausink & Marken, 2005). According to Creswell (2002:16), a researcher conducts a sequential inquiry when (multiple) methods are used for exploratory reasons, followed by quantitative methods "with a large sample so that the researcher can generalize results to a population". The results of the study therefore provided an in-depth description of South African airline pilots' perceptions of the advanced automated aircraft training climate.

- *Cross-sectional research*: Because the survey instrument was only administered to the sample once, it can only provide a once-off or snap-shot view of the theoretical construct. Therefore, an opportunity to examine the structural equivalence of the established scale may exist for future research endeavours.

The overall research design set out a process of constructing and evaluating explanatory statements or theories about the functioning of the real world. Aliseda (2006:6) explains this approach as follows: "[A]n idea leading to a new theory made in science involves a complicated process that goes from the initial conception of an idea throughout its justification and final settlement as a new theory." In determining the steps followed in the scientific theory-building process, one has to differentiate between an adopted methodology that is abductive (where conclusions are based on reasonable estimation), deductive (where conclusions are based on logic) or inductive (where conclusions are based on empirical evidence). This can be viewed as the spectrum of scientific inquiry (that is, abductive-inductive-deductive). In this study, all three methodologies of reasoning were used to draw conclusions.

4.3 REASONING

It goes without saying, that the ability to reason from a set of truths and logical assumptions, is fundamental in order to structure and extract well-formulated conclusions from a scientific research project. Reasoning is the means by which thinking is channelled from one idea to another (Rosenthal, 1994). For instance, Oaksford, Chater and Hahn (2008) discuss the probabilistic approach commonly used in human reasoning, where conditions are set and logical inferences are drawn. In everyday scenarios, a person may, for example, infer that if something is a fish, then it can swim. Therefore, “Nemo can swim” is a logical inference based on a set of conditions and the assumption that the premise is true, that is, “Nemo is a fish” (Oaksford *et al*., 2008:383). However, it was also borne in mind when conducting the current study that human performance errors are unavoidable at times, and can occur as a result of systematic deviations from logic when the wrong normative standard is used. To take it one step further, logic is expounded through reasoning to derive a valid and particularly substantial means of scientific predictability. For instance, the method was followed in building a predictive logistic model of the phenomena under examination in the present study (see Section 5.10).

The importance of the various reasoning techniques is discussed in the subsections below.

4.3.1 Abductive reasoning

Pierce (1901, cited in Pietarinen, 2006:123) describes the term *abduction* as a logical inference based on estimation (in other words, making a reasonable guess). This leads to the argument that although some truth may be attainable, extremely high levels of certainty may not. Therefore some authors have argued that because high levels of certainty are difficult to attain, more evidence is needed and thus, “abductive foundations are stronger than those based on induction” (Josephson & Josephson, 1996:1). For logic-based abduction, scientists pick out an appropriate explanation or prediction, based on a rational theory representing a domain or set of empirical observations. In other words, by systematically eliminating possible competing explanations from evidential data, the plausibility of the preferred explanation may be supported.

Abductive inference has been slow to develop because logicians concentrate mainly on deductive logic and inductive logic, associated primarily with probability theory. For such reasons, an appropriate statistical level of confidence (in the form of a p-value) is required to substantiate any abductive claims in the observation of human behaviour in social sciences research (Oaksford, *et al*., 2008). However, the process of abductive reasoning can help guide or steer a proposition or hypothesis by providing the best explanation. For instance, if *D* is a collection of data, and *H* explains *D*, and, in addition, no other hypothesis can explain *D* as well as *H* does, then *H* is probabilistically true (Josephson & Josephson, 1996:14). Although the process is not strongly advocated in the field of psychology and in the behavioural sciences, it was found that the process is nonetheless gaining considerable momentum in attempts to understand complex phenomena involving computer algorithms and knowledge-based systems at an exploratory level (Aliseda, 2006). Josephson and Josephson (1996) have demonstrated the ability to compute explanatory hypotheses without relying on induction, deduction or probability theory. In formulating a plausible predictive model for this research, abductive reasoning was relied upon to select relevant independent demographic variables for testing.

4.3.2 Inductive reasoning

Pietarinen (2006) argues that the inductive process could in fact actually be considered a sub-class of abductive reasoning. According to Feeney and Heit (2007:1-2), inductive reasoning is “probabilistic, uncertain, approximate reasoning”, but is important for the following reasons:

- Inductive reasoning corresponds to everyday reasoning; for instance, we may use induction to predict the probability of what the weather may be like. The empirical evidence collected could be atmospheric pressure, humidity, wind, etcetera, to induce a prediction.
- Inductive reasoning plays a significant role in the behavioural sciences, because it is a multifaceted cognitive activity. Furthermore, it is central to categorisation, similarity judgments, probability judgements, and decision-making.

By means of induction, one can therefore draw generalised probabilistic conclusions from a set of logical empirical observations. In analysing the current data set, statistical methodologies were used to obtain a p-value from empirical observations, and in turn to draw such probabilistic conclusions. The method was also used extensively in content validation and to understand the latent structure of the research construct (see Section 4.13). Moreover, this process of reasoning affords a researcher an opportunity to explore the chance that the possibility exists for the conclusions that are induced to be false, even though all the *a priori* premises may have been true (De Vaus, 2003:85). This quality provided additional robustness to the conclusions that were drawn in the study. Similarly, how well these premises supported the conclusion was based on the degree to which the sample was in general a good representation of the population (Aliseda, 2006).

4.3.3 Deductive reasoning

Deductive reasoning is concerned with “drawing logically valid conclusions that must follow from a set of premises” (Feeney & Heit, 2007: 2). Thus, a deductive argument is deemed the most sound and irrefutable method of gaining knowledge. Deductive arguments are logically valid if the *a priori* premises are true (Aliseda, 2006). Furthermore, this method relies on developing the conceptual or theoretical framework before actual empirical testing. In developing the initial hypothesised research construct it was necessary that the deductive process follow from seminal theory, so as to produce valid results.

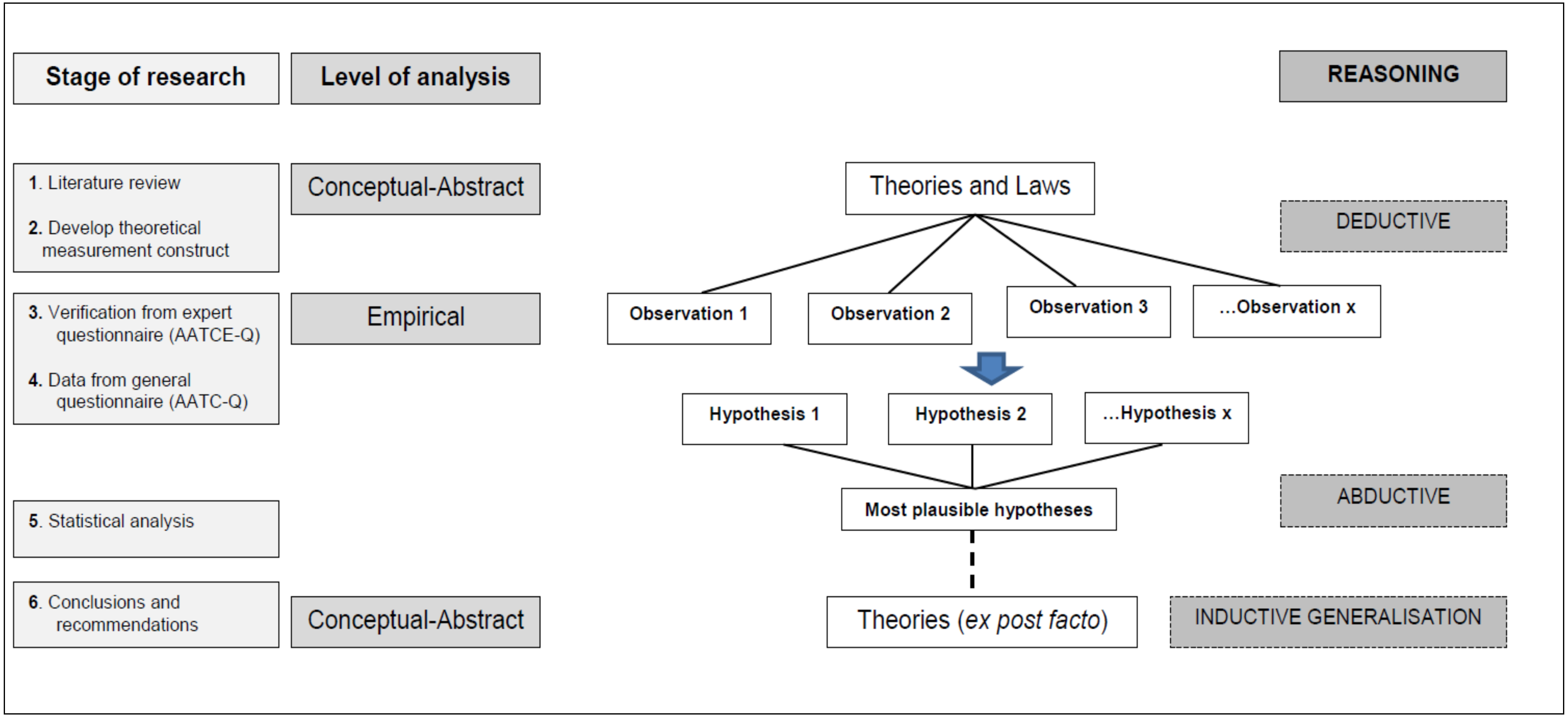
4.3.4 Reasoning followed in the present study

The deductive process has been used successfully in this research to develop an initial conceptualisation of the measurement construct. In the current research, multiple uses of the reasoning spectrum, abductive-inductive-deductive was successfully employed to substantiate conclusions.

Initially, some generalisations were either abduced or deduced from an extensive literature review. The empirical steps of the study involved a combination of both abductive and deductive reasoning, to eventually formulate an *ex post facto* induced theory, and therefore a construction of the final measurement instrument.

Figure 15 contextualises the research design within an eclectic reasoning model. The different stages in the study are contrasted with their level of analysis and the subsequent logical reasoning process.

Figure 15: Integrated model of reasoning used for the study



Source: Adapted from De Vaus (2003), Josephson and Josephson (1996) and Pietarinen (2006)

4.3.5 Ontology

The ontology of a research design has been described in the literature as the philosophical approach taken to the general "nature of existence" (Sullivan, 2009:358). It is a term used to describe categories of entities that may or may not exist in a given domain. For this research, a quantitative ontology was adopted, as it was assumed that the data that was collected could be categorised, classified and counted.

4.3.6 Epistemology

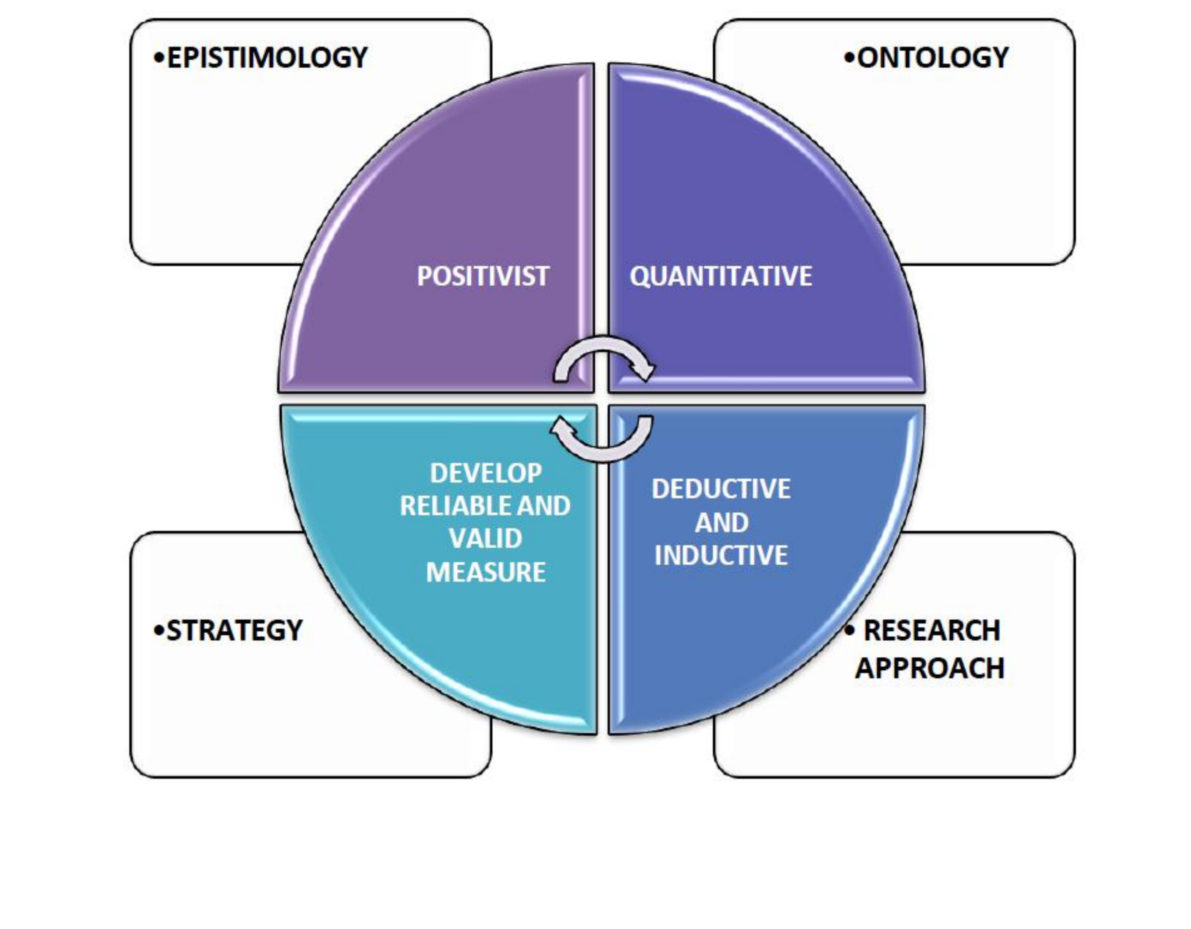
According to Sullivan (2009:180), an epistemology is "the study of knowledge, justification and rationality". A positivist approach excludes speculation as an appropriate origin for an explanation for phenomena. The knowledge gained from this study is of a positivist nature. It can thus be argued that the syllogistic conclusions derived from this study should be considered an objective truth that has been discovered systematically through observation and measurement.

Since, epistemologists argue that simple belief in a proposition is not substantive enough to be an objective truth, that is, only after observable measurement can one therefore deduce a truth (Nelson, 1959). Therefore, any theory of knowledge can only be considered a truth when a specific belief overlaps observable evidence (proof).

4.3.7 Summary of the research design

In measuring latent socio-psychological constructs (in this case, perceptions of the advanced automated aircraft training climate), developing an appropriate method to operationalize the construct was a core process requirement for the quality of the scale's construction (Netemeyer, Bearden & Sharma, 2003). A cycle matrix (see Figure 16) can be used to depict the overall design strategy used.

Figure 16: Research design cycle matrix



Source: Author

4.4 THE EMPIRICAL RESEARCH METHOD: A MULTIPLE METHOD APPROACH

A quantitative research approach based on a positivist paradigm and involving the use of a structured questionnaire to gather data from a sample of airline pilots was employed to meet the research objectives. The ultimate aim of the study was to measure the perceptions of the sample after developing an appropriate instrument.

Concepts applied to human perception are not as clear as concepts related to other fields in psychology, making it more difficult to develop a quantitative measurement (Hakala, 2009). Previous researchers probing the perceptions of large samples of technical professionals found that a structured survey method was by far the most effective way of gathering the necessary empirical evidence (Funk & Lyall, 2000;

James *et al*., 1991; Naidoo, 2008; Sherman, 1997). On the basis of the literature review, it appeared that the most appropriate method to follow in constructing a psychological measurement scale of this nature was to apply a multiple empirical method of inquiry.

The study achieved its objectives by relying on a combination of two separate quantitative (positivist) research approaches. This two-step process resulted in the development of the advanced automated aircraft training climate questionnaire. The approach assisted in eliminating human inquiry errors arising from inaccurate observation, as Clark and Watson (1995) also found in their use of such an approach. The initial survey and subsequent quantitative analysis used expert opinion to validate the content of the theoretical construct, in line with Muijs's (2004:2) contention that "quantitative research is essentially about collecting numerical data to explain a particular phenomenon". This step provided a deeper understanding of the construct by explaining the content of the construct.

Some authors have suggested using a combination of both qualitative and quantitative methods in a single study to benefit from the advantages of "triangulation" (Burns & Grove, 2005:226). However, such an approach was not considered feasible for this study, because using both methods is difficult for an investigator – the data that is extracted needs to be interpreted using two very different philosophical paradigms. Nevertheless, it is always theoretically possible to analyse qualitative data in a quantitative manner, for example, by categorising clustered comments from respondents (Leech, 2004). Moreover, in collecting data at a specific level of measurement, a researcher must extract such information from written words, which are language in an extended (con)text, based on observation, interviews or documents. In the current study, as a secondary source of information, words from both the expert and general surveys were loosely analysed to gain clarity and to guide the study objectives (Denzin & Lincoln, 2005a).

The advantages of combining various methods to triangulate data can be harnessed even within a single ontology. Using two surveys to conduct this study achieved this advantage. The technique has been described as adding a three-dimensional quality to the questionnaire approach (Bergman, 2008). This is particularly true of methodological triangulation, which is generally used to analyse complex phenomena (Burns & Grove, 2005). Strict methodological triangulation in scholarly research is usually divided into two main types. The first is *between-methods* triangulation, which is a complex mixture of the qualitative (interpretivist) and quantitative (positivist) paradigms, and is often difficult to accomplish, as mentioned above. The second is a simpler, *within-method* triangulation, which was used in this study. Burns and Grove (2005:227) explain that within-method triangulation consists of a "multidimensional analysis", or the measurement of a phenomenon using two or even three different quantitative instruments. This was accomplished by initially using a subject matter expert probe, followed by the analysis of data extracted from a refined large sample survey instrument (the so-called two-step process adopted).

The final inquiry approach was based on two elements: a *multiple-methods* triangulation (as opposed to a mixed methods triangulation) with a *within-method* triangulation, as also described by Haworth (1996). The final research design therefore consisted of a sequential approach of quantitative methods without blending the different paradigms *per se*, as is the case in many social sciences research projects.

Data were gleaned from multiple sources and interpreted from the perspective of a positivistic ontology. The complications and sources of human inquiry error inherent in the construction of an effective psychological measurement instrument were mitigated by the advantages of methodological triangulation and a multiple-method system.

According to Bergman (2008:91), the advantages of using methodological triangulation are the following:

- *corroboration* – combining methods mutually confirms results, thus providing greater validity;
- *offsetting* – a study is able to take advantage of the strengths found in two separate inquiries by offsetting any of the disadvantages found in either or both;
- *comprehensiveness* – the researcher is able to provide a more thorough account of the field of inquiry by using a multiple step inquiry;

- *instrument development* – clearer and more structured scale items can be devised from a multiple probe of different sources;
- *credibility* – using a multitude of approaches in the inquiry strategy enhances the integrity of findings; and
- *discovery and confirmation* – this implies using diverse views of the phenomenon to generate objectives and employing quantitative methods to confirm hypotheses.

Teddlie and Tashakkori (2008) argue that research evaluation criteria are vastly improved when the intuitive nature of expert judgement is combined with the robustness of a quantitative analysis. The triangulation of diverse positivist methods can therefore significantly strengthen a researcher's inferences.

