CHAPTER 5

RESEARCH DESIGN AND METHODS

The aim of this research was to explore the feasibility of adapting an existing monitoring system to the South African context. A non-experimental pragmatic approach was used in the research, which utilised a mixed method design, namely a concurrent nested design. Participants in the research included Department of Education officials, principals, educators, and learners. Data was collected by means of a variety of instruments and different validation strategies were used. This study also made use of various data analysis strategies in order to address the research questions identified. The data analysis strategies included thematic content analysis, item response theory, reliability analysis, correlation analysis, and multilevel analysis. The multilevel analysis ascertained the variance in performance explained by factors on a school, classroom, and learner-level. Finally, ethical considerations are discussed and methodological constraints elaborated on.

5.1 Introduction

The foundation of the research process rests on an overarching methodological framework consisting of questions, designs, data structures and decisions about analysis (Heck & Thomas, 2000). The framework draws the various elements of research together into a cohesive, comprehensive whole culminating in a chain of reasoning (Krathwohl, 1998). Furthermore, this framework is rooted in a particular worldview, or a particular way in which truth is perceived and understood (Worthen, Sanders & Fitzpatrick, 1987). This framework or rather the assumption about how things are understood is referred to as a paradigm. This is the lens used to make sense of things (Creswell, 2003; Worthen et al., 1987). Within a paradigmatic framework knowledge claims are made about what knowledge is, how one knows, what values are attached, the way in which we study phenomena and how the phenomena are written about (Creswell, 2003). A pragmatic approach was adopted in this research. The pragmatic approach is discussed below (5.1.1), including a brief overview of different paradigms and a description of the development of pragmatism. The section concludes with a justification of pragmatism as an overarching framework (5.1.2).

5.1.1 The development of pragmatism

Petter and Gallivan (2004) state that until World War II, positivism was the prominent paradigm in the social sciences, which was rooted in the belief that knowledge was based on observable facts. However, after World War II positivism was severely criticised for not producing significant outcomes in social sciences when compared with the physical sciences over and above the argument that human behaviour is complex and needs to be explored more thoroughly. In reaction to this post-positivism was born in order to address the problems experienced in positivism (Petter & Gallivan, 2004).

Post-positivism as an extension of positivism, is characterised by the use of quantitative methods (QUAN), and is governed by the underlying philosophy that causes for effects or outcomes could be obtained. The aim is to examine causes in order to determine the influence on outcomes. Knowledge is developed based on careful observation and the measurement of objective reality by means of developing numeric measures. First, there is theory, then the collection of data or evidence, then the conclusion that the theory is right or wrong (Creswell, 2003).

Discontentment with positivism and post-positivism deepened a need for an alternative approach to research in the social sciences. In the 1970’s qualitative methods became more prominent. Researchers like Lincoln and Guba, Stake and Eisner wrote several books that were critical of the positivist orientation. In response to their criticisms, an alternative in the form of a variety of qualitative methods (QUAL) was proposed. Constructivism was the common name given to the paradigm using qualitative methods (Teddlie & Tashakkori, 2003). During this time, several debates raged in the social sciences regarding the superiority of one or the other of these two paradigms. Numerous attempts were made to make peace between the two paradigms by “pacifists”, such as Tashakkori and Teddlie (1998), Maxcy (2003) as well as Johnson and Onwuegbuzie (2004), who stated that qualitative and quantitative methods are compatible (compatibility thesis). However, paradigm purists, such as Smith and Heshusius (see Teddlie & Tashakkori, 2003) as well as Lincoln and Guba, Popper, Maxwell and Delaney (see Johnson & Onwuegbuzie, 2004), reject the compatibility thesis stating that compatibility between quantitative methods and qualitative methods is impossible due to the knowledge claims made by the different methods (Tashakkori & Teddlie, 1998). Nevertheless, authors promoting the compatibility thesis still support the thesis basing their claims on a different paradigm namely pragmatism (Patton, 1990; Tashakkori & Teddlie, 1998). Pragmatism gained momentum in the 1990’s with researchers becoming frustrated with having to choose between qualitative and quantitative methods (Petter & Gallivan, 2004). Furthermore, knowledge claims for pragmatists arise out of actions and consequences rather than antecedent conditions (as in post-positivism). Pragmatists are concerned with solving problems and applying the most appropriate methods in order to do so. The problem is foremost, followed by an elaboration on the best methods suited to address the problem (Creswell, 2003).

Pragmatism is a philosophical movement that began towards the end of the 19th century (Maxcy, 2003). It originated in the United States unlike most philosophical movements, which originated in Europe (Expers, 2000). It has historical roots with noteworthy contributors such as Charles Sanders Pierce, William James, John Dewey, and George Herbert Mead (Creswell, 2003). A common element of these contributors was their rejection of traditional assumptions about the nature of knowledge, truth and the nature of inquiry. They also rejected the notion that the real world could be accessed solely by means of one scientific method (Maxcy, 2003). Maxcy (2003) states that for these early pragmatists meaningful research was not rooted in the methods employed but rather in ordinary experience and the desire to understand, a desire for a better world. Maxcy goes on to say that two eras exist in the development of pragmatism, namely early pragmatism in which Pierce, James, Dewey, Mead, and Bentley are highlighted and neo-pragmatism in which there were contributors such as Abraham Kaplan, Richard Rorty, Richard Berstein and Hilary Putnam. The neo-pragmatists draw on the work of the early pragmatist as elaborated on in the discussion above. Neo-pragmatists reframe the tenets of early pragmatism stressing the importance of “richer modes of inquiry” (Maxcy, 2003, p. 54) as in “methodological pragmatism” (Maxcy, 2003, p. 81), according to which a researcher should not focus on a single method but instead test different methods of inquiry for their effectiveness in achieving the intended goal. Neo-pragmatists also place emphasis on new ways of communicating by making use of “metaphors”, “stories” and other “narratives” as in the case of Richard Rorty (Maxcy, 2003, p. 80), and attempts to secure forms of common experience (Maxcy, 2003).

Maxcy (2003) is of the opinion that in pragmatism both the meaning and the truth of any idea are functions of its practical outcome. It is the problem that is of importance and not a preoccupation with methods (Creswell, 2003). Outcomes are what counts and not necessarily prior knowledge claims, laws or even what is true (Maxcy, 2003). All principles are viewed as working hypotheses rather than metaphysically binding truths (The Radical Academy, 2002). Subjective and objective perspectives in addition various methods should be used in order to achieve the desired outcome. According to pragmatists, the integration of methods from the different paradigms is a powerful way of enhancing the credibility of findings (Petter & Gallivan, 2004). Pragmatists are of the opinion that there are similarities in the fundamental values between QUAN and QUAL approaches. These beliefs include the value-ladeness of inquiry, theory-ladeness of facts, that reality is multiple and constructed as well as that knowledge is fallible (Tashakkori & Teddlie, 1998). Table 5.1 represents a comparison between the three paradigms namely, post-positivism, constructivism, and pragmatism.

Table 5.1 ***Comparison between post-positivism, constructivism, and pragmatism***

	Post-positivism	**Constructivism**	**Pragmatism**
Methods	Quantitative	Qualitative	Quantitative and Qualitative
Logic	Deductive	Inductive	Deductive and Inductive
Epistemology	Modified dualism, findings probably objectively true	Subjective point of view, knower and the know are inseparable	Both objective and subjective point of view
Axiology	Inquiry involves values but they may be controlled	Inquiry is value bound	Values play a large role in interpreting results
Ontology	Critical or transcendental realism	Relativism	Accept external reality, choose explanations that produce the desired results
Causal links	There are some lawful, reasonably stable relationships among social phenomena, these may be known imperfectly and causes are identifiable in a probabilistic sense that changes over time	All entities simultaneously shaping each other. It is impossible to distinguish causes from effects	There may be causal relationships but will never be able to pin them down

(adapted from Tashakkori and Teddlie, 1998)

The pragmatist rejects an “either or” situation and makes use of both inductive and deductive logic, meaning that one argues from the particular to the general and the general to the particular depending on the problem at hand. Furthermore, pragmatists use both objective and subjective points of view, viewing these on a continuum where one, in the research process, would be more subjective at certain times and more objective at other times. Values in research within pragmatism as a framework play a large role in interpreting results. The pragmatist decides what to research, and makes knowledge claims in terms of what knowledge is, how this can be known as knowledge, the role of values and the methods of study. However, the problem is foremost, followed by discussions on methods which best suit the investigation of the problem (Creswell, 2003). The process is guided by the researchers’ personal value system and they study what they think is important to study. The research methods as well as the research results reflect the researcher’s value system. The pragmatic researcher accepts external reality but chooses explanations that produce the desired results. Pragmatists use both qualitative and quantitative methods in order to obtain the best result. Qualitative refers to research designed to address questions of meaning, interpretation, and realities, which are socially constructed. Quantitative in this context refers to research designed to address questions that formulate hypotheses about relationships among variables, is essentially descriptive in nature, makes use of numeric variables, and attempts to measure the relationships in objective ways (Newton, Ridenour, Newman & DeMarco, 2003). The choice of which method to use, however, depends largely on the research question and with each method one would apply either subjective or objective points of view.