According to Creswell (2002:16), a multiple-method approach with methodological triangulation offers a study the following research options (applications in this study are briefly indicated):

- *Exploration:*

  By using two quantitative techniques, additional theory regarding the unknown prevailing training climate can be generated from subject matter expert opinion. For instance, in the current study, an expert commented on the issue that new navigational procedures, such as area navigation (RNAV) or precision-based navigation (PBN), could influence perceptions of advanced aircraft training. This was of research interest, because it provides new insight into the complexity of two distinct parts of automation, that is, air traffic control (future air navigation) and aircraft operations.
- *Confirmation:*

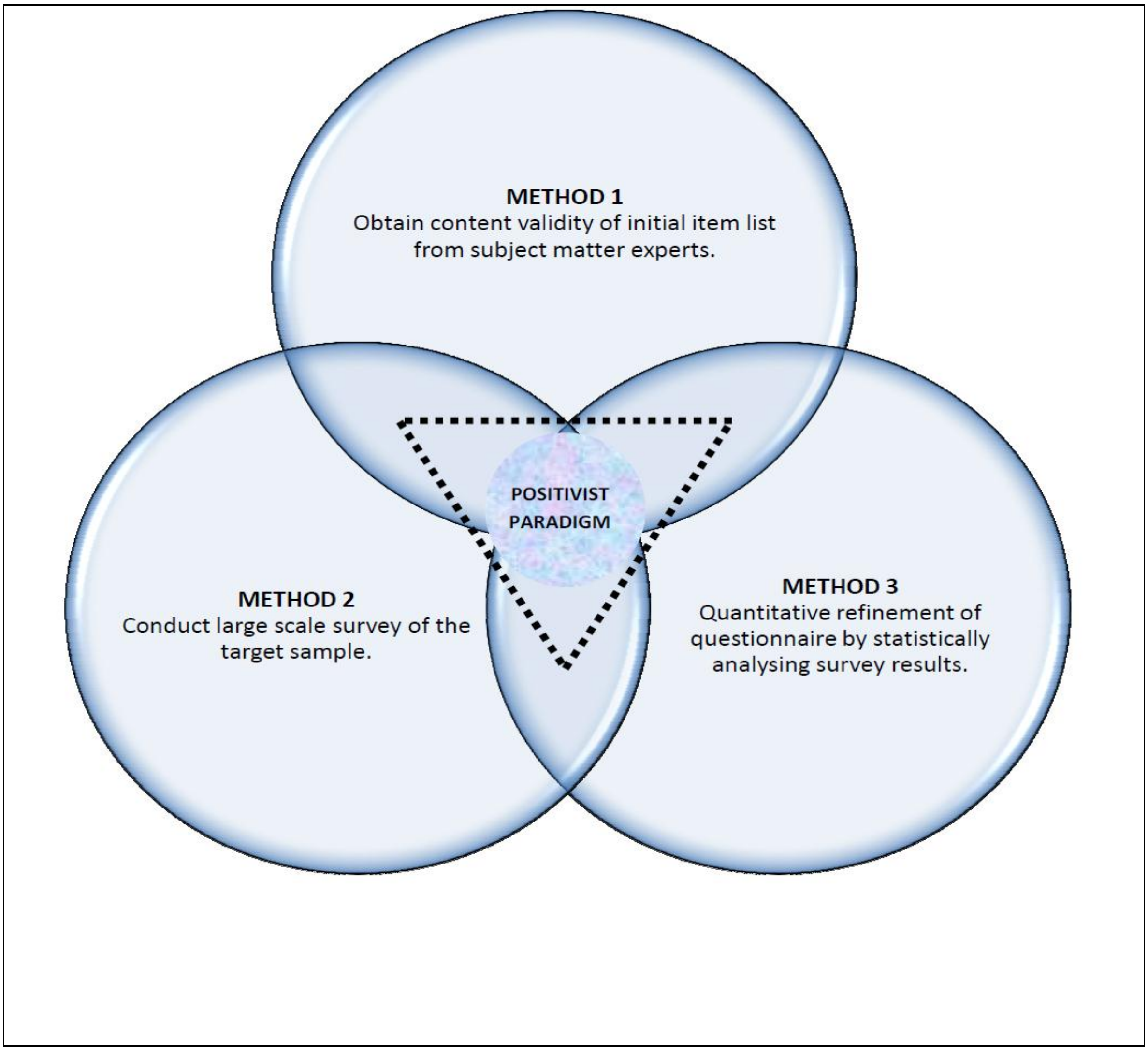
  This involves the quantification of separate findings and statistical analysis to test the theory that has been generated. The results of the general survey in the current study could be traced to aspects mentioned in the content validation of the items in the subject matter expert questionnaire. This provided a level of confirmation that could not necessarily be obtained from only the results of a final survey. For instance, in the final survey, it was found that many respondents were concerned about the loss of their manual handling skills in advanced aircraft. The experts' comments confirmed this finding, in that some flight instructors were of the opinion that new trainees undergoing transition training to highly advanced aircraft had difficulty in selecting appropriate levels of automation (that is, from fully manual, to fully automated) in adverse or non-normal flight situations.

Figure 17 illustrates the non-linear nature of the research design, which remains in the positivist paradigm. A multiple-method approach exhibits the qualities and benefits of positivist quantification, whilst allowing a researcher to gain from dichotomous, yet similar, methods (Haworth, 1996).

Figure 18 then depicts the *sequential* nature of the multiple-method adopted. Read in combination, Figures 17 and 18 show conflicting event lines (in other words, both circular and sequential), however, far from being paradoxical, the contrasting methods proved highly complementary in contributing to the quality of the final research outcome.

Figure 18 depicts the overall research design, which was divided into the two distinct phases, consisting of four stages overall, which maps the present study. Two *questionnaires* gauged the prevailing perceptions of the advanced automated aircraft training climate construct at each stage. The second quantitative probe was used in developing the final measurement scale.

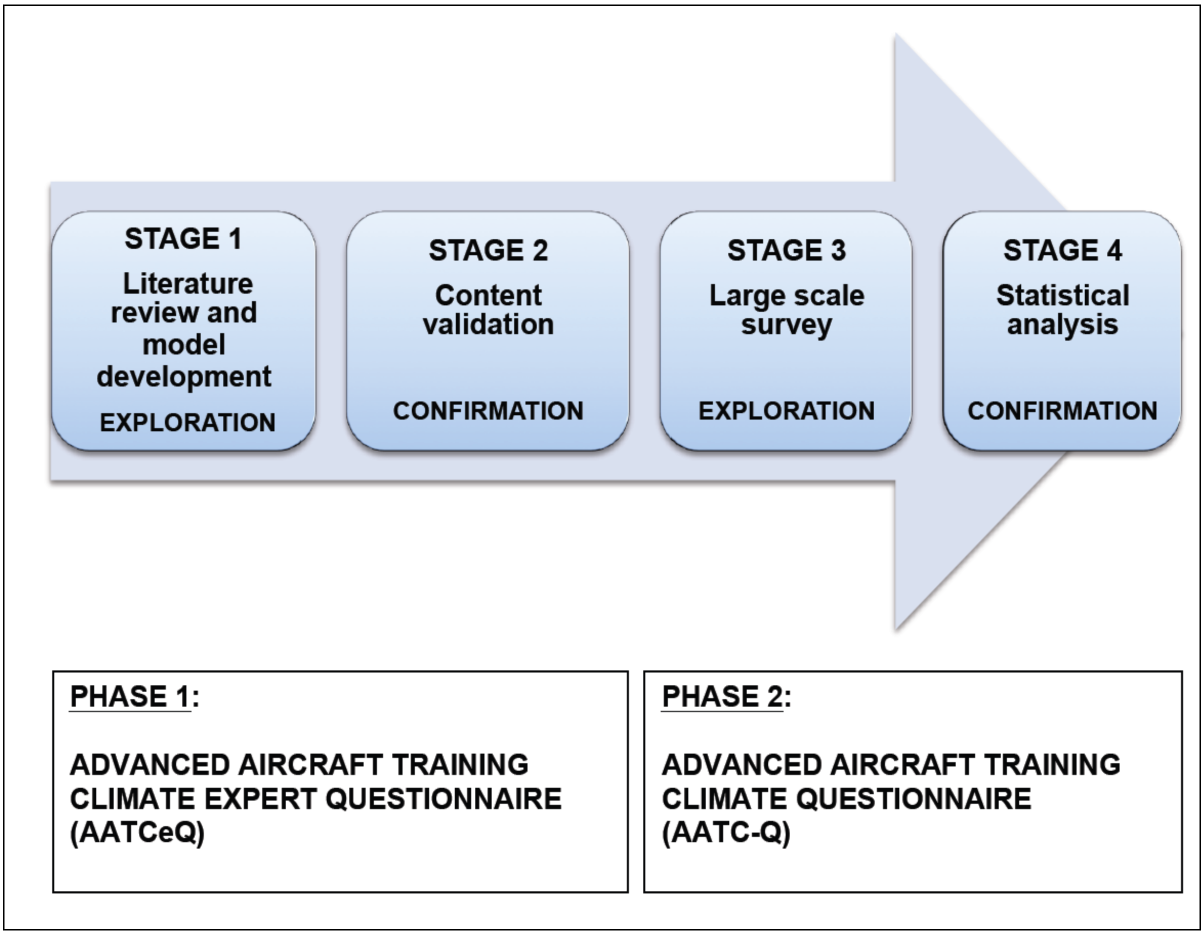
Figure 17: Multiple-method and within-method triangulation


METHOD 1
Obtain content validity of initial item list from subject matter experts.
POSITIVIST PARADIGM
METHOD 2
Conduct large scale survey of the target sample.
METHOD 3
Quantitative refinement of questionnaire by statistically analysing survey results.

Source: Author

Next, Figure 18 shows a four-stage, two-phase process. The questionnaire in Phase 1 was designed to assess the construct by verifying the relevance, conciseness, clarity, and content validity of the deduced pool of initial measurement items. This was possible, because research questionnaires that are quantitative in nature are traditionally refined on the basis of information derived from previous analyses of opinions (words) gained from earlier investigations (Creswell, 2002).

Figure 18: Overall multiple-method research design


STAGE 1
Literature review and model development
EXPLORATION
STAGE 2
Content validation
CONFIRMATION
STAGE 3
Large scale survey
EXPLORATION
STAGE 4
Statistical analysis
CONFIRMATION
PHASE 1:
ADVANCED AIRCRAFT TRAINING CLIMATE EXPERT QUESTIONNAIRE (AATCeQ)
PHASE 2:
ADVANCED AIRCRAFT TRAINING CLIMATE QUESTIONNAIRE (AATC-Q)

Source: Author

The present case made it possible to further abductively explore textual responses to the expert questionnaire and thereby identify any possible additional variables that could be used in the development of the general survey questionnaire. However, to maintain the authenticity of the content validation process, no new items *per se* were added to the final survey (that is, the content of each retained item was unaltered), resulting in a limited number of very high quality final items. Retained items were however, modified for completeness, clarity and comprehension, as recommended in the textual commentary received from subject matter experts (see Table 23).

4.5 MEASURING INSTRUMENTS

Two primary measuring instruments were constructed to gather the data needed to meet the research objectives. The first instrument was a questionnaire sent to a panel of experts (see Appendix A) to validate the hypothetical construct. The final questionnaire (see Appendix F) was used to survey the perceptions of a sample of the target population, namely, pilots with experience of training for advanced automated aircraft.

As discussed earlier, a preliminary literature review served as the basis for developing a hypothetical model of airline pilots' perceptions of the advanced automated aircraft training climate. The main construct had to be operationalized and measured using empirical evidence, data was gathered by means of a psychometrically valid questionnaire designed to identify latent influential factors. In order to construct a measurement instrument from the initial hypothetical model, 17 specific variables were deduced, at three fundamental levels. All these elements fell within the boundaries of three broad areas of organisational behaviour, which are delineated in seminal works from both classical and contemporary theory.

Items in the subject matter expert questionnaire were constructed for the purposes of testing and validating these critical variables. An item list was generated, based on operationalizing the theoretical construct using abductive and deductive reasoning. The following propositions were formulated to then guide the initial item pool construction:

- Airline pilots' perceptions of the advanced automated aircraft training climate manifest themselves at three levels of organisational behaviour, namely the individual, the group and the organisational levels.
- The theoretical model of the construct can be described in terms of 17 core concepts, namely:
  - learning for technology;
  - motivation to train;
  - personality;
  - training stress;

  - training decision-making;
  - training group dynamics;
  - intergroup training behaviour;
  - training teams;
  - training conflict;
  - power;
  - communication;
  - training culture;
  - knowledge environment;
  - structure;
  - training policy;
  - training standards;
  - training planning.
- The demographic characteristics of the sample differ regarding each of the identified criteria derived from the model, thus indicating various levels of the construct.

The tentative item pool used in the Advanced Aircraft Training Climate Expert Questionnaire (AATCe-Q) consisted of 106 positively worded statements, as Barnette (2000) and Gorsuch (1997) recommend. The validation and analysis of the item pool is discussed in Section 4.13. The final items for the general survey (AATC-Q) were retained or discarded based on the significance of Lawshe's (1975) content validity ratio (CVR).

4.5.1 Survey method

Surveys generally fall into two broad categories: questionnaires or interviews. It was decided that relying exclusively on the questionnaire survey method would prove the most effective way to meet the objectives of the study. Cooper and Schindler (2003:325) suggest that the survey method for collecting data be used when one wants to gain "quantitative information about particular phenomena". Creswell (2002) suggests that the survey method be used for comparisons and associations, so as to explore whether relationships between phenomena are present. Generally, a survey is conducted on a fairly large scale, as opposed to a laboratory experiment (which is conducted on a much smaller scale). Cobanoglu, Warde and Moreo (2001) point out that, for the purposes of social surveys, questionnaires, interviews and attitude scales can accurately measure participants' perceptions.

The use of a questionnaire to elicit data from a sample may seem intuitive, however, there are a number of disadvantages associated with this method which did in fact prove challenging for the current study. Welman and Kruger (1999) mention some of these disadvantages:

- There is a possibility of a low response rate. This was a real concern, because "[g]etting pilots to participate in surveys is a general problem in the aviation industry all over the world" (Vermeulen, 2011).
- The researcher has a low level of control over the conditions under which the questionnaire is completed. In this case, because both a web survey and hardcopy questionnaire were distributed, there was a real risk that the survey could possibly be completed by inauthentic (or wrongful) recipients.
- Explanation and clarification of concepts is not possible, because space is limited in questionnaires.
- Anonymity complicates the follow-up on questionnaires. Providing a space for respondents to enter an e-mail address if they wished to receive feedback somewhat mitigated this disadvantage in the current study.
- The survey method is generally used for cross-sectional studies, with mainly closed-ended questions. This can be a disadvantage, because exploration of the phenomena under review may be limited.

The rationale for adopting a questionnaire survey approach for this research despite the above disadvantages was based on two very fundamental advantages found in the technique (Welman & Kruger, 1999):

- a lot of information can be collected within a short time span, thereby saving time; and
- data coding is simplified because the survey is structured and standardised.

Two different questionnaire survey methods (hardcopy and electronic) were used to elicit the data necessary for meeting the research objectives. The hardcopy method consisted of a paper-and-pencil survey, whilst the electronic method was based on either e-mail or hosting on the internet (web-based). Both these methods were used to survey the panel of subject matter experts and the final sample frame.

4.5.2 The paper-and-pencil survey

Traditionally, much psychological and management research (unlike research in other scientific fields) makes extensive use of paper-and-pencil surveys to measure abstract theoretical constructs in order to explore underlying organisational phenomena (Schriesheim *et al.*, 1993). The advantages of the respondent anonymity that can be achieved using this method have been demonstrated in many studies employing this method (Bradburn, 1983). Participants who opted to use this response method were able to record their answers in a questionnaire booklet at any time and without the potential anxiety of having to answer to an interviewer. Schriesheim *et al.* (1993) warn, however, that the quality of measuring instruments may be reduced when there are such high levels of anonymity.

Despite potential disadvantages, due the nature of the target sample (including the fact that they work shifts), the paper-and-pencil questionnaire proved highly useful in gaining adequate coverage of respondents. The general nature of the work involved in operating a commercial aircraft implies that airline pilots do not occupy a traditional office or always work during conventional times. Hence, access to Internet facilities may be limited. However, in order to improve response rates, both an e-mail questionnaire and an internet-hosted questionnaire were constructed.

4.5.3 Electronic surveying

Apart from the advantages of saving time, convenience and coverage, using computer-based questionnaires also eliminates “out-of-range” responses (Bradburn, 1993:333). Such questionnaires allow only pre-determined valid codes to be entered by the respondents, preventing them from marking inapplicable items. Therefore, the use of this type of questionnaire made the management and analysis of the data much easier for the researcher.

To maintain validity, the paper-and-pencil questionnaire was replicated electronically (see Appendix F). First, an electronic questionnaire was constructed, using Microsoft Word's *form* program, and it was then e-mailed to all eligible participants. Secondly, the questionnaire was re-constructed using an open source online survey application, www.limesurvey.org. The advantages of using a web-based survey are legion. For example, the enhanced import and export functions allow a researcher to use statistical and graphical software far more easily than traditional paper methods would (Nunnally & Bernstein, 1994). For this reason, returns from both the pen-and-paper and e-mail questionnaires in the current study were recaptured onto the web-based survey.

Approximately 64% of completed returns came directly from the web-based survey. In making this research choice, during the construction phase, the drawbacks and advantages of web-based surveys were considered, as set out in Table 13.

Table 13: Contrasting the pros and cons of Internet surveys

Advantages/Benefits	Disadvantages/Drawbacks
The researcher is able to tally results instantaneously, as participants submit responses.	Obtaining the correct sample is not an exact science and can become costly or time-consuming.
The ability to conduct a number of surveys over time is enhanced.	Converting a paper-based survey into an electronic format is time-consuming.
It is easier for respondents to remain anonymous.	It takes both research skill and a fair amount of technical ability to conduct a web-based survey.
The turnaround time from drafting a survey to final execution is shortened.	While an internet survey should be compatible with most browsers, the technology is far from perfect, and can result in increased non-response bias.

Source: Cooper and Schindler (2003:340)

4.6 QUESTIONNAIRE CONSTRUCTION

The final data collection instrument was called the *Advanced Aircraft Training Climate Questionnaire (AATC-Q).* In order to partition the instrument, a demographic section and three core dimensions formed the final design, namely:

- Part A (at an individual level);
- Part B (at a group level); and
- Part C (at an organisation level).

This provided a logical flow of the items and created rapport with respondents.

The study set out to develop a measurement *scale*, as opposed to an *index*. Streiner (2003) points out that the items in an index are an important criterion, but that this is not the case in a scale. This was considered in the design of the general instrument. Items in an index are uncorrelated, whereas in an instrument based on a scale design, in general, items tend to be correlated. This scale attribute also suggested that items should be placed in specific and logical groups. Any correlation between items implies that what one item may miss is usually covered by another item. Because the number of potential items capable of reliably tapping a construct is infinite, the researcher has to choose items appropriately. This ensures that as much of the domain is covered as possible and not just one part of it (Comrey & Lee, 1992). In this case, the researcher was confident that the choice of items selected for the general questionnaire would be a valid measure of the theoretical domain, due to the quantitative technique adopted during the first phase of the scale development (computing the content validity ratio from subject matter expert opinion).

The selection of a correct scale is paramount in shaping the questionnaire and the information collected (DeVellis, 2003). The scales used in survey research usually consist of between two and ten points (or categories), depending on how the data collected is intended to be used (Netemeyer *et al.*, 2003; Stevens, 1946). Debate continues regarding the exact number of points that is best for a measurement scale. Arguments against high granularity suggest that respondents cannot discriminate finely enough to justify more than seven points (Bott & Svyantek, 2004).

An objective in the generation of scale items is to have at least “twice as many items” as the final number needed (Nunnally & Bernstein, 1994:128). The current research generated an initial pool of 106 items, guided by a framework derived from the theory that was reviewed. This number was deemed conservative, because some authors have suggested that around 40 items would be appropriate to measure a construct of this nature (Biggs, 1987; Sherman, 1997). Similarly, Nunnally and Bernstein (1994:130) propose that “at least 30 items” are required for a psychometric measure to have a high level of reliability. Items were revised, and some were discarded as unnecessary items after the Lawshe (1975) analysis (see Section 4.13).

A synthesis of the guidelines (see Table 14) followed for developing a perception scale illustrates the fundamental process used by many authors in the literature. The following generic steps guided the development of a quantitative estimate for the theoretical construct of interest:

- Step 1: Develop a theoretical model of the construct;
- Step 2: Generate appropriate items from the theory;
- Step 3: Operationalize the theoretical construct by developing a scale (for instance, using the results from an expert questionnaire); and
- Step 4: Evaluate the robustness of the scale (appropriate statistics to determine validity and reliability).

The development of the instrument for this research was intended to assess the three key perceptual dimensions of the construct (respondents’ perceptions at the individual, group and organisational levels). Alternatively, the developed measurement’s sub-scales assessed concrete variables, which are related to respondents’ perceptions. According to DeVellis (2003), using sub-scales to divide the number of items in a questionnaire allows a researcher to use fewer respondents for factorisation (in other words, fewer than the 300 required for successful factorisation). This was taken into consideration when the response rate turned out to be lower than expected. According to a rule of thumb provided by Cooper and Schindler (2003), the number of respondents in a sampling frame appropriate for a data reduction method is generally five times the number of items in the sub-scale.

The conception or creation of an initial pool of items is a critical stage in questionnaire construction. Clark and Watson (1995) recommend that researchers err on the over-inclusive side of item generation, so as to derive a broader and more comprehensive item pool which goes beyond the researcher's own theoretical view. The design of items used in the questionnaire in this study is discussed more fully in Section 4.6.2.

To guide the questionnaire construction, Table 14 was used to contrast some important recommendations as discussed in the relevant literature.

Table 14: Contrast of scale development guidelines

DeVellis (2003)	Netemeyer *et al.* (2003)	Pett *et al.* (2003)
1. Determine clearly what must be measured.	1. Clearly define the construct and determine its content domain.	1. Clearly identify the measurement framework.
2. Generate an item pool.	2. Generate measurement items.	2. Identify the empirical indicators of the construct.
3. Determine the format for measurement.	3. Judge measurement items.	3. Design and develop the instrument.
4. Have initial item pool reviewed by experts.	4. Design appropriate study to develop the scale.	4. Pilot-test the instrument.
5. Consider inclusion of expertly validated items.	5. Refine the scale.	5. Determine the number of subjects.
6. Administer items to a development sample.	6. Finalise the scale.	6. Administer the instrument.
7. Evaluate the items.		
8. Optimise scale length.		

4.6.1 Scaling procedure

Fiske (2009:449) comments that it “has been said with justification that the history of science could be written in terms of advances in instrumentation”. Furthermore, according to Netemeyer *et al.* (2003), scaling refers to the measurement of a theoretical construct on a multi-item basis. A latent domain is tapped by using a number of alternative items (scale), providing quantitative estimates of the corresponding construct (DeVellis, 2003). Pett, Lackey and Sullivan (2003) contend that in developing a psychological scale, the researcher is more interested in the construct the items endeavour to measure than in the items themselves. For this reason, in the present study, it was important first to validate the quality of the degree to which the items tapped the construct, prior to developing the actual scale. This was achieved by using the technique advocated by Lawshe (1975), as discussed in Section 4.13.

The most appropriate method for extracting the data needed to measure the construct of interest (perceptions of the advanced automated aircraft training climate) was to use a multi-dimensional questionnaire or instrument containing Likert-type (polytomous) items (Likert, 1958; Pett *et al.*, 2003). The items are considered continuous in nature and, in this case, were based on two extreme anchors. The advantage of this technique is that a Likert-type design assumes a latent (continuous) variable with a value that characterises respondents’ attitudes (Likert, 1958). The underlying dependent variable varies quantitatively, as opposed to qualitatively. This is an important quality, which makes the method a popular scale in psychological and behavioural research for measuring opinions, beliefs or attitudes (Cooper & Schindler, 2003; Creswell, 2002; Pett *et al.*, 2003). However, Uebersax (2006) found rampant confusion about the use of Likert-type scales and items in many scholarly articles. With this in mind, Uebersax (2006) pointed out that researchers should take cognisance of the following characteristics that have come to define a Likert-type item-based scale:

- the scale itself consists of several items;
- options are arranged horizontally;
- response options are anchored with consecutive integers;

- the response options should, in addition, be anchored with verbal labels representing evenly spaced gradation;
- response options are symmetrical about a neutral point, which implies that the scale should contain an odd number of responses to induce a natural central point; and
- the scale measures levels of agreement or disagreement in respect of a given statement.