For the pragmatist the research question is more important than the method that is used or the worldview that underlies the method. Furthermore, researchers should address these questions with whatever methodological tool is available. "What works" is what counts. Pragmatists prefer to deal with the practical problem at hand viewing modes of inquiry as tools for better understanding and effective problem-solving (Reeves, 1996). Tashakkori and Teddlie (1998, p. 21) concur and state that pragmatists are "committed to the thorough study of the research problem, method is secondary to the question itself, and the underlying worldview hardly enters the picture, except in the most abstract sense". Moreover, Greene and Caracelli (2003) are of the opinion that applied social inquirers ground decisions primarily on the nature of the phenomenon being investigated, as well as the context in which the investigations are taking place with philosophical assumptions rarely being considered.

5.1.2 Rationale for working within a pragmatist paradigm

Tashakkori and Teddlie (1998) suggest that pragmatism is appealing as:

1) It abstains from metaphysical concepts that have caused endless debates.
2) It presents a practical and applied research philosophy, which states that one should study what is of interest and value to you, study it in different ways, as one deems appropriate and use the results in a way that brings about consequences within the value system one is working in.
3) It provides a paradigm, which philosophically embraces the use of mixed methods.

The primary appeal of pragmatism for this research is not that it abstains from endlessly defending metaphysical concepts or even that it embraces mixed methods, although these are intriguing. Rather, the primary appeal lies with the fact that pragmatism represents a practical and applied research philosophy. It provides the opportunity to investigate what is of importance to the research and broader society in manner, which is compatible with the questions that one wishes to address. Pragmatism makes the investigation of the perceived problem possible without imposing constraints on methods to be used but allows the researcher the option of making use of all possible tools in order to address the problem in a comprehensive manner. Patton (in Teddlie & Tashakkori, 2003, p. 18) states that:

> …in real world practice, methods can be separated from the epistemology out of which they have emerged. One can use statistics in a straightforward way without doing a literature review of logical positivism. One can make an interpretation without studying hermeneutics. In addition, one can conduct open-ended interviews without reading treatises on phenomenology.

Furthermore, pragmatism was chosen as the overarching paradigm for this research as the research aims to adapt a monitoring system for secondary schools. Presently such a system does not formally exist although schools may be making use of various informal systems. In essence, the components and characteristics of a monitoring system, which would be effective and efficient in the South African context, were investigated bearing in mind the vast diversity of secondary schools. Literature offers many alternatives in terms of school-level monitoring systems available. These systems, however, originate in the developed world. The key question was whether these systems or rather one particular system would be applicable within a developing world context, whether the identified monitoring system was the best option, and how this monitoring system would be received by the education system. Pragmatism provides a philosophical framework within which to work, using whichever means necessary in order to establish whether the monitoring system chosen would be a feasible option.

Pragmatism lends itself to the use of mixed methods, which provides the researcher with the opportunity to answer the research questions adequately (Teddlie & Tashakkori, 2003). By using mixed methods, one comes to a more comprehensive understanding of the phenomena under investigation by means of developing a more complete portrayal of the social world and developing fresh perspectives and new ideas. The account of research using mixed methods is also more defensible as the results are credible and there is less bias as the one method compensates for the other method. Thus one is able to develop stronger knowledge claims (Greene, 2005). The researcher for this study was able to choose from various designs in order to investigate whether MidYIS, as a feasible monitoring system, is applicable in the South African context.

Tashakkori and Teddlie (1998) state that when doing research you should study what is of interest and of value to you and undertake the study in a variety of ways appropriate to the identified aims. It is believed that by grounding the research in pragmatism the tools, which could be used to ascertain whether MidYIS is a feasible system increase dramatically. This is due to the use of various methods both qualitative and quantitative, to tease out nuances presented during the course of the research. Thus by means of including various methods and designs the research is reinforced. Moreover, an iterative process exists between the purpose of the research and the research questions. The iterative process facilitates decisions about methods. As various questions are developed from the purpose of the research, so it becomes necessary to make use of mixed methods to address the questions (Newton et al., 2003).

In summary, the first part of the chapter addressed the question of paradigmatic grounding focusing specifically on the use of pragmatism as a philosophical framework. The rest of the chapter provides a detailed explanation of design issues considered (5.2) as well as the methodology employed in the research (5.3). The methodology section of the chapter provides an overview of the sampling techniques used (5.3.1), the instruments used to collect data (5.3.2), validity issues pertaining to the instruments (5.3.3) as well as the data collection procedures undertaken (5.3.4). The methodology section concludes with a detailed description of the data analysis techniques used (5.3.5) and a summary of the research procedures (5.3.6), ending with a section discussing the ethical issues considered in the research (5.4) and the methodological constraints (5.5).

5.2 Design issues

The design issues comprise the specifics of the sample, data collection, instruments and data analysis. These, however, are determined by the research questions. As was discussed in Chapter 1 and elaborated on in Chapters 3 and 4 two main research questions have been identified for this research. The first main research question comprising three specific research questions:

1) **How appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context?**

   1.1 ***How does the Middle Years Information System (MidYIS) compare to other monitoring systems?*** Research question 1.1 was addressed by means of a literature review focusing on the characteristics of different monitoring systems and then by drawing a comparison between these and the characteristics of MidYIS. The literature used to address this question is discussed in Chapters 2, 3 and 4.

1.2 ***How valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** This research question focuses specifically on how valid and reliable, the results of the MidYIS instrument are. However, determining validity is not a straightforward procedure as in the case of reliability, which is a technical procedure (Kline, 1993). As was discussed in Chapter 4 this specific research question can be further refined to address reliability and specific facets of validity.

1.2.1 To what extent are the results obtained on MidYIS reliable?
Reliability analysis was undertaken for the assessment and where possible for the questionnaires. The aim of the analysis was to examine the extent to which the instruments are consistent across contexts as well as to strengthen investigations into construct validity. The results of the reliability analysis can be found in Chapter 7 of the dissertation.

1.2.2 *To what extent are the skills tested by MidYIS valid for the South African curriculum?* From an educational perspective, content-related validity in terms of curriculum or curricular validity is viewed as the extent to which the content of the items can be linked to the South African curriculum. In order to address issues of content-related validity (which included issues of face, content and curriculum validity), literature suggests that information is needed from specialists in the field in which the research is located and thus these specialists should be consulted (Anastasi & Urbina, 1997; Kline, 1993). For this reason, education specialists were consulted. In order to investigate issues relating to the curriculum, document analysis was undertaken (mathematics and language curriculum documents), educational consultants (mathematics and language specialists) as well as Provincial Education Department officials and National Education Department officials were approached in order to adequately explore issues of the intended, implemented, and attained curriculum.

The intended curriculum comprises system-level initiatives from the National and Provincial Departments of Education such as curriculum documents. The implemented curriculum is on a school-level and educator-level and refers to how the school as a whole and educators, in particular, interpret and implement the curriculum. The attained curriculum is what learners have learnt (Van den Akker, 2003). In the case of educational consultants, evaluation reports were required in terms of whether the MidYIS assessment adequately covers the domain, appears valid and corresponds with the intended and implemented curriculum. Provincial Education Department officials working with curriculum issues, who know the intended curriculum well and know when certain skills are introduced to learners and how they should be built upon, were also included, not only to obtain additional information pertaining to the curriculum validity but also as a form of triangulation (Newman & Benz, 1998). Provincial Education Department officials were asked to complete a questionnaire in which specific curriculum related questions were posed. However, to obtain clarity around responses one official was contacted to undertake a follow-up interview. Furthermore, the issue of curriculum and curriculum validity was investigated from an intended curriculum perspective, thus, interviews with National Department of Education officials were undertaken. Two officials were interviewed in the areas of curriculum and assessment (this decision is discussed in more detail in the sections to follow). The results of this exploration are presented in Chapter 6.

1.2.3 To what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa? In order to address issues of content-related validity, it is suggested that specialists in the field evaluate the instrument (Thorndike, 1997; Urbina, 2004). For this reason, psychologists were consulted. Chapter 6 of this dissertation provides the results pertaining to content-related validity.