Gerbing and Anderson (1988) found that a respondent's behaviour in complying with the internal consistency of Likert's criterion would tend to exhibit a linear and continuous relationship to the score, making it advantageous to use it for statistical analysis. Exploring the latent structure of a construct provided by Likert-type items provides a more robust factor analytic option than other alternatives, such as Thurstone's approach to scaling (Andrich, 1978; DeVellis, 2003).

4.6.2 Item design

In constructing a perception or attitude measurement instrument, "items tend to be very narrow and specific, developed to match a particular situation" (Pett *et al*., 2003:15). Thus, Kline (2000a, 2000b, cited in Pett *et al*., 2003) points out that the *quality* of the items tapping the domain of interest is a far more important criterion early in the exploration than psychometric virtues such as validity or reliability. The quality of a scale's inter-item correlations depends, to a large extent, on the number of response options in an item using a Likert-type design (Streiner, 2003).

In deciding on the number of response options that may be appropriate to this study, the researcher followed the steps recommended by Pett *et al*. (2003) and by Gerbing and Anderson (1988):

- Step 1: Decide on what number is appropriate, depending on how well subjects are deemed to be able to discriminate meaningfully between response options relating to statements. Since typical advanced aircraft airline pilots have many years of experience, it was assumed that they have mastered their skill to some degree of expertise. Therefore, potential respondents were assumed to have the ability to discriminate on each item at a far deeper level than the average layman.

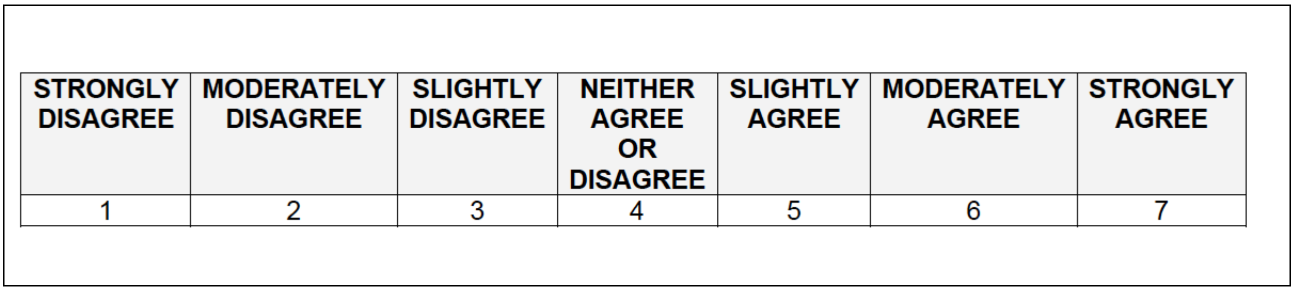
- Step 2: Determine whether or not the sample is able to distinguish a construct finely.
- Step 3: Decide how precise the responses should be.

The literature review revealed very little consensus regarding the optimum number of response points to include in an item in a multidimensional Likert-item questionnaire (Streiner, 2003). There are also various advantages and disadvantages to offering respondents an even- or an odd-numbered item scale (Cooper & Schindler, 2003). Creswell (2002) claims that an even-numbered item scale forces subjects to either agree or disagree with the statement, but that this may lead to frustration or even to their discarding the questionnaire altogether. By contrast, an odd-numbered scale may entice some respondents to neglect careful consideration of the statement and continuously give neutral or middle responses (DeVellis, 2003). Nevertheless, several authors, including DeVellis (2003), Field (2009) and Streiner (2003), present convincing arguments in support of the use of an odd number of response options in psychological instrument development. Uebersax (2006) advises researchers to use a neutral point in the design of a Likert-based item, because this method has the advantage of mitigating respondents' frustration levels at not being able to choose a middle stance when they may be unsure of their decision.

The literature review suggested that the majority of perception and attitude measures which used an odd number of item categories in a Likert (1932) design multi-item scale demonstrated very high levels of reliability (Gliem & Gliem, 2003). Furthermore, according to Masters (1974), research findings have shown empirically that for respondents whose opinions do not diverge widely (a relatively homogeneous sample of respondents), the internal consistency of a measure improves as a direct function of the number of categories employed in the item.

With the above argument in mind, a seven-point Likert-type item measuring scale was designed for the present study. For each statement in the scale, the respondents indicated the degree to which they disagreed or agreed with the item. Therefore, high scores would indicate that a respondent held a positive perception of the construct. An example of the seven-point anchored item used in the general survey is depicted in Figure 19.

Figure 19: Seven-point Likert-type item



STRONGLY DISAGREE	MODERATELY DISAGREE	SLIGHTLY DISAGREE	NEITHER AGREE OR DISAGREE	SLIGHTLY AGREE	MODERATELY AGREE	STRONGLY AGREE
1	2	3	4	5	6	7

Source: Adapted from DeVellis (2003) and Likert (1932, 1958)

4.6.3 Rationale for using only positively worded items

Acquiescence bias is another issue of contention when designing a questionnaire. A handful of scholars have demonstrated that a balanced range of items will prevent difficulties encountered by researchers when respondents tend to one extreme of the item scale (Billiet & McClendon, 2000). However, Kristovics (2010) pointed out that these arguments engage in “statistical play”, and are therefore not based on truly scientific reasoning. Kristovics (2010) suggests that researchers should instead maintain a pool of unidirectional statements in scale development.

Welman and Kruger (1999) found that negatively worded items create the error of central tendency. To eliminate this bias, researchers should endeavour to avoid statements that reflect extreme negative positions, for instance “I *never* enjoy simulator training”. Negative items may also tend to frustrate participants or result in their abandoning the questionnaire altogether, especially if the sample comes from a professional population who take pride in the topic under review (Pololi & Price, 2000). Furthermore, such sentiments are corroborated in statistical practicality. For example, an exploratory study by Vermeulen (2009) which used a bi-directional item pool to survey flight instructors’ attitudes towards gender issues required changing the final items to reflect perceptions in a more logical manner – the researcher had to recode negatively worded items so that high scores related to positive perceptions, while the “inverse would be true for low scores” (Vermeulen, 2009:131).

Another danger of using bi-directional items is that such statements may not be true opposites of each other. One can then argue that a truly *balanced* item pool may be very difficult or impossible to construct. Analysing internal consistency, factor structures and other statistics when negatively worded items are used, either together or separately, can be problematic for any researcher (Barnette, 2000; Kristovics, 2010; Vermeulen, 2009). Additionally, in "situations in which respondents can be expected to provide reasoned responses and are willing participants, the need for such a practice would seem to be minimal and may actually be detrimental to the validity and reliability of survey scores" (Barnette, 2000:362).

A primary objective of the study was to determine the underlying structure of the research construct, based on a sample of highly experienced automated aircraft pilots. Participants in the research sampling frame were regarded as professional, as they all hold the necessary licences and certificates as regulated by the Civil Aviation Authority of South Africa, which can be obtained only after acquiring the mandatory levels of training and experience stipulated (CAA, 2011). For this reason, designing positively worded quality items was highly appropriate. The recommendations of Barnette (2000) also played a decisive role in the decision to use only positively worded item statements for the final scale.

4.6.4 Rationale used in the clustering of questionnaire items

Various authors have discussed the advantages of applying item response theory to the structure of scales and the exploration of datasets (Hambleton & Rogers, 1989; Meijer & Baneke, 2004). The origins of item response theory can be traced to the seminal work of Lawley (1943) and Ledyard (1966). Cronbach (1942:299) defines a "response set" as the tendency for a participant to agree or disagree with an item statement, independently from its content. According to Goldstein and Wood (1989:140), "[i]tem response theory (IRT) hinges crucially on the assumption that only a single latent trait underlies performance on an item". This assumption raises the question of whether grouping items into themes or clusters can influence or bias participants' responses. Determining the level of inherent response bias designed into a questionnaire is important, because research into item response theory indicates that the way the items in a survey are constructed can significantly influence the quality of statistical computations.

The order in which to place statements in a questionnaire remains a contentious issue (Simon, Little, Birtwistle & Kendrick, 2003). For instance, Ballinger and Davey (1998) propose a funnelling approach, where questions become progressively narrower in scope. By contrast, Wilson and McClean (1994) suggest grouping statements or questions with a similar topic coverage. The literature also provides examples of the necessity of keeping sensitive statements only in the middle of the questionnaire so as to avoid participant embarrassment early in their participation (Walker, 1996). To achieve the objectives for this research, the questionnaire items were placed in dimensional groupings, as suggested by Wilson and McClean (1994). Items were grouped according to their levels of analysis, that is, either at the micro (individual or person), meso (group) or macro (organisational) level.

The importance of the ordering of the items and its impact on response bias should not be underestimated (Simon, *et al.*, 2003). Ordering is important because it presents a contextual effect, which may or may not influence the responses to particular items (Hambleton & Rogers, 1989). Therefore, the design of the current study's structured questionnaire involved clustering items to counter this kind of bias by not *labelling* the underlying or grouping theme. That is, respondents were not explicitly made aware of any particular grouping. Furthermore, only positively worded statements were placed in each latent group to limit the likelihood of bias between the dimensions of the questionnaire.

The main hypothesised construct in this study is systemic, and is therefore comprised of three core dimensions, or sub-constructs (an organisational behaviour approach analyses variables at an individual, group and organisational level). The operationalization of the construct resulted in a questionnaire that extracted both demographic and perceptual data (opinions, beliefs, attitudes). The aim of the Phase 2 questionnaire development (the AATC-Q) was to produce statements that represented each of the three dimensions and 17 conceptual themes identified from the literature study. A structured questionnaire with clustered or grouped items was deemed the most appropriate method for capturing such perceptual data.

The structuring of the survey provided respondents with alternatives to each question in a Likert-type item. The current items of the questionnaire were captured in three latent themes (the individual, the group, and the organisation). Clustering items according to such themes in scale development is adequate because, according to Bejar (1983), dimensionality is situation-specific. This means that dimensionality is not purely a property of the items itself, but rather a response to the items under a specific set of conditions. This approach resulted in an accurate assessment of the latent structure (see Section 5.2.4). Therefore, it can be said that a response set provides better data when it remains in its natural thematic setting, as opposed to being randomised (Wilson & McClean, 1994).

Alternatively, a review of the literature revealed that there are as many reports of no or trivial order effects as there are of significant or important order effects – "[a]t present, therefore, the frequency, size, and nature of question-order effects in standard surveys of the general population are matters of considerable uncertainty" (Schuman & Presser, 1996:24). The decision to maintain underlying themes from the order of items for this research was based on the original intent of organisational behaviour analysis, which implies measurement at three distinct levels. Furthermore, maintaining specific themes or dimensions within a questionnaire is based on the premise that the groupings themselves admit items that are *only peripherally* related to the underlying unitary theme. Fundamental to item theory is the notion that psychological constructs are "latent" (Meijer & Baneke, 2004:354). The perceptions of these constructs can only be obtained from the manifest responses from participants to a set of items. According to Meijer and Baneke (2004), the structure of a research questionnaire assumes the existence of a latent trait on which persons and items have an opinion or take a position. In the current study, this implied the need to group items in line with the assumption of the existence of latent themes (traits) prior to a factorial data exploration. Randomisation of items may have dissolved the assumed structure. Item clustering then provided a more accurate description of what the variables were actually doing and, more specifically, acknowledged the nature of organisational behaviour theory as *substantive*. In addition, it will be observed that the results of the factor analysis (see Section 5.2) revealed a latent underlying structure of the items which themselves correlated across latent themes. Factor analysis was the *statistical* method of choice, which determined that clusters of items were actually related to one another. Furthermore, according to Goldstein and Wood (1989:164), "unidimensionality in the presence of multidimensionality will produce a composite dimension".

The item grouping choices made for the purposes of the current research can be summarised according to the position of Schuman and Presser (1996), who argue that grouping similar questions together presents a smoother organisation of the questionnaire and appears sensible or coherent to respondents. The negative effects of any ordering sequence are far too inconclusive to warrant a randomised set of questionnaire items. In addition, many scholars are fairly confident that major findings in their research were not due to response order effects as such in any case (Schuman & Presser, 1996).

4.7 STRUCTURE AND LAYOUT OF THE QUESTIONNAIRE USED IN THE STUDY

The overall structure and layout of a questionnaire has been known to influence the responses participants are willing to give, as well as the overall response rate (Cooper & Schindler, 2003). An introductory letter (see Appendix B) was therefore attached to each paper-based questionnaire, and a similar introductory letter preceded the web-based survey (Appendix F). The main body of the questionnaire was highly structured. This entailed that alternatives were provided to the respondents, who had the simplified task of marking only the appropriate answers. The purpose of the questionnaire was to elicit data from the sample with regard to their demographic particulars and their perceptions of, or attitude towards, advanced automated aircraft training. The questions found in most of the questionnaire were closed-ended, and took the form of Likert-type items. According to Babbie (2010:256), closed-ended questions can be "easily" processed and provide for better uniformity of responses, as opposed to the alternative, which is open-ended questions.

Table 15 depicts the layout of the final questionnaire (AATC-Q).

Table 15: Questionnaire structure

Section	Topic of section	Number of questions
A	Demographic information	22
B	Perceptions of the advanced automated aircraft training climate	42
C	Participants' comments and feedback	2
Total number of questions		66

Section A consisted of questions related to the demographics of each participant. Specific questions referred to the person's age, gender, educational qualifications, levels of experience as a pilot in terms of years and hours, type of aircraft operated, perceived level of computer literacy, and whether the person had enjoyed his or her most recent flight simulator and route training experience.

The airline pilot's experiences, opinions and perceptions of their training were then gauged in Section B. Each perception statement was presented as a seven-point Likert-type item. The items were also clustered according to the level of measurement at a micro, meso and macro level. To limit any response bias associated with clustered items (see Section 4.6.4), the various categories of analysis (micro, meso, macro analysis) were not indicated to the respondents in the questionnaire itself.

Section C provided an area for the participant to interact with the researcher if the participant wished to do so. Comments by respondents were recorded here. Textual data is an important source of information collaboration and can be used to verify or clarify ambiguous findings. Participants were also given an opportunity to provide their e-mail addresses for future correspondence on the study results and to communicate any interesting recommendations. This option was intended to allow the possibility of providing feedback and close the knowledge loop (closure).

4.8 LEVELS OF MEASUREMENT

Many researchers have had difficulty in deciding whether data extracted from items in a Likert-type design are "continuous, categorical or rank ordered" (Stevens, 1946:677). Clason and Dormody (2001) argue that it is highly probable that the summated items from a Likert-type designed questionnaire are ordinal or interval, and thus approximate a continuous scale. In addition, it is generally assumed that an ordinal or interval Likert-type item is continuous, because, according to Nunnally and Bernstein (1994), behavioural research scales measuring perceptions assume an approximately equal interval scale with considerable assurance. Since the study was intended to measure the perceptions of the automated aircraft training climate construct and its associated variables, the study was designed to measure airline pilots' attitudes by means of an interval scale (Likert-type, continuous data).

Likert (1932) originally constructed five-point items in a summated scale to assess survey participants' attitudes. However, Likert (1932) admitted that the number of intervals in the item might be open to manipulation, and subsequently no fixed number of intervals was recommended in the original Likert-type item. The confusion in the literature regarding Likert-type scales and Likert-type items still persists. Clason and Dormody (2001) point out that Likert's seminal work was not intended to develop a summated scale in the first place, although the questionnaire items appeared as a scale of some sort. With this in mind, in the current study, individual seven-point Likert-type statements were adopted in which a rating from 1 to 7 implies varying levels of disagreement or agreement with the statement (that is; strongly disagree, moderately disagree, slightly disagree, neither agree or disagree, slightly agree, moderately agree, strongly agree).

A level of measurement stems from the granularity of the items (number of intervals). High granularity is based on the assumption that respondents in a sample can discriminate fairly accurately due to their enhanced levels of experience in the given field (as was the case among the respondents in the current study).

Cooper and Schindler (2003:223), Morgan, Leech, Gloeckner and Barrett (2007:42), describe the characteristics of four different types and levels of measurement, in terms of “ratio”, “interval”, “ordinal” and “nominal”. However, variables’ levels of measurement were originally contemplated by Stevens (1946:678) to clarify and determine the nature of data. Such clarification improved computational quality by guiding the selection of appropriate types of statistics to be used to explore the data further. The details pertaining to the levels of measurement according to Stevens (1946) are:

- *Nominal* (categorical scale) measurement is used for the empirical determination of equality, as in gender (male or female). Permissible statistics for this measurement level are the number of cases, the chi-square, McNemar, phi or Cramer’s V, and discriminant analysis.
- *Ordinal* (rank ordered scale) measurement is used for the empirical determination of greater or lesser value, as in the perceived quality of training received (very good, good, average, poor). Permissible statistics for this measurement level are rankings, mean rank, median and mode. The Mann Whitney-U, Kruskal-Wallis, Spearman, rank order correlation or Kendall Tau are preferred measurement tests.
- *Interval* (continuous scale) measurements are used for the empirical determination of scores that are ordered from low to high in categories that are evenly spaced. For example, a summated Likert-type designed scale of which the items measure on a “strongly agree” to a “strongly disagree” continuous seven-point scale would be considered an interval level measure. Permissible statistics for this measurement level are mean, standard deviation, factor analysis, Student’s t-test, one-way analysis of variance (ANOVA), Pearson’s correlation, regression analysis, multiple regression analysis, factorial ANOVA and multivariate analysis of variance (MANOVA).
- *Ratio* (continuous scale) measurements are used for the empirical determination of the equality of ratios, as in most physical measurements, for example, age in years or hours of experience in advanced aircraft. These measures have equal intervals between the levels or scores and a true zero level. The permissible statistics used for each measure are cumulative, in other words, all operations discussed above can be used for calculations involving ratio type data.

4.9 RESEARCH POPULATION AND SAMPLING STRATEGY

The “universe” of elements in which a researcher happens to be interested is commonly referred to as a “population” (Butcher, 1966:3). Many scholars stress the importance of defining the population correctly, because doing so determines the level of the statistical accuracy of the final sample. For this reason, Cooper and Schindler (2003:181) propose that the “ultimate test of a sample design is how well it represents the characteristics of the population it purports to represent. In measurement terms, the sample must be valid”. The validity of a sample is highly dependent on its accuracy (absence of bias) and precision (degree of error).

The target population for this study consisted of individual persons, in particular, qualified South African airline pilots who have some level of experience with advanced commercial aircraft.

4.9.1 Determining the sample size

Determining a sample size that makes it possible to extract sufficient data for statistical analysis can be a difficult exercise for researchers. In determining the most suitable sample size, three criteria are usually specified (Kalton, 1999):

- the level of precision required (for the social sciences, an acceptable level of error is 3%);
- the level of confidence or risk accepted (in social sciences research, an alpha level of 0.05 at the *a priori* level is acceptable where the *ex post facto* effect size is evaluated); and
- the degree of variability in the attributes being measured (designed using Likert-type items, provided a level of continuous data).

Bott and Svyantek (2004) have suggested two fundamental reasons for ensuring an accurate sample size for conducting scientific research:

- a minimum number of cases is required to analyse sub-group *relationships* adequately. For factor analysis (discussed in Section 4.17.5) around 200 cases are required if several items are used to define each construct); and
- in order to draw associational and comparative conclusions, the sample must, as far as possible, *represent* the population under scrutiny.

Two separate sampling procedures were conducted during the study. The first step called for experts to provide statistical validation of the questionnaire items of the training climate dimensions and their descriptive elements, which were initially identified theoretically. For a classical statistical analysis of expert judgements, Lawshe (1975) strongly suggests a minimum of 15 panellists for quantitative validation.

The second part of the research relied substantially on the results of an exploratory factor analysis. It was noted that DeVellis (2003) claims that a large number of unspoilt returns (around 300) are required for a factor analysis to be reliable (in other words, to uncover dimensional clusters). There are, however, other opinions on this topic, so it was explored further to determine the most appropriate path to follow in order to obtain a workable sampling frame for the study.

An analysis of the literature presented conflicting and varying propositions on determining the most appropriate number of elements to provide a good sample (see Table 16). For instance, Stoker (1981) suggests that the sample size should be proportional to the number of elements contained in the population size (N), whereas Welman and Kruger (1999) argue that, irrespective of the size of the population, it is not necessary to use a sample larger than 500 units for the analysis. This suggestion is in line with findings reported by Gravetter and Wallnau (2008), who have demonstrated that the standard error in a sample size is reduced exponentially, and not in a linear fashion. Therefore, the standard distance between a sample mean and the population mean tends to be reduced with larger samples, although it will never drop to zero even for extremely large samples.

Table 16 synthesises some important authors' sample size requirements for the development of a valid and reliable psychometric measurement instrument. It is, furthermore, important that the sampling method adopted be reported accurately, so that readers can draw their own conclusions (Bartlett, Kotrlik & Higgins, 2001). It is clear from the comparison of methodologies that a sample of 200 to 300 observations is adequate to provide a stable factor solution for the instrument. Comrey and Lee (1992) suggest that 200 elements in a sample is a fair to adequate number for obtaining relatively stable solutions.

Table 16: Contrasting notions of what constitutes a good sample size

Source	Recommendation
Stoker (1981)	Proportional to /N (for example, when N=1000, minimum sample size=141)
Arrindell and Van der Ender (1985)	20 times the number of factors
Comrey and Lee (1992)	100:poor; 300:good; 1 000:excellent
Nunnally and Bernstein (1994)	10 observations per variable
Welman and Kruger (1999)	Not necessary to have more than 500 observations
DeVellis (2003)	At least 300 observations required to conduct factor analysis
Netemeyer *et al*. (2003)	5 to 10 observations per parameter estimated
Pett *et al*. (2003)	10 to 15 subjects per item
Tabachnick and Fidell (2007)	A minimum of 300 cases
Saunders, Lewis and Thornhill (2007)	278 cases in 1 000 will provide a 5% margin of error

Alternatively, using Cochran's (1954) sample size formula for scales based on seven-point Likert-type items, Bartlett *et al*. (2001) calculated that for a finite population of around 1 400 elements (which was the target population for the current study), the required sample size was only 118, and corrected to 111 (when the sample size exceeds 5% of N).