1.2.4 How well do the items per sub-test function and do they form well-defined constructs? The prominence of construct validity was highlighted in Chapter 3 and in order to investigate construct validity literature suggests undertaking inferential statistics (Gronlund, 1998; Suen, 1990). Data from the assessment and questionnaires was needed in order to investigate construct validity. Statistics procedures used to investigate construct validity included Rasch analysis to examine item characteristics as well as reliability analysis. The results of the analyses pertaining to construct validity are presented in Chapter 7 of this dissertation.

1.2.5 To what extent do the data predict future achievement?

The extent to which the instruments predict performance in academic subjects is another aspect of validity, namely predictive validity, as highlighted by the specific research question. Literature suggests that predictive validity is investigated by means of correlating two sets of scores from different instruments in order to investigate the relationship between them (Grunlund, 1998; Kline, 1993). In order to explore the predictive validity, correlation analysis was undertaken where the results of the assessment instrument were correlated with the language and mathematics results as obtained at the end of the school year. Chapter 7 of the dissertation provides the analyses pertaining to the predictive validity.

1.3 ***What adaptations are needed to transform MidYIS into SASSIS, a monitoring system for the South African context?*** Based on the investigations of specific research question 1.2.2 and 1.2.3 changes and adaptations are put forward in order to transform the MidYIS monitoring system into SASSIS or the South African Secondary School Information System. Sub-questions can be identified based on the discussion in Chapter 4.

1.3.1 To what extent are the administration procedures appropriate and if not, how can they be adjusted? Recommendations of the specialists in the field were used to address this research question. However, this research question can also be linked to the sub-research question related to reliability as was discussed in Chapter 4.

1.3.2 To what extent is the content in MidYIS appropriate for second language learners? Initial results are presented in Chapter 6 of the dissertation based on the recommendation of the specialist in the field. However, this question is also addressed in Chapter 7 of the dissertation to a lesser degree.

1.3.3 To what extent is the format of the assessment appropriate and if not, how can it be changed? This research question is addressed in Chapter 6 as the format was one of the aspects which specialists in the field were asked to evaluate.

1.3.4 To what extent are the time allocations appropriate and if not, what adjustments are needed? The time allowed for each section of the assessment has implications for the quality of data. Specialists in the field were asked to evaluate the time constraints allowed for each section. The results of this evaluation are presented in Chapter 6.

1.3.5 To what extent is the feedback given in MidYIS appropriate for South Africa and how can this format be improved upon? The feedback reports, which form part of MidYIS, were addressed in Chapter 4. This question is elaborated on in Chapter 9 of the dissertation.

The first main research question addresses issues of validity and reliability. The second main research question is an extension of the first main research question if MidYIS is valid, with the necessary adaptations, and reliable then which factors on a school, classroom and learner-level could have influenced learner performance.

2) **Which factors could have an effect on learner performance and therefore inform the design of the monitoring system?** This research question addresses the exploration of variables that could possibly have a significant effect on the performance of learners. When investigating the effect of factors on performance, literature suggests that some form of regression analysis be undertaken (Newton & Rudestam, 1999). In school effectiveness research multilevel analysis, which is a form of regression analysis, is undertaken to investigate which factors are associated with performance of learners (Scheerens, 1990; Scheerens & Bosker, 1997; Riddell, 1997; Sammons, 1999; Scheerens, 2001a). Likewise, in this research, multilevel analysis was undertaken to investigate factors on a school, classroom, and learner-level that have an effect on learner performance. Four specific research questions can be identified namely:

2.1 ***What factors on a school-level affect the performance of learners on the assessment?*** Prominent school-level factors were identified from literature (Scheerens, 1990; Scheerens & Bosker, 1997; Bosker & Visscher, 1999; Marsh & Willis, 2003) and included in the principal questionnaire. The factors not only are prominent in literature but also had to correlate with the results of the assessment. Significant factors were retained and were included for exploration by means of multilevel analysis (see Chapter 8 for details).

2.2 ***What factors on a classroom-level affect the performance of learners on the assessment?*** Relevant factors from literature were identified (Scheerens & Bosker, 1997; Bosker & Visscher, 1999; Sammons, 1999; Marsh & Willis, 2003) for inclusion in the educator questionnaire. The results from the educator questionnaire were correlated with assessment data. Only the most significant factors were retained for further analysis using multilevel modelling (see Chapter 8 for details).

2.3 ***What factors on a learner-level affect performance of learners on the assessment?*** Literature suggests that several factors affect performance (Anderson, 1988; Lens, 1994; Anderson, 1994; Howie, 2002). Some of these factors were already present in the questionnaire as designed as part of the MidYIS system (Extended MidYIS and SATIS). Additional information pertaining to frequency of events in the classroom was added. The factors in the questionnaire were explored and correlated with the assessment data. Significant factors were included in the multilevel model (see Chapter 8 for details).

2.4 ***How can the identified factors be included in the design of the monitoring system?*** This research question draws on the results presented in Chapter 8. Suggestions on how the significant factors can be incorporated into the monitoring system are presented.

Given the exploratory nature of the research and aims identified in Chapter 1, the research was approached with an open mind in terms of using complementary methods. With pragmatism underpinning the research design, both quantitative and qualitative methods were used to answer the identified questions and mixed methods were adopted. According to Johnson and Onwuegbuzie (2004, p. 16), “the bottom line is that research approaches should be mixed in ways that offer the best opportunities for answering important research questions”. Johnson and Onwuegbuzie (2004, p. 17-18), go on to say “…research methods should follow research questions in a way that offers the best chance to obtain useful answers”. Figure 5.1 illustrates a mixed method process model.

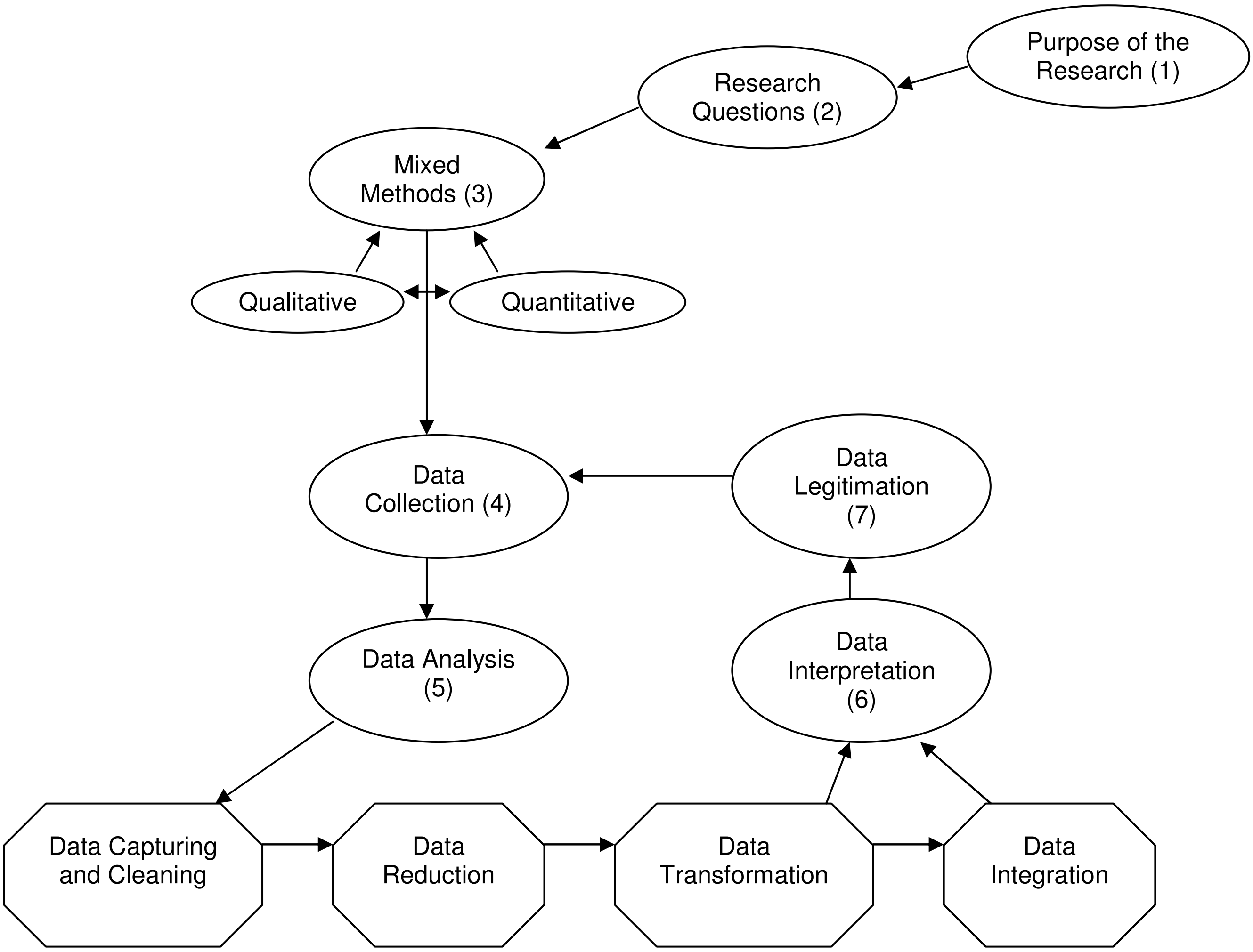

Figure 5.1 **A mixed method process model (adapted from Johnson and Onwuegbuzie, 2004)**

Mixed methods intentionally combines different tools and techniques to gather, structure, analyse and interpret quantitative and qualitative data (Williams, 1999), as illustrated in Figure 5.1. Mixed methods can answer questions, which other methodologies in isolation cannot, for example in the investigation of validity issues, which in itself is a complex task (Kline, 1993), where a combination of methods can be used in order to provide different perspectives on the same issue making inferences stronger. Thus the issue of validity can be addressed quantitatively by using inferential statistics to investigate construct validity, to undertake reliability analysis, and to investigate factors affecting performance. Validity can be addressed qualitatively by undertaking an analysis of curriculum documents, follow-up interviews with the Provincial Department of Education officials and interviews with National Department of Education officials.

Furthermore, mixed methods comprise various design dimensions (Morse, 2003, Greene, 2005). The typology used for this research was a concurrent nested strategy. A concurrent nested strategy implies that there is a dominant method that guides the research. In the case of this research a quantitative approach is the dominant method. The qualitative component was given lesser priority but was nested within the quantitative approach. The qualitative approach was embedded in the quantitative approach as the method addresses a different aspect of the question relating to