Oversampling may be necessary when return rates are expected to be low, as is typical with survey research of this nature. Therefore Bartlett *et al*. (2001) determined that for a population of 1 400, it may be necessary to send out a minimum of 171 questionnaires if the anticipated return rate is estimated at 65%. With this in mind, in addition to the electronic means used for data collection, a total of 700 hardcopy surveys were distributed to ensure a good response rate. This suggestion also requires a sample of no less than 5% of the population for acceptable accuracy and precision in social sciences research.

Stoker (1981) and, more recently, Streiner (2003) recommend that researchers bear in mind the three boundary constraints when considering the size of the sampling frame (that is, level of precision, confidence interval and degree of variability). The main concern for researchers determining the ideal sampling size for an exploratory factor analysis stems from sampling error (Osborne & Costello, 2004). However, when item communalities (the amount of variance explained by common factors) are relatively high (0.6 and above, as was the case in the present study's data set), sampling error is somewhat reduced, and an exploratory factor analysis produces a fairly stable factor solution using smaller frames of between 200 and 300 elements (MacCallum, Widaman, Zhang & Hong, 1999). Moreover, in their analysis, Osborne and Costello (2004) found that neither the number of variables nor the size of *N* had any significant unique effect when all other variables were kept constant. The levels of variable communalities were high in the present study because the scale development began with content validation from subject matter experts.

4.9.2 Sampling frame based on the response rate

The actual cohort of the final sample in the current study was related to the response rate. The response or return rate is usually expressed as a percentage (the ratio of the number of questionnaires sent out divided by the number of usable questionnaires returned). A multitude of factors may influence the response rate. Haworth (1996) suggests that around half the final response will be obtained without the need to send participants a reminder. Another third of the responses can be obtained from a first reminder. This technique was used to elicit additional returns from participants in the current study too. The technique resulted in a response rate of approximately 33%. In social sciences research similar to that in the current study, the average response rate was found to vary around 30% (Osborne & Costello, 2004). Therefore the response rate for the present research was satisfactorily typical.

After reviewing the results of a rigorous versus a standardised survey methodology, a response rate of 33% was not completely disappointing. To answer the question of "[w]hat differences arise in point estimates subject to different response rates", Keeter *et al*. (2000) compared two surveys. In their study, they completed two surveys: a rigorous survey conducted over five days, which obtained a 60.6% response rate, and a standard one, which obtained a 36% response rate (in other words, nearly half the rate obtained in the five-day survey). Perhaps surprisingly, Keeter *et al*.'s (2000) study found that, despite the differences between the two survey responses or return rates, both achieved very similar statistical results. Their survey with a lower response rate was only minimally less accurate than its more rigorous counterpart. However, it was nonetheless borne in mind that variances increase when samples are smaller than the target number of minimum returns (Bartlett *et al*., 2001).

In summary, Haworth (1996) provided some very important methods to obtain good response rates and reduce non-response bias. These include

- concentrating on the design of the questionnaire (careful layout);
- maintaining a logical ordering of questions;
- clear phrasing of statements, combined with an attractive presentation; and
- endeavouring to keep participants interested in the topic to elicit greater participation.

4.9.3 Sampling procedure

According to Kalton (1999), in order to conduct replicable scientific research, it is necessary to clearly state and implement definitive statistical reasoning when selecting only some elements from a population. Therefore, in order to draw conclusions, make inferences or devise theories about a population, one must have a mathematically sound basis. Researchers usually use two fundamental sampling scheme categories, first, random or probabilistic, and, second, non-random or non-probabilistic (Cooper & Schindler, 2003). A probabilistic method requires that each element of the population frame have an equal chance (a non-zero probability) of being selected for inclusion in the final sample. This method requires an accurate list of the elements in the population and can prove expensive. A non-probability method was used for this study, based on the fact that a list of the entire population was unobtainable. A guideline on the actual numbers of eligible pilots was, however, obtained from both the Civil Aviation Authority and the Airline Pilots' Association of South Africa. These numbers were used to determine an appropriate size for the sampling frame.

The judgement, quota, snowballing or convenience sampling methodologies are examples of the most common non-probabilistic methods used in similar research (Creswell, 2002). A purposive judgemental method was used in the current study, based on the guidelines, using the pilot numbers from the Civil Aviation Authority and Airline Pilots' Association. In addition, after interviewing and then using the judgement of experts who are particularly knowledgeable about the field and phenomena under study, due consideration was given to the systematic inclusion and exclusion of certain elements from the population, as recommended by Babbie (2010). In order to extract a representative sample for the South African situation, the population was stratified according to the various major airline companies based in the country. Cooper and Schindler (2003:193) describe such a stratification method as partitioning the population into mutually exclusive "sub-populations or strata".

The primary unit of analysis was the perceptions of *airline pilots*; hence, the target population consisted of only those South African airline pilots who held a current licence to operate advanced automated aircraft at the time of the survey. According to the figures provided by both the Civil Aviation Authority and the Pilots' Association of South Africa, the population was estimated at approximately 1 400 pilots.

A non-probability method was used to gain access to a convenient sample. Questionnaires were purposefully distributed to the stratified groups of individuals in accordance with Haworth's (1996:47) suggestion. The probability of selecting a particular entity from the sub-population for the sample frame using this method was unknown in terms of the criteria proposed by Bott and Svyantek (2004). In other words, systematic randomisation (where each entity has a known non-zero chance of selection) was not obtained. According to Kalton (1999), non-representativeness is a distinct disadvantage when using such sampling techniques. Obtaining the required sample size by targeting elements in a stratum of interest offsets some of the disadvantages found in the non-probability sampling technique and provided a level of control and precision, as described by DeVellis (2003). In this case, elements in the population of interest (airline pilots) could be regarded as highly homogeneous by nature, with very little significant variation in opinion, as was the case in a prior similar study (Naidoo, 2008). This premise also reduced sampling error in the final sample frame, because "how large a sample should be is a function of the variation in the population parameters under study" (Cooper & Schindler, 2003:190).

The non-probability, convenience and purposive *stratified sampling* technique (Cooper & Schindler, 2003) used in this study entailed dividing the population into several strata or groups. Stratification is the process of partitioning members of the population into relatively homogeneous subgroups before sampling (Kalton, 1999). In this case, the homogeneous strata were based on the specific airline company to which each element belonged. Saunders *et al*. (2007) suggest that convenience sampling be used when there is very little variation in the population, as was the case in this target population.

The population itself was deemed to contain little variation, because it is common knowledge that all airline pilots employed at major carriers are selected only after a battery of tests, and after complying with the certification requirements stipulated by the South African Civil Aviation Authority (SACAA). These tests serve as a filtering mechanism for each organisation to ensure that only those candidates who fit the corporate culture of the particular airline are selected. Hence, it was reasonable to assume that the *source* of the data was limited to a fairly homogeneous cross-section of qualified airline pilots flying advanced automated aircraft in various South African airlines.

In order to target specific strata, a number of airline organisations were also approached for assistance in maximising the response rate. These organisations are regarded as the largest airlines in South Africa and fit into the Airline Pilots' Association of South Africa (ALPA-SA) portfolio, namely (also see Table 17):

- South African Airways (SAA);
- British Airways Comair (BA Comair);
- South African Express Airways (SAX);
- Mango Airlines (Mango);
- South African Airlink (Airlink); and
- 1Time Airlines (1Time).

4.9.4 Stratification in terms of airline pilot unionisation

To explore other stratification options, pilot unionisation was considered, because, amongst the airline pilot group in South Africa, unions play a major role in organisational perception. Also, to maintain some level of anonymity for the organisations under study, it was decided to partition the six participating airlines into groupings according to whether the pilots were unionised or not. The major carriers in South Africa can easily be separated into organisations, which have large numbers of unionised pilots (membership of more than 60% of the pilots employed at the organisation) on the one hand, and those which do not on the other hand. The population was partitioned in this manner to allow for easier categorical comparisons. Airline pilots tend to gravitate towards those organisations that boast larger numbers of unionised members due to perceived improved working, training and safety standards (Walsh, 1994). Such perceptions may also have a significant influence on opinions of the training and overall organisational climate (Olney, 1996).

Airline pilot unions are considered separately from traditional industrial unions. They are generally considered professional bodies and are regarded more in terms of an association (ALPA-SA, 2011). In the case of some smaller airlines, airline pilots may be represented by large industrial unions, such as Solidarity. Be that as it may, the current numbers of job applicants are higher at airline companies with ALPA-SA membership, and labour turnover at the airlines without such representation is higher. In South Africa, the legacy airlines report the largest number of unionised members (ALPA-SA, 2011). At the two oldest and largest airline companies in South Africa (SAA and BA Comair), at least 99% of the pilots are unionised. Higher salaries, pension and provident funds, coupled with a significantly better safety record and a non-punitive organisational culture appear to be the primary attraction (Walsh, 1994). Olney (1996) postulates that structured unification of professional employees improves training standards and subsequently organisational climates, because many professional associations regard themselves as an integral part of efficient enterprises.

According to ALPA-SA (2011), currently, half of South African airline organisations are unionised and half are not. SAA, SA Express and BA Comair account for the bulk of the unionisation, while the smaller carriers – South African Air Link, Mango Airlines and 1Time Airlines – are non-unionised companies.

4.10 BASIC DEMOGRAPHIC INFORMATION ON THE FINAL SAMPLE

Aaker, Kumar and Day (1995) propose that the representation of the population within the sampling frame has more significance on post analytical results than the actual response rate in itself. In addition, Cooper and Schindler (2003) also point out that relying on sheer magnitude from numbers, would not guarantee a representative sample. A primary disadvantage from using a convenience sampling method, is that population representation within the sampling frame is compromised. However, by targeting specific sections of the population of interest, selection bias was to a certain degree, mitigated.

The decision to utilise an Internet based survey method resulted in a level of unavoidable under coverage of the target population, leaving certain demographics underrepresented. It was hoped that by using a hybrid data collection method (paper-based and web-based surveying), the adverse effect of the Internet for surveying, would be reduced.

Table 17 shows that, in general, the population was well represented. The majority of the participants in the sample frame (48.7%) are positioned within the organisation employing the largest number of airline pilots in South Africa (approximately 800 pilots). The concept of representation is especially important when a stratified sampling method has been adopted, as in the case of the current study.

In addition, the desired categories were well represented, apart from gender (see Table 17). The distinct inequity in the distribution of male and female pilots is nonetheless an accurate reflection of the current status of the aviation industry, as there are very few female airline pilots. Previously, South African legislation prevented potential female candidates from pursuing a career in aviation, but change, albeit slow, is now occurring at many airlines. However, because the current study was not focused on gender issues or gender phenomena as such, the skewed distribution within the gender category was not regarded as an aggravation in terms of the analyses of results.

Table 17: Respondent sample frame (N=229)

VARIABLE	FREQUENCY	PROPORTION	MEAN (S.D)
ORGANISATION			
1 (SAA)	112	48.7%	
2 (BA Comair)	23	10.0%	
3 (SAX)	14	6.1%	
4 (Airlink)	34	14.8%	
5 (Mango)	11	4.8%	
6 (1Time)	10	4.3%	
7 (Other)	25	10.9%	
SIZE OF AIRLINE COMPANY			
Large (1+2)	135	58.5%	
Medium (3+4)	48	21.4%	
Small (5+6+7)	46	20.1%	
MAIN AIRCRAFT MANUFACTURER			
Boeing	57	24.9%	
Airbus	95	41.5%	
Other	77	33.6%	

Table 17: Continued

VARIABLE	FREQUENCY	PROPORTION	MEAN (S.D)
GENDER			
Male	212	92.6%	
Female	17	7.4%	
AGE (years)			41.28 years (11.359)
Below 30	38	16.6%	
30 – 40	81	35.4%	
41 – 51	56	24.5%	
52 – 63	51	22.3%	
Above 63	3	1.3%	
EDUCATION LEVEL			
No tertiary education	131	57.2%	
Tertiary education	98	42.8%	
INSTRUCTOR RATED			
No	102	44.5%	
Yes	127	55.5%	
FLYING EXPERIENCE (hours)			9 753.3 hours (6116.719)
Below 2000	7	3.0%	
2 001 – 5 000	58	25.3%	
5 001 – 7 000	30	13.1%	
7 001 – 10 000	39	17.0%	
10 001 – 15 000	57	24.9%	
Above 15 000	38	16.6%	
COMPANY STATUS			
Captain	120	52.4%	
Co-pilot	109	47.6%	
COMPUTER LITERACY			
Poor	5	2.2%	
Average	87	38.0%	
Above average	92	40.2%	
Excellent	45	19.6%	
INITIAL TRAINING			
Military	81	35.4%	
Airline cadet	18	7.9%	
Self-sponsored (part-time)	64	27.9%	
Self-sponsored (full-time)	66	28.8%	

Table 17 clearly shows that in terms of the general flight experience levels of the group, the sample was fairly well distributed, with the majority of respondents above the 5 000 hour mark (Mean=9753.29; SD=6116.719). However, the dispersion of the participants in terms of flight experience was large – the majority of the sample had between 3 000 and 16 000 flight hours. The high standard deviation of this descriptor is a testament to the heterogeneity of pilots found in the South African airline industry. Most of the pilots with the national carrier regard their present organisation as the final step in their career progression and some will retire after spending almost 40 years there (ALPA-SA, 2011). This is a further indication of high levels of industry experience. This was expected, as the target population were all qualified airline pilots operating advanced aircraft. Airline companies operating such aircraft tend to hire very experienced pilots.

The experience of the group can also be reflected in the mean age of 41 years (SD=11.359). The distribution of the participants' ages ranged from the mid-20s to the late 60s, indicating that, in terms of generational analysis, the airline pilot group is a fairly disparate one. This provided a good area for further statistical analysis, which was then undertaken as reported in Chapter 5 where the age category was sub-divided or combined as required, for an in-depth exploration of the relevant phenomena.

In general, most of the respondents (59.8%) perceived their levels of computer literacy as better than average. It may be hypothesised that by virtue of the fact that participants operate relatively superior machinery, their presumed technological acumen becomes pervasive. Secondly, one major South African carrier provides company laptop computers to its pilots, therefore possibly facilitating improved perceptions of computer abilities and skill within the target group.

The airline organisations in South Africa were further categorised in terms of size. The size of an organisation is generally determined from the sheer number of employees, the market reach or market share it enjoys and the extent of its operations (Desler, 2002; Drucker, 1946). According to Pitfield, Caves and Quddus (2010), a large major carrier is described as operating a fleet of aircraft with the company brand and identity in terms of the ICAO (International Civil Aviation Organisation) or IATA (International

Air Transport Association) code. The major airline also has a unique call sign associated with it. For instance, South African Airways has the call sign "Springbok", whilst British Airways has adopted the call sign, "Speedbird" (IATA, 2012). However, in order to define the various airlines in South Africa in terms of being either large or small, it was considered whether the organisation operates at least one fleet of more than 10 aircraft, which is capable of carrying more than 99 passengers upon its national operating certificate (Child, 1973). As a middling category however, it was necessary that the well-known regional carriers be positioned in the medium size airline group (SA Express and SA Airlink).

Most airline pilots employed at the largest organisation (in this case, South African Airways) operated Airbus-manufactured advanced aircraft (41.5%), which was reflected in the skewed proportions regarding aircraft type and manufacturer category. Appropriate non-parametric methodologies were subsequently employed (see Chapter 5) in the data analysis to understand and further explore the phenomena associated with the aircraft type sub-groupings. Employing more robust statistical methods (non-parametric procedures) mitigated the impact of any adverse effect emanating from the fact that only 24.9% of the participants indicated that they operated Boeing-manufactured advanced aircraft.

Instructors (non-rated or rated), level of education (tertiary or no tertiary) and company status (captain or co-pilot) were relatively well balanced, providing for good comparative examinations later in the thesis.

4.11 DATA COLLECTION PROCEDURES

Cooper and Schindler (2003:87) define data as "the facts presented to the researcher from the study's environment". There are many methods to extract raw data from the field. Such methods include, but are not limited to, questionnaires, standardised tests, observational forms, laboratory notes, and instrument calibration logs (Cooper & Schindler, 2003). Alternatively, collection methods for large-scale surveys include electronic mailing (e-mail), internet-based e-survey submissions (for improved response rates), together with traditional paper-based questionnaires (Cobanoglu *et al*., 2001).

Because empirical research requires data to be collected, in this case, first from a group of subject matter experts, and thereafter from a number of respondents in the target population, it was decided that a structured self-administered questionnaire would be used in both cases. The description, design and administration of the subject matter expert questionnaire are discussed later in Section 4.13. The advantages of self-administered questionnaires, as described by Cooper and Schindler (2003), include the benefits of expanded geographic coverage, minimal staff requirements, and the use of complex instruments, allowing respondents time to think about questions. The greatest disadvantage, however, was a low response rate (apathy).

Apart from the conventional survey distribution methods currently used in the airline industry, such as box dropping (personal letter boxes), the assessment instrument (AATC-Q) was administered to the sample population via the distribution channels used at ALPA-SA, namely its web page and e-mail contact list. Both the Association's executive committee and the different airline management groups graciously offered their assistance to maximise the response rate. Correspondence regarding the goals and intentions of the research project was communicated directly through email, telephone and one-on-one interaction with both airline management and pilots' association executives, so as to gain the necessary support and endorsement of the present study. In order to ensure an adequate response rate and a greater number of unspoilt returns, the instrument was also hosted on the World Wide Web as a dedicated e-survey that replicated the hardcopy questionnaire.

A cover letter explaining the purpose of the survey (see Appendix B), together with a note of the endorsement from both ALPA-SA and the company's management, accompanied each questionnaire in an attempt to entice participation and therefore improve the response rate. For data collection purposes, both the expert group and the target population were nonetheless readily accessible to the research team.

Subsequent to the development of a draft of the aforementioned large sample survey instrument (the questionnaire), the validation of a pool of items (constructed after an in-depth literature review) was analysed by a purposive group of subject matter experts. An expert in the area of modern advanced automated aircraft training is defined, for the purposes of this study, as an academic experienced in the field of aviation management, or a highly experienced flight instructor (with an advanced licence rating). It is generally accepted in the aviation industry that there is a positive correlation between total flying time and mastery of skill (Sherman, 1997; Telfer & Moore, 1997). To contrast the construct validation, it was necessary that a proportion of the expert panel be current academics (advanced educational credentials) in the field of interest. Therefore each subject matter expert was either an academic in the field, or held many thousands of hours flying instructional experience on advanced automated aircraft. The next section reports on this item validation process.

4.12 CONTENT VALIDATION

The quality of a perception measurement instrument rests on the level of validity in its content (Cooper & Schindler, 2003). The first step in the research plan required the use of expert opinion in refining the derived questionnaire items and thereby obtaining a valid content that could operationalize the construct of interest. It is a challenge to ascertain the content validity of a measurement scale based on the opinions of experts in the field (Landis & Koch, 1977) – Hardesty and Bearden (2004:98) suggest that there is “a lack of consistency and guidance regarding how to use the expertise of judges to determine whether an item should be retained for further analysis in the scale development process”.

After obtaining sufficient data from the judges, there were two areas of potential inaccuracy, which may have affected the quality of the measurement scale. The first of these inaccuracies stemmed from potential *inter-observer bias*, which consists of differences between the marginal distributions of the response variable associated with each of the observers (Altman, 1991; Fleiss, Levin & Paik, 2003; Karlsson, 2008). Cochran’s Q-test, a test in the analysis of variances, was subsequently used to test the hypothesis that inter-observer bias was absent.

The second inaccuracy was *observer disagreement*, which reflects the fact that observers may classify individual items in the same category of the measurement scale. Karlsson (2008) suggests that the computation of the Kappa test statistic coefficient can be used to determine the level of inaccuracy associated with the data set. However, an alternative method based on the value of a ratio calculated according to Lawshe's (1975) formula (discussed in this section) was used to determine the level of agreement between the judges' categorisation of items in the current study.

To determine whether the measure's items actually capture a proper sample of the theoretical content domain, opinions from experts and/or inter-rater agreement were sought, in line with Karlsson's (2008) suggestion. In order to gain a representative sample of the content domain of the unobserved construct of interest, judgements regarding whether possible items may actually represent the intended construct were then validated. According to Hardesty and Bearden (2004), there is some confusion between face validity and content validity, which are terms that are also, to some degree, used interchangeably in the literature. Some authors suggest that expert opinion primarily evaluates *face validity.*

According to Fleiss *et al*. (2003), the *content* validity of a measure can then be validated indirectly through a statistically significant inter-rater agreement calculation. Hence, for the purposes of meeting the research objectives, the first phase of the study evaluated the statistical significance of inter-rater agreement as an indication of the content validity of each item's relationship with the construct and sub-constructs, this was calculated using Lawshe's (1975) method and Cochran's Q statistic.

In an attempt to gain a more in-depth understanding of how the research process acquired content validity, an analogy was constructed. The universal domain of the construct under study is represented by the contents (universe of acceptable items) entering a funnel (which represents the inter-rater or expert judgement filtering procedure). In order to obtain a proper representation of the main construct of interest, items are hypothesised to belong to one of three sub-constructs (consisting of specific item clusters). This was based in accordance to Hardesty and Bearden's (2004) premise that a measurement scale (measuring the main or super-construct) would not have the required content validity if its items accounted for the variability in only one exclusive sub-construct. Hence, if items appeared to fall into the opening of the analogical funnel, it would have face-validity. In other words, according to the expert judgement method, these items were actually measuring the main construct at some level. With this analogy it is easy to imagine how a different researcher measuring the same construct of interest, may obtain different indicators (items) to the ones discovered in this specific study. It is possible that different items can measure (tap) the same construct, because an infinite number of indicators manifest in one variable space.

The purpose of expert validation in this study was therefore to ensure that the items in the initial pool reflected the desired main hypothesised construct of interest. Eventually, after Lawshe's method of item analysis, and a statistical assessment of inter-rater bias, the final item pool consisted of fragments from the universal domain of available items (see Figure 20). The aim of the filtering phase in this scale development was to ask subject matter experts to judge whether,

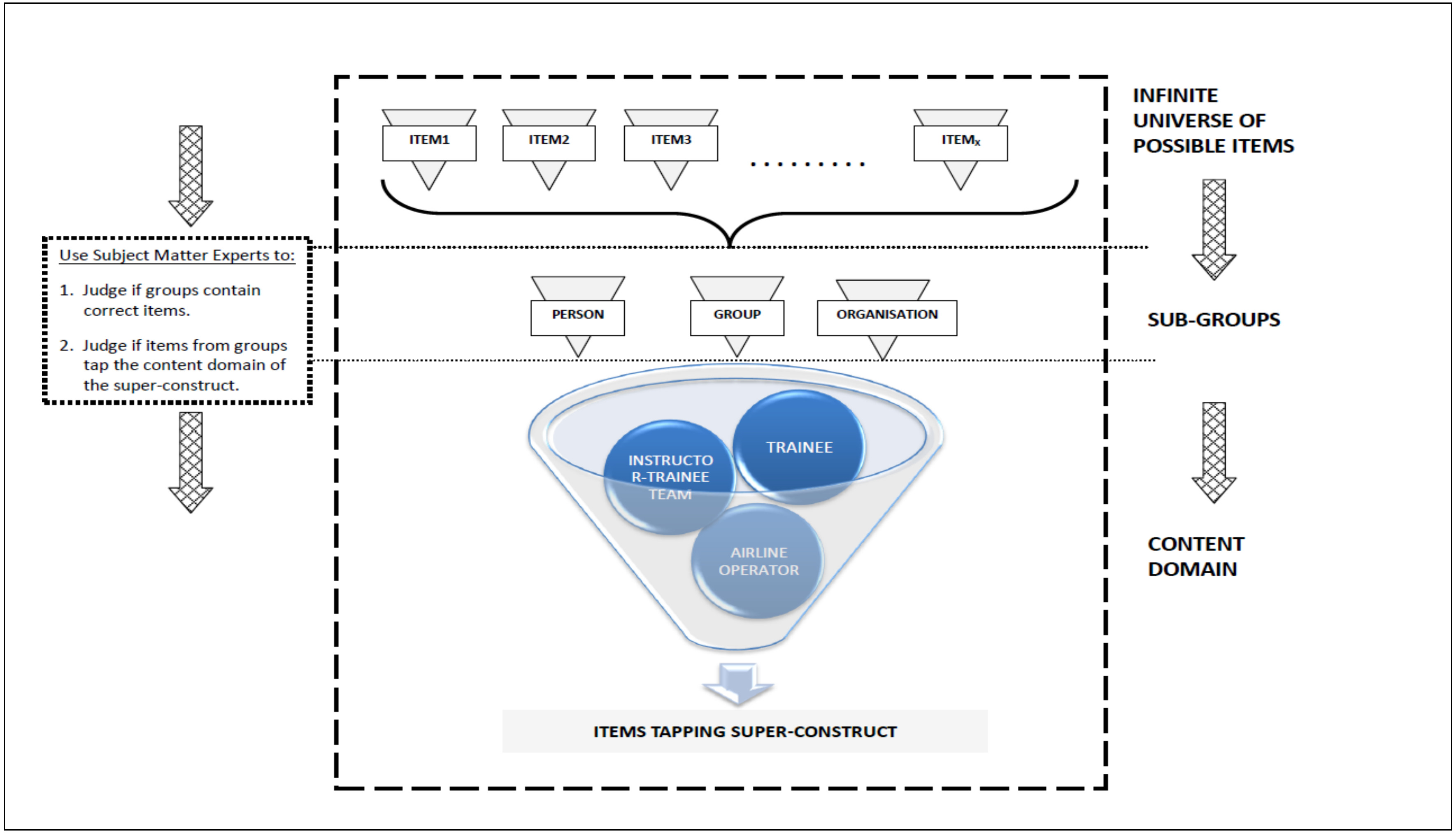
- groups contained correct items; and
- items from groups tapped the content domain of the super-construct.

Analysing expert judgment is a validity process undertaken before data collection, therefore "the development of a new measurement instrument is [generally] a two-stage process" (Karlsson, 2008:110). Altman (1991) warns that there are many problems in causally determining levels of agreement between judges during initial scale development. Inter-rater agreement is a process rife with systematic error. It was found that many researchers conducting scale development use associational statistics incorrectly in an attempt to obtain content validation (Landis & Koch, 1977), therefore it was necessary to also check the level inter-rater bias post-validation.

Fleiss *et al*. (2003) found that the percentage of agreement between judges or correlations determined using a Pearson's coefficient could be highly misleading. These and other reasons prompted the pursuit of a more robust content validation method. To develop a valid and reliable scale, it was decided that the technique proposed by Lawshe (1975) was the optimum solution in the initial stages of scale development.

Figure 20 was developed to propose an analogy that illustrates the process followed to validate the content of the hypothesised construct. Content validation was deemed an important early step in the study, as it created the foundation for subsequent data collection, analyses and final discussion of phenomena.

Figure 20: Content validation analogy


Use Subject Matter Experts to:
1. Judge if groups contain correct items.
2. Judge if items from groups tap the content domain of the super-construct.
ITEM1
ITEM2
ITEM3
ITEMx
INFINITE UNIVERSE OF POSSIBLE ITEMS
PERSON
GROUP
ORGANISATION
SUB-GROUPS
INSTRUCTOR-TRAINEE TEAM
TRAINEE
AIRLINE OPERATOR
CONTENT DOMAIN
ITEMS TAPPING SUPER-CONSTRUCT

Source: Author

The level of agreement between subject matter experts was based on Lawshe's (1975) method of content validity because the method is regarded as *mathematically* sound. Judges were asked to determine how "essential" an item cluster is to a specific sub-construct representing the content domain. Independent views were elicited from the experts by asking each expert to respond to the following question in terms of the measurement of the hypothesised construct: "Is the knowledge measured by this item cluster: essential, useful but not essential, not necessary?"

Lawshe (1975:567) developed the following formula for the computation of the minimum content validity for different panel sizes based on a one-tailed test at a significance level of α = 0.05:

Content Validity Ratio (CVR) = $(n_{e\,-}N/2)/(N/2)$,

where:

n_e = number of experts indicating "essential"; and

N = total number of expert panellists.

Lawshe (1975:566-567) suggests that a minimum CVR value of "0.49 is required from 12 to 15 subject matter experts" to ensure that agreement is unlikely to have been due to chance. Alternatively, Fleiss *et al*. (2003) suggest that the statistical value of a Kappa coefficient would also confirm significance (this method was not pursued in the current study). For this study, 36 subject experts were approached to participate, and 17 usable sets of responses to the questionnaires were returned (a response rate of 47%, which was deemed fair, and therefore adequate for the analysis to continue [Streiner, 2003]).

A CVR value of 0.46 is required to obtain the necessary validity when using a panel of 17 experts (Lawshe, 1975). A more conservative cut-off point of 0.49 was however, subsequently used (see Section 4.14).

A non-exhaustive list of 106 items was generated from the literature review to hypothesise the operationalization of a model of the construct. Of the 106 items, 64 were deemed not essential or necessary for having some degree of content validity.

Thus, 39.62% of the original item list was retained after analysis of the opinions from the panel of subject experts. The next section describes these results in more detail.

4.13 RESULTS OF LAWSHE'S TECHNIQUE

A final cohort of 17 highly experienced airline flight instructors and university academics participated in the expert validation process. An instrument in the form of a survey questionnaire was developed to extract data from the sample of experts (see Appendix A). The instrument contained five main sections as follows:

- *Section 1:*
  This part of the instrument contained an introductory letter and information regarding respondent consent. It introduced the research to the expert and provided the contact details of the researchers.
- *Section 2:*
  This part of the instrument contained information about the background literature review on the topic of interest. More importantly, this section of the survey showed the expert respondent what the hypothesised model of the construct consisted of (as discussed in Chapter 3).
- *Section 3:*
  This part of the survey asked for the respondent's demographic information.
- *Section 4:*
  The important data collection statements were contained in this section of the expert survey. This part of the expert instrument was further divided into three dimensions. The first dimension (27 statements) solicited information about the organisational level (the airline) of the construct. The second dimension (27 statements) asked experts about their opinions regarding statements related to the group level of analysis (the instructor-trainee team). Finally, the third dimension (52 statements) in this part of the instrument solicited information about the individual level of analysis on the construct (the trainee), from the expert respondent.
- Section 5:
  The final part of the expert survey was qualitative in nature. Here, experts were

asked about their opinions regarding the clarity and comprehensiveness of the items.

The subject matter expert questionnaires were distributed electronically and in hardcopy format. A follow-up request was made to the experts after two weeks. Due to the length and depth of the subject expert questionnaire, it was difficult to convince participants to complete the request timeously. Of the 36 questionnaires distributed, 17 were returned, giving a response rate of 47%. This response rate is regarded as average for studies of this nature (Streiner, 2003).

Table 18 (demographic data) and Figure 21 (distributions) show that the mean age of the panel was 54.23 years (SD=7.64). The participants displayed a high degree of industry experience, with a mean of 30.65 years (SD=10.82). The mean instructional experience of the airline pilots was 3 780.64 hours (SD=2023.97), indicating a very high level of expertise in the subject.

A minimum of 15 panellists were required to attain a CVR of 0.49 in order to accept an item as essential in tapping the construct of interest. The distributions depicted in Figure 21 furthermore show clearly that the data are skewed. Skewed distributions in this context thereby confirm that the experts come from the tail of a normal curve. This was expected, as subject matter experts cannot be regarded as being in the same category as the average large survey respondent.

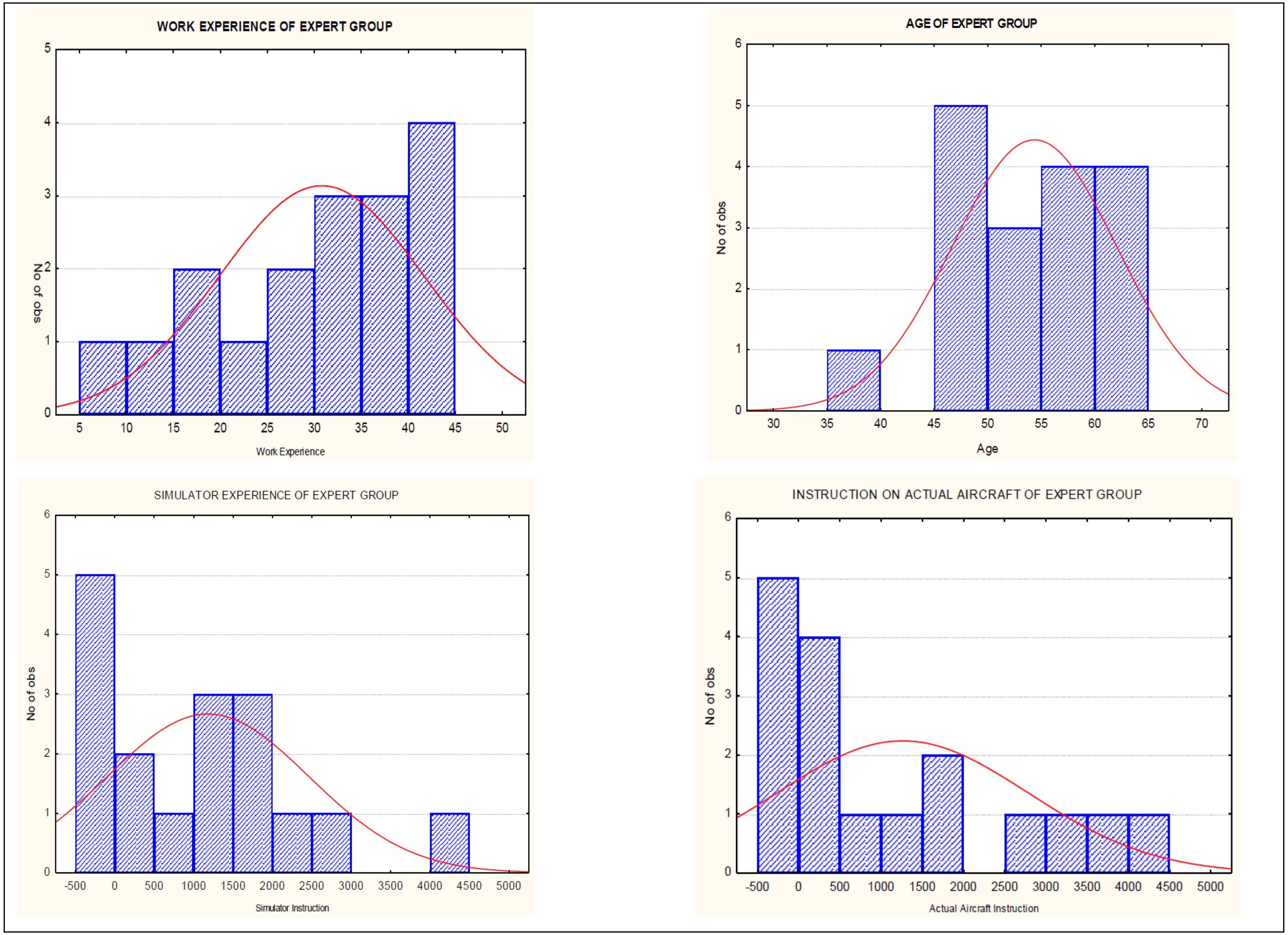
Table 18: Demographic data of the subject matter experts (N=17)

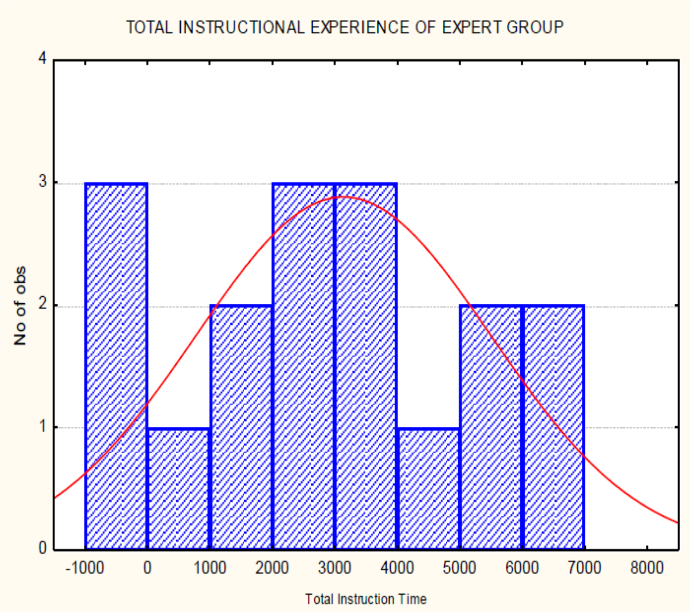
DEMOGRAPHIC VARIABLE	COUNT	PERCENTAGE
AGE (years)		
31-40	1	5.88
41-50	5	29.41
51-60	7	41.18
61+	4	23.53
INDUSTRY EXPERIENCE (years)		
10-14	1	5.88
15-20	3	17.64
21-25	1	5.88
26-30	2	11.76
31-35	4	23.53
36-40	2	11.76
41+	4	23.53
TITLE		
Airline Pilot (training instructor)	12	70.59
Academic	3	17.65
Airline Pilot and Academic	1	5.88
None (indicated)	1	5.88
HIGHEST EDUCATION ATTAINED		
Secondary School	5	29.41
Diploma	3	17.65
Bachelor Degree	2	11.76
Honours Degree	1	5.88
Masters Degree	2	11.76
Doctoral Degree	4	23.54

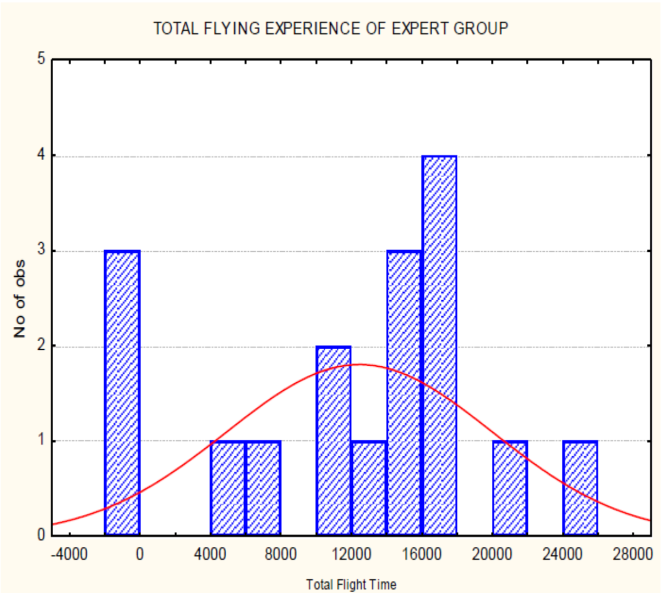
Table 18: Continued

DEMOGRAPHIC VARIABLE	COUNT	PERCENTAGE
FLIGHT SIMULATOR INSTRUCTION (hours)		
0-500	7	41.17
501-1000	1	5.88
1001-1500	3	17.65
1501-2000	3	17.65
2001+	3	17.65
ACTUAL AIRCRAFT INSTRUCTION (hours)		
0-500	9	52.94
501-1000	1	5.88
1001-1500	1	5.88
1501-2000	2	11.76
2001+	4	23.54
TOTAL FLIGHT TIME (hours)		
<5000	3	17.64
5001-10000	2	29.41
10001-15000	4	23.53
15001-20000	4	23.53
20001+	4	23.53

Figure 21: Distribution of subject matter expert demographic variables




TOTAL INSTRUCTIONAL EXPERIENCE OF EXPERT GROUP
No of obs
0
1
2
3
4
-1000
0
1000
2000
3000
4000
5000
6000
7000
8000
Total Instruction Time


TOTAL FLYING EXPERIENCE OF EXPERT GROUP
No of obs
0
1
2
3
4
5
-4000
0
4000
8000
12000
16000
20000
24000
28000
Total Flight Time

Tables 19 to 21 show the results of Lawshe's (1975) technique to assess content validity.

Table 19: Lawshe test results for Domain A

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
A1	Pilot training at my airline is in line with company goals.	17	0	1.000	Y
A2	My company's training produces world-class pilots.	15	2	0.764	Y
A3	I have noticed a steady improvement with regard to pilot training at this company.	11	6	0.294	N
A4	I know what my company's training goals are.	15	2	0.764	Y
A5	My company has talented people managing airline pilots' training.	15	2	0.764	Y
A6	Pilot training at this company is professional.	15	2	0.764	Y
A7	Management follows the regulator rules appropriately.	15	2	0.764	Y
A8	Pilot training on this aircraft is well organised at this company.	17	0	1.000	Y
A9	Pilots who are engaged in simulator training are professionally attired.	3	14	-0.647	N
A10	I understand what the company expects of me when I am in training.	16	1	0.882	Y
A11	It is easy to share my training experiences with colleagues at this company.	7	10	-0.176	N
A12	Training at my airline produces safe pilots.	16	1	0.882	Y
A13	There is a well-established chain of authority for pilot training on this aircraft.	12	5	0.411	N
A14	This airline gives its pilots an appropriate amount of preparation work before training.	13	4	0.529	Y
A15	The paperwork involved in training for this aircraft is appropriate.	11	6	0.294	N
A16	It is easy for me to appeal for assistance if I encounter a training problem at this airline.	16	1	0.882	Y
A17	There is sufficient training guidance from the company.	16	1	0.882	Y

Table 19: Continued

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
A18	**The standard operating procedures (SOPs) for learning to fly this aircraft are adequate.**	17	0	1.000	Y
A19	**The company provided me with sufficient time to prepare for training on this aircraft.**	17	0	1.000	Y
A20	**The simulators my company uses to train its pilots are in good condition.**	14	3	0.647	Y
A21	**I feel motivated by my airline to train for this aircraft.**	8	9	-0.058	N
A22	**The training department at my company is flexible.**	6	11	-0.294	N
A23	**The airline is very supportive of its pilots' learning requirements for this aircraft.**	16	1	0.882	Y
A24	**My company's culture supports training for new technology aircraft.**	16	1	0.882	Y
A25	**There is sufficient feedback about my training on this aircraft.**	17	0	1.000	Y
A26	**Pilot training at my airline follows civil aviation requirements.**	16	1	0.882	Y
A27	**My company uses only current training material.**	15	2	0.764	Y
	AVERAGE NUMBER OF ENDORSEMENTS	13.778	3.222		Total (Y)=20
	AVERAGE PERCENTAGE	81.047	18.953		74.074

Table 20: Lawshe test results for Domain B

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
B1	I find it easy to identify with my instructor.	11	6	0.294	N
B2	I can easily identify with my simulator partner.	8	9	-0.058	N
B3	I work well with others during simulator training exercises.	9	8	0.058	N
B4	Instructors communicate their expectations effectively.	11	6	0.294	N
B5	I learn better when I work as a member of the crew.	17	0	1.000	Y
B6	I am always at ease when interacting with my flight instructor.	9	8	0.058	N
B7	I always find my simulator partner prepared for training.	8	9	-0.058	N
B8	I trust my simulator partner.	5	12	-0.411	N
B9	I am confident that my instructor will be fair.	7	10	-0.176	N
B10	I operate well as a crew member in the simulator.	15	2	0.764	Y
B11	My instructor is willing to listen.	14	3	0.647	Y
B12	I communicate well with my simulator partner.	14	3	0.647	Y
B13	I feel secure in the decisions made by my simulator partner.	9	8	0.058	N
B14	I make good decisions with my partner in the simulator.	6	11	0.294	N
B15	I find that decision-making with my simulator partner is equitable.	9	8	0.058	N
B16	I am motivated by my instructor.	5	12	-0.411	N
B17	When training for this aircraft, I feel that I am part of a team.	12	5	0.411	N

Table 20: Continued

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
B18	**The instructors on this aircraft are committed.**	13	4	0.529	Y
B19	**Instructors are similar in how they teach pilots to fly this aircraft.**	16	1	0.882	Y
B20	**I am always paired with someone who is committed to performing well.**	10	6	0.176	N
B21	**I enjoy being evaluated as a member of a crew.**	3	14	-0.647	N
B22	**Instructors on this fleet follow company policy.**	9	8	0.058	N
B23	**The instructors on this aircraft avoid overloading pilots with unnecessary information.**	14	3	0.647	Y
B24	**I always bond well with my simulator partner.**	9	8	0.058	N
B25	**Decisions made in flight simulator training exercises are team-based.**	5	12	-0.411	N
B26	**The instructors on this aircraft are friendly.**	8	9	-0.058	N
B27	**I get sufficient feedback on my flight training performance.**	6	11	-0.294	N
	AVERAGE NUMBER OF ENDORSEMENTS	9.703	7.259		Total (Y)=7
	AVERAGE PERCENTAGE	57.076	42.924		25.926

Table 21: Lawshe test results for Domain C

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
C1	Pilots are in direct control of the training outcome.	17	0	1.000	Y
C2	A good training session on this aircraft is a result of the trainee's actions.	11	6	0.294	N
C3	Evaluation of my flight training is objective.	11	6	0.294	N
C4	Adequate preparation improves flight training performance.	16	1	0.882	Y
C5	I am always on time for a flight training session.	14	3	0.647	Y
C6	I co-operate well when training in a simulator.	13	4	0.529	Y
C7	I never feel rushed in the flight simulator.	12	5	0.411	N
C8	I easily express my opinion during flight training.	5	12	-0.411	N
C9	I prepare sufficiently for training on this aircraft.	11	6	0.294	N
C10	After flight training, I feel a sense of mastery.	16	1	0.882	Y
C11	I enjoy learning about this aircraft.	7	10	-0.176	N
C12	Simulator training affects behaviour on the actual aircraft.	10	7	0.176	N
C13	I get along well with my flight simulator partners.	11	6	0.294	N
C14	I found my transition to advanced automated aircraft easy.	5	12	-0.411	N
C15	I believe that if pilots do well in training, overall flight safety improves.	10	7	0.176	N
C16	I am happy with simulator training on this aircraft.	11	6	0.294	N
C17	I aim to do better at my next flight simulator training session by learning from my mistakes.	11	6	0.294	N

Table 21: Continued

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
C18	I have a positive relationship with my colleagues.	14	3	0.647	Y
C19	The workload between trainees is balanced during a flight simulator training session.	9	8	0.058	N
C20	Pilots are judged as members of a team when they train in the flight simulator.	7	10	-0.176	N
C21	I feel rewarded for the amount of work I put into flight training.	10	7	0.176	N
C22	The more work I put into my preparation for training on this aircraft, the better I will perform.	11	6	0.294	N
C23	Pilots who are prepared have no problems training for this aircraft.	14	3	0.647	Y
C24	It is essential that pilots prepare adequately to pass a rating on this aircraft.	11	6	0.294	N
C25	I am in control of the outcome of my flight training on this aircraft.	16	1	0.882	Y
C26	I enjoy studying the technical aspects of the aircraft.	15	2	0.764	Y
C27	I always learn something new after undergoing training on this aircraft.	11	6	0.294	N
C28	I focus on the pertinent and relevant topics when learning about this aircraft.	12	5	0.411	N
C29	I reflect on my learning after a flight training experience.	14	3	0.647	Y
C30	I look for additional information so as to gain a deeper understanding of this aircraft's systems.	16	1	0.882	Y
C31	I know where to find specific information for this aircraft.	11	6	0.294	N
C32	It is important to know more than just what is required to pass.	16	1	0.882	Y
C33	I aim to gain a deeper understanding of this aircraft.	14	3	0.647	Y
C34	I learn more than is required of me from the company.	6	11	-0.294	N

Table 21: Continued

ITEM	ELEMENT	Endorsement of statement		CVR	RETAIN (Y/N) (Reject if CVR < 0.49)
		Essential	Not essential or not necessary		
C35	**I find the training on this aircraft easy.**	9	8	0.058	N
C36	**I do well in training for this aircraft.**	11	6	0.294	N
C37	**I look forward to my next flight training session.**	10	7	0.176	N
C38	**I sleep well the night before training on this aircraft.**	10	7	0.176	N
C39	**An appropriate level of stress helps me perform well in flight training for this aircraft.**	6	11	-0.294	N
C40	**I'm comfortable undergoing training for this aircraft.**	14	3	0.647	Y
C41	**I can control my anxiety so as to perform well in training.**	13	4	0.529	Y
C42	**I enjoy spending extra time flight training.**	11	6	0.294	N
C43	**I am motivated to learn more about this aircraft.**				
C44	**I am happy to be subjected to regular flight checks.**	12	5	0.411	N
C45	**I enjoy route training on this aircraft.**	9	8	0.058	N
C46	**I enjoy simulator training for this aircraft.**	3	14	-0.064	N
C47	**If my simulator partner is having a bad day, I am not affected.**	8	9	-0.058	N
C48	**I create a relaxed atmosphere in the flight simulator.**	11	6	0.0294	N
C49	**The length of time spent simulator training is appropriate for this aircraft.**	5	12	-0411	N
C50	**I enjoy the free play flight simulator time on this aircraft.**	11	6	0.0294	N
C51	**I aim to gain a deeper understanding of this aircraft.**	6	11	-0294	N
C52	**I learn more than the company requires me to.**	12	5	0.411	N
	AVERAGE NUMBER OF ENDORSEMENTS	11.038	5.961		Total (Y)=15
	AVERAGE PERCENTAGE	64.93	35.07		28.846

The hypothesised constructed consists of three separate dimensions at an organisational (airline), group (instructor-trainee team) and individual (trainee) level of analysis. Thus, the statements were clustered accordingly. Tables 19 to 21 show the level of expert endorsement of each item within each dimension of the hypothesised construct. More importantly, the last columns in each table report on whether the item was retained or discarded, based on its content validity ratio.

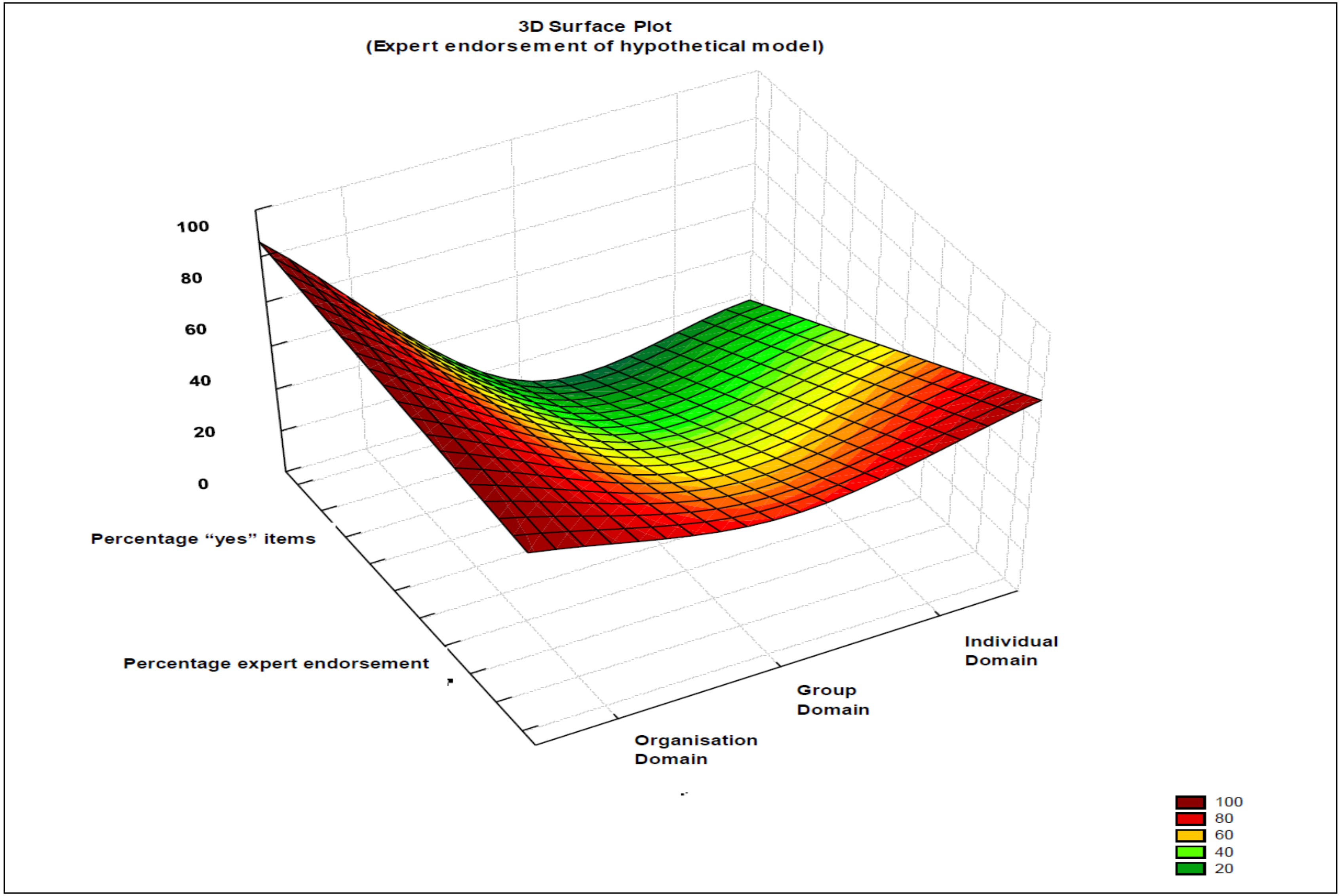
4.14 ITEM RETENTION RESULTING FROM THE APPLICATION OF LAWSHE'S TECHNIQUE

The computation achieved from the application of the Lawshe method resulted in the retention of 42 items. It appears that, out of the three dimensions of the hypothetical construct, Domain A, which assesses the organisational level of analysis (airline), was by far the most endorsed section, achieving an 81.047% proportion of expert endorsement and a 74.074% item retention level. Domain B, the group level of analysis of the hypothetical construct (the instructor-trainee dimension) received middling support from the panel of experts, with an endorsement proportion of 57.076% and an item retention level of 25.926%. Domain C, which assessed the trainee at the individual level of analysis, received a slightly higher level of support from the panel of experts, with an overall 64.93% acceptance of items, but only a 28.846% item retention level.

Additional clarity regarding the level of endorsement of items was mapped in a surface plot (see Figure 22). Myers, Montgomery and Cook (2009) refer to this kind of empirical evidence as a response surface model. The response given by the panel of 17 experts plotted on a three-dimensional surface suggests that items operationalizing the construct at a macro (organisational) level received far more support than the other two levels or dimensions. Red peaks suggest more support, whilst green troughs imply support to a lesser degree.

The 42 items extracted from the expert survey are considered very robust, due to the stringent criteria of the Lawshe method (Streiner, 2003). The next phase of scale development required an assessment of the authenticity of the data obtained from the item retention method followed.

Figure 22: Subject matter expert response surface model



4.15 ASSESSMENT OF INTER-RATER BIAS

Figure 22 usefully depicts the expert support for the 42 items that were retained. However, a level of inter-rater bias could not be eliminated and may have affected the analysis. Cochran's Q statistic was consequently calculated using the software package Statistica 7 to examine this possibility further. A matrix was produced in a one-way frequency table. A judge was given a score of 1 if he or she endorsed the proposed item, or conversely a score of 0 when the opposite was true (Table 22 provides a summary of this data). Therefore, a dichotomous variable was measured several times across differing conditions. According to Karlsson (2008), and Landis and Koch (1977), Cochran's Q test is an appropriate measure to determine whether the marginal probability of a positive response (that is, 1) is unchanged across the panel of judges. Cochran's Q test produced a very small P value (Q [16] = 201.3697, $p < 0.001$). Thus providing sufficient empirical evidence to conclude that the cohort of 42 essential or endorsed statements retained was of statistical importance.

Table 22: Summary of expert endorsement from Cochran's Q test

Expert	Sum	Percentage of 0s	Percentage of 1s
1	57	46.22	53.78
2	75	29.24	70.76
3	67	36.79	63.21
4	68	35.84	64.16
5	50	52.83	47.17
6	82	22.64	77.36
7	73	31.13	68.87
8	83	21.69	78.31
9	87	17.92	82.08
10	51	51.88	48.12
11	71	33.01	66.99
12	81	23.58	76.42
13	80	24.52	75.48
14	33	68.86	31.14
15	85	19.81	80.19
16	93	12.26	87.74
17	72	32.07	67.93
Mean	71.059	32.958	67.042

4.15.1 Final item retention

Based on the aforementioned analyses and further commentary from the group of subject matter experts with regards to the clarity and comprehensiveness of each retained item, Table 23 was produced after minor adjustments on selected statements. The final large survey item cohort is therefore found in the last column of Table 23.

Table 23: Comparison of items retained after applying Lawshe's method

Item	Retained statement based on Lawshe's method	Adjusted final large survey item
1	Pilot training at my airline is in line with company goals.	Training at my airline is in line with company goals.
2	My company's training produces world-class pilots.	My company's training produces world-class pilots.
3	I know what my company's training goals are.	I know what my company's training goals are.
4	My company has talented people managing airline pilots' training.	My company has talented people in training.
5	Pilot training at this company is professional.	Training on this aircraft is professional.
6	Management follows the regulator rules appropriately.	Management follows the rules and regulations appropriately.
7	Pilot training on this aircraft is well organised at this company.	Training on this aircraft is well organised.
8	I understand what the company expects of me when I am in training.	I understand what the company expects of me when training.
9	Training at my airline produces safe pilots.	Training at my airline produces safe pilots.
10	This airline gives its pilots an appropriate amount of preparation work before training.	The airline gives its pilots an appropriate amount of preparation work for training.
11	It is easy for me to appeal for assistance if I encounter a training problem at this airline.	If I had to experience a problem in training, it's easy for me to appeal.
12	There is sufficient training guidance from the company.	There is sufficient training guidance from the company.
13	The standard operating procedures (SOPs) for learning to fly this aircraft are adequate.	The standard operating procedures (SOPs) for learning to fly this aircraft is adequate.
14	The company provided me with sufficient time to prepare for training on this aircraft.	I'm given sufficient time to prepare for training on this aircraft.
15	The simulators my company uses to train its pilots are in good condition.	The simulators my company trains its pilots in are in good condition.
16	The airline is very supportive of its pilots' learning requirements for this aircraft.	The airline is very supportive of its pilots' learning requirements for this aircraft.
17	My company's culture supports training for new technology aircraft.	My company's culture supports training for new technology aircraft.
18	There is sufficient feedback about my training on this aircraft.	There is sufficient feedback about my training on this aircraft.
19	Pilot training at my airline follows civil aviation requirements.	Training is in line with civil aviation regulations.
20	My company uses only current training material.	My company uses only current training material.

Table 23: Continued

Item	Retained expert survey statement	Adjusted final large survey item
21	I learn better when I work as a member of the crew.	I learn better when I work as a member of the crew.
22	I operate well as a crewmember in the simulator.	I operate well as a crewmember in the simulator.
23	My instructor is willing to listen.	My instructor is willing to listen.
24	I communicate well with my simulator partner.	I tend to communicate well with my simulator partner.
25	The instructors on this aircraft are committed.	The instructor is committed.
26	Instructors are similar in how they teach pilots to fly this aircraft.	Instructors are very similar in how they teach pilots to fly this aircraft.
27	The instructors on this aircraft avoid overloading pilots with unnecessary information.	The instructors on this aircraft don’t overload us with information.
28	Pilots are in direct control of the training outcome.	Pilots are in direct control of the training outcome.
29	Adequate preparation improves flight training performance.	Preparation improves performance.
30	I am always on time for a flight training session.	I try never to be late for a training session.
31	I co-operate well when training in a simulator.	I co-operate when training in a simulator.
32	After flight training, I feel a sense of mastery.	After training I feel a sense of mastery.
33	I have a positive relationship with my colleagues.	I have a positive relationship with my colleagues.
34	Pilots who are prepared have no problems training for this aircraft.	Pilots who come prepared have no problems training for this aircraft.
35	I am in control of the outcome of my flight training on this aircraft.	I’m in control of the outcome of a training session.
36	I enjoy studying the technical aspects of the aircraft.	I enjoy studying the technical aspects of the aircraft.
37	I reflect on my learning after a flight training experience.	I reflect on my learning experience after a simulator session.
38	I look for additional information so as to gain a deeper understanding of this aircraft’s systems.	I read to understand so as to gain a deeper understanding of this aircraft’s systems.
39	It is important to know more than just what is required to pass.	It’s a good idea to know more than what is required.
40	I aim to gain a deeper understanding of this aircraft.	I aim to gain a deeper understanding of this aircraft.
41	I’m comfortable undergoing training for this aircraft.	I’m comfortable undergoing training for this aircraft.
42	I can control my anxiety so as to perform well in training.	I can control my anxiety so as to perform well in training.

4.15.2 Data collection

The self-administered survey (AATC-Q) was adopted for this part of the study. Three methods were used to distribute the large sample survey questionnaire to potential participants:

- Firstly, respondents were e-mailed a copy of the questionnaire, which they could answer, and then return to a specified e-mail address.
- Secondly, an electronic version of the questionnaire was hosted on the World Wide Web (see Appendix F). Additionally, the survey questionnaire web site was linked to the ALPA-SA home page and each potential respondent was requested to follow the link advertised.
- Finally, hardcopy questionnaire booklets were box-dropped in such a manner as to cover each pilot stratum. All responses to the hardcopy questionnaire were then subsequently recaptured on to the electronic version of the survey (World Wide Web).

4.16 DATA ANALYSIS

According to Cooper and Schindler (2003:87), "[d]ata analysis usually involves reducing accumulated data to a manageable size, developing summaries, looking for patterns, and applying statistical techniques".

In this section, the main approaches and techniques used to analyse the data that were collected are explained. One of the objectives of the study was to determine whether or not multiple variables contained in the measurement instrument could be reduced to a fundamental or latent factorial structure that may account for the majority of the variability found between respondents' replies.

In order to achieve the core research objective, the construct "perceptions of the advanced automated aircraft training climate" was operationalization and captured via an appropriate questionnaire as mentioned in Section 14.16.2.

4.16.1 Computerisation and coding of the data

Preparation of data requires concise editing, coding and statistical adjustment on the part of the researcher (Aaker *et al.*, 1995). The paper-based returns in this study required initial editing to identify omissions, ambiguities and errors. Answers that were deemed illegible or contained nonsensical responses were coded as “missing”. To ensure that this did not distort any interpretations of the data, the overall answers found in the returned questionnaires were reviewed. The paper-based returns were then recaptured electronically onto the web-based version of the survey to simplify data analysis. Coding the closed-ended questions from the web base was fairly straightforward, because the instrument made provision for response values and a column that was used for variable identification. The response values were then exported to a spread sheet and then entered into a computer software program. The Statistical Package for the Social Sciences (SPSS version 17) was employed to generate the statistical diagnostic information in most cases.

4.16.2 Statistical analyses

The purpose of conducting a statistical data analysis is to summarise univariate or multivariate data, to explore relationships between variables and to test the significance of these differences (Corston & Colman, 2003). The results obtained from the survey instrument were interpreted using appropriate statistical techniques for

- summary statistical descriptions;
- factor analysis;
- item analyses;
- reliability and homogeneity analysis;
- scale description;
- comparative analyses; and
- associational analyses.

The levels of measurement achieved at each stage also determined the choice of statistics used. The study treated the construct “perceptions of the automated aircraft training climate" as the dependent variable. In instances where the demographic variables of the sample frame were used to determine the effect of perceptions, these variables and situational categories then became the independent variables, and the hypothetical construct became the dependent variable, for example, the analysis of data would then indicate that more experienced airline pilots (the independent demographic variable) have a more favourable perception of the training climate (the dependent variable) than less experienced junior pilots have.

4.16.3 Analysis of compliance with specific assumptions

To assess compliance with the distribution requirements for factor analysis, Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy were used in this study. Morgan and Griego (1998:15) suggest that data are likely to factor well with a measure of sampling adequacy (MSA) of around "0.70". Table 24 shows the level of acceptability for the calculated measure of sampling adequacy according to Gravetter and Wallnau (2008).

Table 24: Acceptance levels for the measure of sampling adequacy

Acceptance	Measure of sampling adequacy (MSA)
Outstanding	0.90 to 1.0
Meritorious	0.80 to 0.89
Middling	0.70 to 0.79
Mediocre	0.60 to 0.69
Miserable	0.50 to 0.59
Unacceptable	Less than 0.50

As a general rule, for unequal sample sizes in social science research of this nature, Vermeulen (2009) strongly advocates computing Levene's test of homogeneity and Box's M-test for homoscedasticity. These diagnostic tests are administered to test for the assumption of equality of variance across groups. Such tests are a recommended requirement when conducting an analysis of variance or ANOVA when there is an assumption of the equality of covariance. In addition, favourable outcomes of these tests are sought for the parametric versions of a multivariate analysis of variance or MANOVA (Tabachnick & Fidell, 2007).

The Kolmogorov-Smirnov test (often called the K-S test) was used to analyse the normality of distributions and is generally regarded as the statistic of choice for such requirements in the behavioural sciences (Lilliefors, 1967). For instance, Field (2009) proposes that the K-S test be applied to determine whether a sample comes from a population with a specific distribution or can comply with a set of assumptions. The hypothesis regarding the distributional form (that is, the data following a specified pattern) is then rejected if the test statistic is greater than the critical value obtained from the SPSS-generated output table. Alternatively, Lilliefors (1967) and Pett *et al.* (2003), suggest conducting the *chi square goodness-of-fit* test to determine whether the observed frequency distribution of the respondents could reasonably have arisen from the expected sample frame distribution.

Both the K-S test and an analysis of the skewness and kurtosis of the data assisted the researcher in choosing between the two families of statistical methods, because choosing between a parametric and a non-parametric test can be difficult (Corston & Colman, 2003). Because one of the continued issues raised in survey research is the choice of statistics employed (Cohen & Lea, 2004), it was deemed important to critique the various methods available and to defend the final choices made, for achieving the goals in the present study.

Depending on the distribution pattern of the data received, appropriate parametric and non-parametric methods were considered at each analytical stage. According to Cohen and Lea (2004:222), a number of assumptions are generally made regarding the distribution of parametric variables:

- observations are independent;
- observations must be drawn from a normally distributed population;
- populations must have the same variances; and
- the means of these normal populations must be linear combinations of effects due to columns and/or rows.

Similarly, non-parametric assumptions may also have restricting requirements, such as, that

- observations must be independent; and/or
- the variable under study should have underlying continuity.

However, non-parametric testing tends to be far less restrictive than parametric procedures (Field, 2009). Stevens (1946) suggests that, instead of using actual measurements, the rank orders of measurements be used when conducting a non-parametric analysis. Depending on the situation, data was ranked from the highest to the lowest or vice versa (see Chapter 5).

Statistical tests fall into various categories of analyses, such as tests of differences and tests of relationships between groups or variables. Corston and Colman (2003) add that there is at least one non-parametric test that is the equivalent to any given parametric test (see Table 25). The categorisation and labels of some of these methods used in the final data analysis are summarised in Table 25.

Table 25: Comparison of statistical tests

	Parametric tests	Non-parametric tests
Differences between independent groups	• T-test for independent samples • ANOVA • MANOVA	• Chi square goodness of fit • The Kruskal-Wallis analysis of ranks • Mann-Whitney U • Non-parametric MANOVA
Differences between dependent groups	• T-test for dependent samples • Repeated measures ANOVA	• Wilcoxon's matched pairs test • Friedman's two-way analysis of variance
Relationships between variables	• Pearson's correlation coefficient • Probability regression analysis	• Spearman's Rho • Phi or Cramer's V • Kendall's Tau • Partial eta square • The chi square test

Source: Adapted from Cohen and Lea (2004), Field (2009) and Lilliefors (1967)

4.16.4 Descriptive statistics

Descriptive statistics were used to summarise the data. The essence of this statistical technique is to describe the sample and to calculate the mean, standard deviation, skewness and kurtosis of the sample scores (Corston & Colman, 2003; Gerbing & Anderson, 1988). An item analysis was then pursued to determine the initial item mean, item variance, standard deviation and item-scale correlation (Cooper & Schindler, 2003).

In order to analyse the distribution of each item as a percentage of respondents included in the different sub-dimensions, the descriptive statistical techniques mentioned were used where necessary. Any problems associated with the data that were collected (such as miscoded values or missing data) were discovered with the aid of summary statistics. Table 26 sets out the descriptive statistics that were applied in analysing the data obtained from the survey.

Table 26: Descriptive statistics

Summary statistic	Computation
Central tendency of variables	• Average or mean • Median • Mode
Measures of spread	• Variance • Standard deviation • Range • Inter-quartile range • Quartile deviation
Measures of shape	• Skewness • Kurtosis (platykurtic, leptokurtic, mesokurtic)

Source: Adapted from Cooper and Schindler (2003); Field (2009)

4.16.5 Factor analysis

Charles Spearman has been largely credited as the inventor of factor analysis (Cattell, 1987). Factor analysis is the preferred technique used to mathematically reduce a large amount of data into smaller more manageable clusters of related variables (Gerbing & Anderson, 1988). The method is commonly used in the behavioural sciences to uncover the latent dimensions when one is faced with a matrix of correlation coefficients (Cattell, 1987). Statistically clustering the common variables therefore informed the researcher of whether the instrument was a valid measure of the substantive constructs.

Two types of factor analysis were considered, namely exploratory factor analysis (which attempts to discover the nature of the underlying dimensions influencing a set of variables), and confirmatory factor analysis (which tests whether a set of variables is influenced by specific constructs in a predictive manner). Because the current study was an attempt to discover phenomena associated with a relatively unknown construct, an exploratory factor analysis was conducted on the dataset.

After receiving the questionnaires, respondents' answers were analysed by inserting the answers into a data matrix. The factor analysis used *heavy-duty* matrix algebra, because such a data matrix consists of as many rows as subjects (respondents), and as many columns as questionnaire items (Cohen & Lea, 2004; Pett *et al.*, 2003). In order to determine the interrelationships amongst these items, the data presented in the matrix took the form of Pearson product moment correlations or Pearson r (r_{xy}). According to Pett *et al.* (2003), the number corresponding to each row and column ranges from -1.00 to +1.00, where a negative value represents a negative correlation between the items, and a positive value represents the opposite. The subsequent meaning of relational strengths was then assessed in the context of the research topic.

Because one of the objectives in scale development is to determine the latent structure of a hypothetical construct, factor analysis was the analytical tool of choice to explain the variation and co-variation in a set of observed *variables* in terms of a set of unobserved *factors* (Field, 2009). Tabachnick and Fidell (2007) point out that reducing a complex array of data into statistically relevant correlates yields factors with some commonality. Sub-dimensions of the construct "perceptions of the advanced automated aircraft training climate" were then uncovered by using an exploratory factor analysis method. This technique is commonly employed when the exact number of factors that can accurately describe the construct of interest is unknown. In this case, there was no prior theory of the factorial structure of the construct that could be referred to. The basic aim of subjecting the data to an exploratory factor analysis was therefore to determine the relationship between observed, empirical evidence (survey results) and the latent factors (scale variables).

Field (2009) suggests that researchers use an appropriate statistical software package such as SPSS when conducting complex analyses. Many similar alternative software packages are also available in the market, such as Statistica. The program algorithm in many of the software packages calculates the interrelatedness between the factor, factors and/or other variables in the data space. This interrelatedness is then presented as a numerical value or correlation coefficient, referred to in exploratory factor analysis as a "loading" (Field, 2009). The loadings were used to find sub-constructs measuring the super-construct by rotating the loadings in order to find a pattern. The aim in this case was therefore to ascertain whether the variables of the measurement tool could be reduced (clustered) to yield an appropriate sub-structure.

4.16.6 Factor extraction

The retention of the correct number of factors has a significant bearing on the overall quality of a psychological scale (DeVellis, 2003; Gorsuch, 1997). Cohen and Lea (2004) contend that factors are generally extracted from a data set that represents the variance accounted for in each underlying factor. Two primary methods of extracting factors from the data space are available. Principal components analysis (PCA) or principal factor analysis (PFA) are the two methods of factor extraction often used in research similar to that in the current study. Schaap (2010) suggests that an exploratory factor analysis with principal factor analysis extraction is the preferred method. Additionally, according to Gorsuch (1997:534), "component analysis gives inflated loadings". In the current study, therefore, a principal factor analysis was conducted on the combined questionnaire item response variables.

In principal factor analysis, the diagonal of the so-called big-R matrix is replaced by estimates of communalities, which may be the reason it is considered a more prudent method than principal component analysis (Schaap, 2010). According to Field (2009), the communality of a variable is the proportion of the variance that is produced by the common factors underlying the set of variables. However, the actual difference in results (the number and nature of the factors) obtained when contrasting a component's factoring extraction method can be small for data obtained in the social sciences (Warner, 2008).

When one needs to uncover the shared variance in a set of variables, Horn's (1965) parallel analysis, eigenvalues or Cattell's scree plot are techniques available to determine the number of factors that are to be retained. Gorsuch (1997) recommends that, if the sample is large enough (at least 300 cases), it should be divided into two sub-samples, and then each sub-sample should be subjected to a factor analysis. This comparison improves understanding of the extracted factors. For the purposes of this study, both parallel analysis and scree plots (which involve studying the slope of the plotted eigenvalues) were considered as a retention method. It was found in this study that the calculation of eigenvalues tended to produce many unnecessary factors that became difficult to examine without a scree plot or parallel analysis, because each eigenvalue is the percentage of total variance accounted for by a corresponding component.

The eigenvalue of each factor accounts for the number of variance units out of the total number of items that are being measured and that yield the approximate percentage of variance accounted for by a specific factor (Brown, Hendrix, Hedges & Smith, 2012). Furthermore, for a particular class of square matrices (A), it is possible to find vectors (eigenvectors, x) such that when said square matrix is multiplied by its associated eigenvector, the resultant product provides a scalar or constant value (λ), referred to in this case as the eigenvalue; that is, A.x = x. λ (Brown *et al.*, 2012). Each element (item) of the square matrix would thus be associated with its own eigenvalue. In the current study, a major application of matrices was to represent such linear transformations, and therefore all the complex matrix algebra used in the study was conducted using commercially available computer software packages. In addition, the computed eigenvalues and eigenvectors provided an insight into the geometry of the transformations needed for factor analysis.

There are a number of criteria to determine the number of factors that should be retained (Gorsuch, 1983). Each of the following methods was considered based on both the advantages and disadvantages associated with the technique, and the requirements of the study:

- *Kaiser's criterion* – according to Kaiser's rule, components with eigenvalues under 1.0 should be rejected. However, this method tends to over-extract factors (Pett *et al.,* 2003).
- *"Variance explained" criteria* – this involves retaining enough factors to explain at least 90% or 80% of the variance in the data. This method is not recommended, given the level of subjectivity involved, and was thus discarded.
- *Cattell's scree test* – because a mathematical approach may lead to extracting factors of trivial importance (Gorsuch, 1983:167), Cattell (1987:16) suggests plotting the eigenvalues graphically. In this method, the components are plotted on the x-axis, with the corresponding eigenvalues plotted on the y-axis. The technique involves plotting the components as a diminishing series according to sizes and joining the points through the variables concerned. Where the number of factors ends due to certain error factors, a sharp break (or elbow) in the graph appears. This is why Cattell (1966:245) uses the analogy of "scree", which is a term that describes the broken rock fragments at the foot of a hill where it collects. Hayton, Allen and Scarpello (2004:192) point out that there is some subjectivity involved in determining the "sharp break" or when there are several "elbows" in the plot.
- *Comprehensibility* – the use of this method in *isolation* is not recommended as a scientific technique for answering the question of how many factors to retain. The non-mathematical nature of this process induces a fair amount of subjectivity when a researcher limits the number of factors to retain based on prior knowledge and comprehensibility.
- *Horn's Parallel Analysis (PA)* – a parallel analysis is one of the most robust and objective methods for retaining factors; however, very few computer programmes offer this solution, so the technique is rarely used. Because a parallel analysis

requires a Monte Carlo method to simulate mathematical systems (Glorfeld, 1995), Horn suggests comparing eigenvalues obtained from uncorrelated normal variables to the observed eigenvalues (such a comparison can only be achieved efficiently using a computational algorithm). Due to sampling error, sole reliance on Kaiser's criterion often overestimates the number of factors to retain; therefore Horn's method was considered in mitigation of this fundamental limitation (Hayton *et al*., 2004). In order to determine the number of factors to extract without over- or underestimating the quantity, a modified version of Horn's parallel analysis was conducted in the current study, based on a Monte Carlo simulation in SPSS using the syntax developed by O'Connor (2000).

According to Sawilowsky (2003), a Monte Carlo simulation determines the properties of a phenomenon from repeated sets of random or permutated samples. Thousands of random or permutated samples can be easily generated using specialised computer-based algorithms. The choice between selecting randomised or permutated data sets is based on the level of robustness sort by the researcher, where, permutated parallel data sets are considered robust (more complex mathematical formulae) and therefore less susceptible to error. Also, the choice of selecting a method can be affected by the availability of the appropriate computer software programmes. However, in the current study, permutated sets of the original data were generated for comparison with the real data, as the option was available, and provides a more accurate solution. The steps followed in performing the analysis were the following, as recommended by Hayton *et al*. (2004):

- Permutated data sets were generated quickly, based on the same dimensions as that those analysed. This was possible using the syntax provided by O'Connor (2000), which produced the Monte Carlo type simulation.
- Next, eigenvalues from the permutated data correlation matrix were extracted, based on the principal axis option.
- The mean and 95th percentile of all eigenvalues generated from the permutated data sets resulted in a vector of the same size as the number of variables, and diminishing in value.

- Finally, the real data were compared in parallel to the permutations; in other words, eigenvalues from the real data and generated sets were compared. Statistically significant factors (in this case $p < 0.05$) were retained, based on the fact that they are greater than the eigenvalues from the permutated sets. The factors retained were therefore statistically significant and were due to more than chance.

Plotting the actual eigenvalues versus randomly generated or permutated eigenvalues can give a clearer picture of the solution. In this case, the graphical plots generated were also compared to Cattell's scree plot (discussed later in Section 5.2.2), providing a substantive level of confidence for factor retention.

For the purposes of the current study, Kaiser's criterion, Cattell's scree plot, Horn's parallel analysis and comprehensibility of factors were used in combination to determine the number of factors that should be extracted and subsequently retained.

4.16.7 Factor rotation

Once factors have been extracted, it is plausible that some variables have high loadings on one important factor and small loadings on all other factors, causing some confusion with the interpretability of the variables and their subsequent latent structures (Cohen & Lea, 2004; Field, 2009). Two methods of rotation were considered for this study, namely orthogonal rotations (varimax, quatimax, equamax) and oblique (oblimin, promax, direct quartimin). The goal of all rotation strategies is to obtain a clear pattern of loadings. A comprehensive discussion of the various sub-rotation methods is beyond the scope of the current study, and therefore only the rotation method selected for the present research is discussed.

Pett *et al*. (2003) advocate the use of Kaiser's normalisation method for factor rotation. This technique, which is the most common, and generally the default setting in commercial statistical software packages such as SPSS, was deemed appropriate for analysing the present data set (Kaiser, 1961). In a study similar in nature to the present one, Rogers, Monteiro and Nora (2008) found that Kaiser's normalisation decreased the standard errors of the loadings for the variables that had lower communalities and economised the correlations among oblique factors. However, some authors have justified the use of orthogonal rotations when employing a Kaiser normalisation, so as to enhance a better understanding of the latent factorial structure present within a data set, in particular, a varimax rotation is punted (Govindarajulu, 2001; Green & Salkind, 2008). This then leads to orthogonal factors and variables that are regarded as completely independent from each other (Glorfeld, 1995; Gravetter & Wallnau, 2008; Ho, 2006). Employing a varimax rotation with Kaiser's normalisation is still the most common technique in use. The method can result in lower eigenvalues; however, interpretability of the final factors is slightly superior. Nonetheless, Field (2009:643) warns that the choice of the type of rotation "depends on whether there is a good theoretical reason to suppose that the factors should be related or independent", and by observing the nature of variable clustering before rotation.

The behavioural sciences are considered an interdisciplinary genre and theoretical constructs are consequently related at some fundamental or root level. "…i[I]f one expects that the factors would be related among themselves, then an oblique rotation is appropriate" (Rogers, *et al*., 2008:261). Furthermore, employing specifically a promax oblique rotation "…maintains the same high loadings as the first orthogonal solution in a varimax factor analysis" (Rogers, *et al*., 2008:261). Therefore, an oblique rotation (promax) method as opposed to the orthogonal (varimax) method was deemed most applicable for a study of the present theoretical nature. The original premise then holds that the factors in a latent structure of the data are related.

The current study used a promax rotation with Kaiser's normalisation raised to a Kappa 4 (see also section 5.2.2 for a more in depth discussion), which included several rounds of exploratory factor analysis, until a definite structure of latent factors explaining the majority of the variability within the dependent construct was obtained. The term "promax" reflects that the new axes after rotation are free to take any position in the factor space (Corston & Colman, 2003:55). The method seeks a rotation (linear combination) where the degree of correlations among the factors is allowed, in general, to be relatively small because a pair of highly correlated factors should be interpreted as a single factor (Thurstone, 1947).

Due to the number of rotation and extraction methods available, some scholars have questioned the true objectivity of factor analysis (Hayton *et al*., 2004). Nonetheless, it is apparent that linear factor analysis continues to dominate the behavioural and psychological sciences as a method to assess dimensionality among a set of correlated or uncorrelated variables. To assess the latent structure of the data by determining which variables to retain, the current study considered factor loadings of 0.40 and above, as well as the cross loading of items on more than one factor, and the reliability and importance of each variable.

The information for interpreting a factor analysis was obtained from the summary table output produced by the SPSS software programme (version 17), and is reported in Section 5.2. The summary table (see Table 35) relates factor loadings, communalities, eigenvalues, and the cumulative percentages of variance.

4.16.8 Reliability

In general, a test's reliability is examined by estimating the amount of error associated with its scores: "One of the central tenets of classical test theory is that scales should have a high degree of *internal consistency*, as evidenced by Cronbach's alpha (α), the mean inter-item correlation, and a strong first component...is used in establishing the reliability of the scale" (Streiner, 2003:217). Cronbach's coefficient alpha is the most commonly used statistic in determining the *reliability* of a scale (Field, 2009:674). Reliability in the current study implies that the measure that was developed consistently reflected the construct that it purported to measure. The alpha statistic is mathematically defined in the reliability formula (Field, 2009; Streiner, 2003; Zeller & Carmines, 1980):

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N}\sigma^2_{Y_i}}{\sigma^2_X}\right)$$

Where:

α = Cronbach's coefficient alpha;

N = The number of items

$\sigma^2 X$ = The variance of the observed total item scores

$\sigma^2 Yi$ = The variance of component i

i = Component i

The above equation clearly shows that α increases when the correlations between items increase. In theory, Cronbach's coefficient alpha should not be considered a statistical test *per se*, but should instead be related to a coefficient of reliability or indication of consistency (Zeller & Carmines, 1980).

In essence, by determining the reliability of an instrument, it is established just how well the items reflect the same construct and may consistently produce similarity in results (Saunders *et al*., 2007). The coefficient was used in the study to provide a mathematical value of how well the set of items measured the latent construct. The alpha statistic is a good measure of reliability because it relates directly to the average co-variance of all the items and is inversely proportional to the sum of all the item variances and co-variances (DeVellis, 2003). This therefore also provided the statistical level at which the items in the current research scale (or sub-scale) correlated with each other. "The high correlation tells us that there is similarity (or homogeneity) among the items" (Cooper & Schindler, 2003:239). This is an important concept that was a basic requirement in the development of the current psychological and behavioural measurement instrument.

An examination of the literature to determine the best value for Cronbach's coefficient alpha revealed a lack of complete clarity and also presented some confounding arguments (see Table 27). Most authors however, are comfortable in converging on a cut-off value of 0.70 as a reliable measure of perception (Clark & Watson, 1995:310; Cooper & Schindler, 2003:629; DeVellis, 2003:28; Field, 2009:675). In addition it can be argued that values in excess of 0.90 do not necessarily demonstrate good reliability, but rather point to the redundancy of items in a scale (Streiner, 2003).

The raw Cronbach's coefficient alpha values were compared to the values of the alpha value pertaining to the group of items in each sub-scale by deleting the item under analysis. An increase in the overall value of alpha indicates that the variable is neither reliable nor valid, and that it should be excluded from further examination (Cooper & Schindler, 2003; Streiner, 2003).

Table 27 was used as a guide in determining the level of item internal consistency for the developed measurement scale, as it provides a comparison of acceptable alpha values.

Table 27: A contrast of relevant Cronbach's coefficient alpha values

Source	Rationale
Cortina (1993:102)	α = 0.70: If the scale contains more than 14 items, or when the scale has at least two orthogonal dimensions with modest (0.30) inter-item correlations, 0.70 is a good alpha value for the test of reliability.
Field (2009:679)	α = 0.70 to 0.80: This shows that a questionnaire has good overall reliability; and 0.70 is needed for ability tests.
Netemeyer *et al*. (2003:102)	α = 0.70: 0.70 is widely advocated as an adequate alpha measure, and the statistic is directly related to the number of items and inter-item correlations.
Streiner (2003:102)	α = 0.70 to 0.90: An alpha higher than 0.90 may indicate redundancy rather than improved levels of scale reliability

4.16.9 Homogeneity

Within the framework of the theory of tests and measurements, homogeneity relates to the degree to which the items approximate a hierarchical scale (Krus, 2006). Alternatively, homogeneity can be used to test the inter-relatedness of each item with another and therefore it was used to determine the efficiency of the AATC-Q in measuring particular constructs (Oosterhof, 1976).

A comparison of the coefficients of reliability (r_{xx}) and homogeneity (h_{xx}) is provided in Table 28.

The formulae depicted in Table 28 show that both reliability and homogeneity are a function of the mean squares (MS), which represent the average squared deviations of an effect of interest around the grand mean. The mean inter-item correlations of the items in each factor were used to establish the unidimensionality or homogeneity of the scale (Clark & Watson, 1995). Inter-item correlations for the current study exceeded 0.20 and were thus regarded as acceptable in terms of the criteria advocated in the literature (De Vellis, 2003).

Table 28: Summary of reliability and homogeneity coefficients

Reliability	Homogeneity
Spearman-Brown's coefficient of reliability	Guttman's coefficient of reproducibility
Kuder-Ruchardson's formula K-R 20	Loevinger's (1957) coefficient of homogeneity
Cronbach's coefficient of reliability (alpha)	Cliff's coefficient of homogeneity
Hoyt-Jackson's coefficient of reliability $r_{xx} = \frac{MS_R - MS_I}{MS_I}$	Homogeneity as formulated by Krus and Blackman $h_{xx} = \frac{MS_R - MS_I}{MS_R^{max} - MS_I^{max}}$

Source: Adapted from Krus (2006)

4.16.10 Item discrimination analysis

After determining the appropriate items that contributed statistically to the latent factors of the measurement construct, it was necessary to examine the dispersion of the scores. A discriminant analysis provided evidence that the items in the developed scale were effective, in other words, the items differentiated adequately between high scores and low scores. In the study, item discrimination indexes were calculated by separating the item mean scores of the top half and lower half scores of responses from each item in the main interacting construct factors as suggested in Oosterhof (1976).

Furthermore, the upper bound scoring participants were separated from the lower bound scoring participants by appropriate dummy variables (1 or 0), and their response patterns were explicitly modelled based on the difference between the data classes. In discriminant analysis, according to Leech, Barrett and Morgan (2005:131), "one is trying to devise one or more predictive equations to maximally discriminate people in one group from those in another group". Statistically significant differences between discriminant item groups provided some evidence that the AATC-Q is a highly effective scale.

Due to the potential violations in the assumptions of multivariate normality in the collected data, poor accuracy of the estimates of the probability of correct classification was anticipated and expected. However, the method was still highly valid in achieving the study's scale development goals, because discriminant analysis is fairly robust to the assumptions of linearity, normality and equivalence of covariance across groups (Embretson & Reise, 2000). Nonetheless, both a matrix scatter plot and Box's M-test was used to determine the assumption of homogeneity of variance-covariance matrices in the data. Where the scatter plots were fairly equal, homogeneity of variance-covariance was assumed (Leech *et al*., 2005) and the power of the discriminant model was considered relatively stable.

4.16.11 Comparative statistics

To compare the relationship between the demographic dimensions of the respondents in the sample and the main and interacting constructs and sub-constructs, it was necessary to utilise various univariate and multivariate procedures (Field, 2005; 2009). The following measures of analysis of variance were selected as the basis for making inferences about the current data set.

Multivariate analysis of variance (MANOVA) was used to determine the main and interaction effects of categorical variables on multiple dependent interval variables. MANOVA makes use of one or more categorical independents as factor variables, with two or more dependent variables (Morgan & Griego, 1998). MANOVA tests the differences in the centroid (vector) of means of the multiple interval dependents, for various categories of the independent(s). After an overall F-test had shown significance, *post hoc* tests where then used to enable a more precise evaluation of differences between specific centroids. It has been suggested that the *post hoc* multiple comparison tests be performed separately for each dependent variable (Field, 2005:571-595). For example, the categorical subgroups (such as Boeing, Airbus, male, female) of the sample group were compared independently using the proposed *post hoc* procedures.

Because the data set did not meet the assumption of normality or homogeneity of variance-covariance (see Table 54 for more detail), the "non-parametric MANOVA with rank order data" was performed in the present study (Zwick, 1985:148-152). Box's M test was used to ascertain the homogeneity of the variance-covariance matrices (Anderson, 2001; Clark & Watson, 1995), because Zwick (1985) warns that the power of a significance test can be severely aggravated whenever sample sizes vary across cells. SPSS (version 17) was used to determine Box's statistic, and a violation of the assumption of homogeneity would then be accompanied with a low p-value, therefore this then further substantiated the employment of a non-parametric MANOVA.

To determine the significance of differences between categories (based on the multivariate Pillai-Bartlett trace, V), the degrees of freedom are initially computed by multiplying the number of dependent variables to the number of groupings minus one (Anderson, 2001; Zwick, 1985). Based on an appropriate alpha level, a chi-square table was consulted to locate the chi-square value that pertains to these degrees of freedom. Whenever the difference between the test statistic and critical chi-square value was exceeded, the test would be considered significant (Field, 2005; Zwick. 1985). Furthermore, the familiar Mann-Whitney non-parametric *post hoc* tests were applied to calculate the differences between the rank ordered means with only two categories or sub-groups (Morgan *et al.*, 2007).

Due to the skewness of the distribution of the responses in this study, it was also decided to use the Mann-Whitney U test and Kruskall-Wallis test to compare the mean rank order scores of different groups (see Section 5.6 for detail). These tests are commonly referred to as distribution-free tests (Rencher, 2002; Rosenthal, 1994). As non-parametric methods, their applicability is much wider than the corresponding parametric methods. Because non-parametric methods rely on fewer assumptions, they are also more robust (Field, 2005; Stuart, Ord & Arnold, 1999). The final analyses employed non-parametric or distribution-free statistical tests, because these tests do not depend on any assumptions about the form of the sample population or the values of the population parameters, making the method less limiting.

The Mann-Whitney U (M-W) test is a non-parametric test used to assess whether two samples of observations come from the same distribution (Tabachnick & Fidell, 2007). This test was used because the assumptions of the t-test were violated, in that the dependent variable data set was non-normally distributed or ordinal in nature. The Mann-Whitney U test is only slightly less powerful then Student's t-test (Field, 2005; Morgan *et al.*, 2007). For the M-W test, Z-values are calculated that are used to "approximate" the statistical level of significance for the test (Winks, 2008:108). The method was found to be ideal for unequal and small sample sizes (Babbie, 2010).

The Kruskall-Wallis (K-W) test is a non-parametric test that can be applied to assess whether three or more independent samples of observations have the same distribution (Rencher, 2002). The Kruskall-Wallis test was used as an alternative to its parametric one-factor ANOVA counterpart because the ANOVA's normality assumptions were not completely met in the present data set; also the data were ordinal. The K-W test uses mean ranks to determine whether scores differ across groups and a chi-square distribution to estimate the statistical level of significance for the test (Field, 2005; Morgan *et al.*, 2007).

4.16.12 Associational statistics

When one explores the relationship between variables, one should quantify the degree of linear relationship between two variables at an ordinal or interval level of measurement (Embretson & Reise, 2000). For an ordinal level of measurement, Spearman's Rho can be used, and for an interval level of measurement, Pearson's Correlation Coefficient can be used, thereby generating a value of association between parametric variables (Field, 2009). Alternatively, Netemeyer *et al.* (2003) suggests the chi square test and Kendall's Tau-b statistic as a more robust determination of association when dealing with non-parametric data. These techniques (see Table 29) were kept in cognisance when attempting to understand the phenomena within the data set.

Correlational research is used to discover how the status on one variable tends to reflect the status on another (Babbie, 2010). Mainly non-parametric correlations were used in this study to predict the effect of one variable on another; and examine related events, conditions or the behaviour of the population sample. According to Morgan and Griego (1998), the predictor variable (independent) is believed to produce an outcome in the dependent or criterion variable. In order to establish the most appropriate correlation statistic to use in determining association, Table 29 was consulted for contrasting the available options.

Table 29: Correlation statistic guideline

Predictor variable	Criterion variable	Correlation to use
Interval (continuous)	Interval (continuous)	Pearson
Real dichotomy	Interval (continuous)	Point biserial
Artificial dichotomy	Interval (continuous)	Biserial
Real dichotomy	Real dichotomy	Phi
Artificial dichotomy	Artificial dichotomy	Tetra choric
Ranking	Ranking	Spearman's rho
Ranking	Ranking	Kendall's tau

Source: Adapted from Field (2009), Leech *et al.* (2005) and Stuart *et al.* (1999)

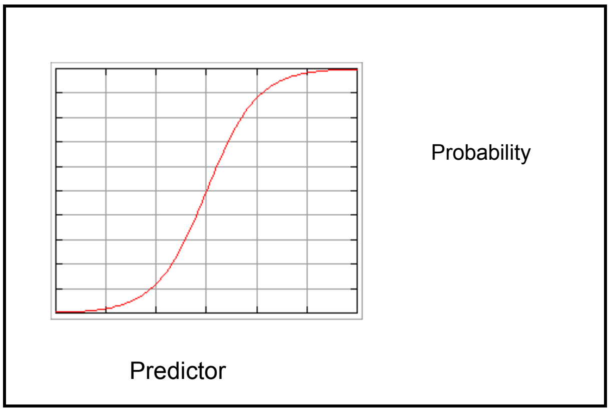
4.16.13 Logistic regression analysis

Logistic regression was the method of choice in developing a predictive model from the data set, because "[r]egression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables" (Hosmer & Lemeshow, 2000:1). The probability of a binary outcome on a discrete variable was modelled from the most likely relationship between demographic covariates as a last step in the final statistical analysis of the observations.

"Logistic regression allows one to predict a discrete outcome such as group membership from a set of variables that may be continuous, discrete, dichotomous or a mix" (Tabachnick & Fidell, 2007:437). A stepwise method was adopted, where all predictors were initially placed in the logistic model and eliminated sequentially in subsequent steps of the analysis. Stepwise regression is used in the exploratory phase of research (Field, 2005). The options available in a logistic regression and justification for the final method selected is discussed in Section 5.10.

In the current study, a dichotomous dependent variable was constructed, based on the level of favourability perceived by the respondents. Similar to discriminant analysis, a dummy variable (1 or 0) was allocated to the dichotomy of perceiving a favourable or unfavourable training climate. The regression based on the logit is particularly useful in this case, as the distribution of responses on the dependent variable was non-linear, with one or more of the independent cases or co-variates. In other words, the probability that the criterion variable will have one outcome rather than another is based on a non-linear function of the best linear combination of the predictors (Tabachnick & Fidell, 2007). The logistic curve can be considered a generalised sigmoid or “S” curve (Govindarajulu, 2001; Morgan *et al*., 2007). An “S” curve begins exponentially and thereafter begins to taper off (see Figure 23). The curve is based on a mathematical concept that has been widely used to model the natural life cycle of many phenomena (Cohen & Lea, 2004). For instance, in the case of the current study, the plotted logistic regression model clearly showed an initial exponential change in probability between the levels of perceived computer competence in the interaction effect between flying experience and computer ability, and a slowing down or tapering off later in the curve.

Figure 23: General shape of the common sigmoid curve used in logistic regression



Source: Adapted from Govindarajulu (2001)

The non-parametric nature of the empirical data that was collected in the current study meant that an alternative multiple regression analysis technique was required. According to Field (2009), the logistic function provides a value of probability between 0 and 1 based on the logit formula or curve (non-parametric), where logit (p) = ln (p/1-p), also referred to as the log odds (Tabachnick & Fidell, 2007). This is in contrast to the probit curve (parametric), which is based on a probability unit or normal distribution (Govindarajulu, 2001). The primary reason for using a logistic regression analysis to model the data in this study was that the method offers several distinct advantages (Tabachnick & Fidell, 2007):

- logistic regression is robust, in that the independent variables in the equation do not have to be normally distributed or display equal variances within groups;
- there is no assumption of linearity between the predictor and criterion variables;
- the outcome variable can be binary; and
- there is no requirement for continuous or interval independent variables.

However, Field (2005) points out that there are also some distinct disadvantages associated with the logistic regression technique. For instance, the method requires a higher number of data points to produce meaningful and stable results.

To determine how powerful the developed regression equation was at predicting the variable of interest (proportion of variance in the criterion variable associated with the predictor variable), a pseudo R^2 was computed, based on the methods of Cox and Snell's R-Square, Nagelkerke's R^2, and McFadden's (adjusted) R^2. A pseudo R^2 is computed to evaluate the goodness-of-fit of the logistic model (Tabachnick & Fidell, 2007). In general terms, a correlation coefficient can range between -1 and 1. However, in the case of the R^2 coefficient, the value computed in the current study ranged from 0 to 1 (because squaring the correlation between the predicted values and the actual values of the regression model produced a positive value). This value is referred to as the so-called pseudo R^2. According to Govindarajulu (2001), a high value of the pseudo R^2 indicates that there is a greater magnitude of the correlation between the predicted values and the actual values. Ho (2006) cautions, however, that when using different pseudo R^2s, it is possible that one may arrive at very conflicting results.

Cox and Snell's R^2 was calculated based on the following formula (Long, 1997):

$$R^2 = 1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}$$

Where L(M) is the conditional probability of the dependent variable, given the independent variables (the model *intercept* contains no predictors or independent variables). When there are N observations, L(M) is the product of N probabilities. The formula shows clearly that even if the regression model were a perfect fit, the R^2 value can never attain a value of 1 (Field, 2005). Cox and Snell's R^2 indicated the improvement from the null model (intercept only) to the derived or fitted model.

Nagelkerke’s R^2 was calculated based on the following formula (Long, 1997):

$$R^2 = \frac{1 - \left\{ \frac{L(M_{Intercept})}{L(M_{Full})} \right\}^{2/N}}{1 - L(M_{Intercept})^{2/N}}$$

Like Cox and Snell’s R^2, Nagelkerke’s R^2 provides a value which indicates the improvement of the full model from the null or intercept only model. However, the formula indicates clearly that the range of the R^2 value can achieve 1 in a prediction model with a perfect fit.

McFadden's (adjusted) R^2 was calculated based on the following formula (Long, 1997):

$$R_{adj}{}^2 = 1 - \frac{\ln \hat{L}(M_{Full}) - K}{\ln \hat{L}(M_{Intercept})}$$

In the above formula, L-hat refers to the estimated likelihood. The adjusted formula is useful in that it indicates whether particular predictors add value to the model or not. The model is penalised by a reduced R^2 value when there may be too many ineffective predictors (K). The formula clearly shows that a negative R^2 is possible when using McFadden's adjusted method.

Several significance tests are available to determine the inclusion or exclusion of co-variates from a logistic regression model (Hosmer & Lemeshow, 2000). The Wald test was used to test the statistical significance of each of the coefficients in the regression model. A Z-score (Z = coefficient [B]/SE) was calculated. The hypothesis of inclusion or exclusion of the coefficients was thus based on the subsequent chi-square fit (Tabachnick & Fidell, 2007). For smaller sample sizes, Agresti (1996) suggests the likelihood-ratio test. Because a backward stepwise elimination was followed in building the final model, the likelihood-ratio test statistic used in the current study was based on the following formula:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1)$$

The above equation shows that the maximised value of the likelihood function for the full model (L_1) was compared to the maximized value of the likelihood function for the simpler or null model (L_0), associated with a chi-square goodness-of-fit.

The Hosmer-Lemeshow goodness of fit test was applied to determine whether the model prediction did not differ significantly from the observed number of subjects in each group. It was desirable to achieve a non-significant chi-square test statistic in a case such as this, as recommended by Agresti (1996). A good model would be effective in categorising most successful subjects into an upper level and placing failures into a lower level (Hosmer & Lemeshow, 2000).

The change in odds (odds ratio) based on a unit change in the predictor indicated the impact of each predictor on the regression model. The odds ratio and the other tests mentioned which were conducted on the final logistic regression model are discussed further in Section 5.10.

4.16.14 Practical significance and effect size

When a significant result is reported from the research, it is accompanied by a p-value and confidence can be established in the assumption that the results are not simply due to chance (Muijs, 2004). However, a p-value has more meaning when it is accompanied by an effect size value. Cohen and Lea (2004) are critical of studies that assess significance without an accompanying practical or effect size report. For instance, if a study looks for a 95% confidence interval, a p-value of less than 0.05 implies that there is less than a 5% probability that the result may occur by chance. However, this would not indicate how large the significance actually is.

The goal of practical significance computations of research results is an interpretation to present significant conclusions, which may be more meaningful to a non-statistician (Cohen, 1992). In other words, statistical significance shows that the results are unlikely to have occurred by chance, whereas practically significant and effect size results are more "meaningful in the real world" (Ellis, 2010:15). It is important to note that in this context effect size does not refer to cause and effect relationships between variables, but merely provides a value that quantifies the practical significance of findings (Rosenthal, Rosnow & Rubin, 2000).

In many cases, it is necessary to know whether a relationship between two variables is practically significant – for example, between pilots' level of education and their perceptions of advanced automation training climate. The statistical significance of such relationships can be determined by using the correlation coefficients (r). In this case, the effect size was determined by using the absolute value of r and relating it to the cut-off points for practical significance recommended by Cohen (1988), where

- r = 0.10 suggests a small effect;
- r = 0.30 suggests a medium effect; and
- r = 0.50 suggests a large effect.

To assess the significance of the z-statistic of the Mann-Whitney test the coefficient 'r' was computed by using the conversion formula, $r = z/\sqrt{(N)}$ suggested by Field (2005) and Morgan *et al*. (2007).

Pett *et al*. (2003) recommend the calculation and use of the partial eta square (η^2) to determine the effect sizes or strength of relationship between demographic variables and the construct of interest. The results of the η^2 provide a value that quantifies the practical significance of the findings (Cohen, 1988). Where MANOVAs and ANOVAs are implemented and statistically significant main and interaction effects are found, the partial eta squared is calculated to determine the practical effect size. Partial eta squared (η^2) is the proportion of the effect, plus the error variance attributable to the effect. In the current study, Field's (2009) formula, $\eta^2 = (SS_{effect})/(SS_{effect} + SS_{error})$, was used to determine partial eta squared.

According to Cohen's (1988) effect size criteria, the following cut-off points normally apply if partial eta squared is to be of practical significance:

- $\eta^2 = 0.01$ suggests a small effect;
- $\eta^2 = 0.06$ suggests a medium effect; and
- $\eta^2 = 0.14$ suggests a large effect.

4.17 RESEARCH ETHICS

An area of concern for many researchers conducting empirical studies using primary data in the social sciences is gaining access to participants (Saunders *et al*., 2007). To overcome this problem, permission was obtained from ALPA-SA, which gave its consent for the researcher to access its database of over 1 000 airline pilots. Airline management at the various organisations also endorsed the project. In addition, all participants were asked to acknowledge a consent form prior to commencing with the survey (see Appendix D). In accessing these participants, a quality scientific inquiry was conducted, adhering strictly to the moral and ethical research principles applicable at the University of Pretoria.

The *quality* of a research design is directly related to the *ethical* standing of the final research (Rosenthal, 1994:127). Researchers in the social and psychological sciences face many dilemmas that may have an impact on the morality of their studies. Some important issues in ethics that can become a problem if these issues are not avoided are the following (Rosenthal, 1994):

- *hyper-claiming* – claiming that the research will achieve specific goals and objectives that it cannot achieve;
- *causation* – claiming that there is a causal link between variables when there is actually none;
- *data dropping* – analysing data that never existed or removing data that conflicts with the researcher's hypotheses; and
- *questionable generalisations* – failing to pay careful attention to not inferring findings on a population without sufficient empirical evidence or an adequate sampling technique.

An outline of the principles involved at each stage of the research process to ensure that the study was conducted in an ethical manner is provided in Table 30. The columns show the stage of the research at which a particular principle was applied and thereafter what techniques were used to mitigate any adverse morality issues.

Table 30: Ethical issues considered in the research process

Stage of research	Possible ethical issue	Specific issues addressed
Exploration	Confidentiality	Sponsor non-disclosure.
Research proposal	Informed consent	Participants' and sponsors' right to quality research.
Research design	Informed consent and confidentiality	Deception of respondents. The right to privacy. Immoral coercion from parties with ulterior agendas.
Data collection/preparation	Confidentiality	Participant privacy issues. Data exploitation.
Data analysis/reporting	Confidentiality and Data completeness/integrity	Confidentiality of participants. Censoring of results. Meta-analysis.

Source: Adapted from APA (1994), Cooper and Schindler (2003), and Rosenthal (1994)

The issues highlighted in Table 30 were dealt with on a case-by-case basis during the writing and distribution of the final completed questionnaire.

According to Cooper and Schindler (2003:120), "[e]thics are norms or standards of behaviour that guide moral choices about our behaviour and relationships with others" – hence, any research conducted in the field of psychology affecting human subjects requires consent from the organisation responsible. The American Psychological Association's (APA) guidelines were strictly adhered to in order to maintain ethical standards required by the University of Pretoria. These principles, according to the APA (2002:3-5) are

- *beneficence and non-maleficence* – psychologists should maximise the benefits of participants and minimise any harm that may result from their research;
- *fidelity and responsibility* – scientists must establish lines of trust between themselves and participants, and must ensure the highest standards of professionalism by maintaining objectivity;
- *integrity* – scientists must promote honesty, accuracy and the truthfulness of their research;
- *justice* – all persons participating should be entitled to access the research that is being conducted; any unjust practices must be prevented by guarding against potential biases; and
- *respect for people's rights and dignity* – researchers must adhere to the requirements of participants' right to privacy, confidentiality and self-determination at all times.

The final instrument was accompanied by a cover letter introducing the study. The letter assured respondents who volunteered to take part in the study of the confidentiality of their responses and their anonymity. The current study endeavoured to maintain academic objectivity at all times by following a structured design and methodology, and complying with the ethical requirements for research of this kind. Annexure D contains a draft of the informed consent form that was used.

4.18 SUMMARY

The chapter focused on the methodologies and statistical applications required to design and construct a valid and reliable psychometric instrument for the advanced aircraft industry. The present research was completed by means of abductive, inductive and deductive reasoning processes grounded in prior theory. The study design consisted of an empirical quantitative approach based on a positivist paradigm, which resulted in the development of a new measurement tool in the advanced aircraft training environment.

The chapter also discussed the population, method of sampling, the design and layout of the questionnaire, the type of questionnaire used, the design of the questionnaire items, as well as the correlations, factor analysis, comparative, associational and regression modelling technique used in the study. Practical and effect sizes were discussed with regard to reporting any tests of significance.

The statistics used in the research were discussed rather in detail, because they form the foundation for the reporting of results and recommendations set out in the subsequent chapters.

The following general quantitative steps in the second phase of the research were undertaken to meet the research objectives:

- Step 1: Determine the content validity of each item's relationship with the construct and sub-constructs. This was calculated using,
  - Lawshe's (1975) method, and
  - Cochran's Q statistic.

- Step 2: Explore the data using a factor analytic technique;

- Step 3: Applying an appropriate rotation method to extract an optimum number of factors;

- Step 4: Analyse clusters of items;

- Step 5: Determine the level of reliability and homogeneity present;

- Step 6: Summarise the data;

- Step 7: Determine the distribution and normality status of the data set; and

- Step 8: Explore possible relationships and phenomena of the latent structure by means of the following non-parametric statistical procedures:
  - the non-parametric MANOVA;
  - the Kruskall-Wallis test;
  - the Mann-Whitney U-test;
  - Spearman’s rank order correlation and Kendall’s tau; and
  - Logistic regression.