validity and seeks information from a different level. While the quantitative approach makes use of information at the school, classroom, and learner-level, the qualitative approach makes use of information at the provincial and national-levels. According to Creswell (2003) the use of a concurrent nested strategy provides a broader perspective and makes it possible to study different groups and different levels simultaneously. In terms of this research for example, a quantitative approach was used to investigate issues of construct validity, predictive validity, and significant factors influencing performance all of which can be attributed to a school, classroom, and learner-level. However, a qualitative approach was used to investigate issues of curriculum validity, specifically at national and provincial levels focusing on the intended curriculum. In both instances, information from one level builds upon information from the other levels (Creswell, Plano Clark, Gutman & Hanson, 2003).

Several advantages can be identified when using a concurrent nested strategy:

- The researcher can collect quantitative and qualitative data at the same time (Creswell et al., 2003).
- The integrity of both methods is maintained as the assumptions underlying each method are not violated (Morse, 2003).
- The advantages of using quantitative and qualitative approaches can be exploited (Creswell et al., 2003).
- The researcher can provide different perspectives from different types of data and from different levels within one study (Creswell et al., 2003) working both inductively and deductively to accomplish the aims of the research (Morse, 2003).

5.3 Methodology

In the section to follow issues pertaining to the methodology followed in the research is discussed namely the sample included (5.3.1); instruments used (5.3.2), validity issues (5.3.3), data collection (5.3.4) and data analysis (5.3.5).

5.3.1 Sample

Eleven schools were purposefully selected in the Gauteng Province to participate in this project, the schools were sample for maximum variation (according to Patton (2002) this is called a maximum variation sample which is a form of purposive sampling). Purposive sampling is a non-probability sampling technique. Its aim is to select candidates with a specific purpose in mind (Neuman, 1997). In this case, due to financial constraints only a limited number of schools could be accommodated. As the aim of the research was to develop a monitoring system, which would be appropriate for secondary schools regardless of the diversity of schools, it is was imperative to include schools from a variety of environments, including demographic variations in learners, educators, surrounding communities and access to funding. Thus three former White suburban schools of which two were English medium and one school dual medium (English-Afrikaans) were included as well as three former African township schools (ex-Department of Education and Training), two former Indian schools (ex-House of Delegates) and finally two former Coloured schools (ex-House of Representatives). Two Grade 8 classes from every school were randomly selected by means of WinW3S (IEA, 2005). Therefore, all learners had an equal and independent chance of being selected (Gay & Airasian, 2003). WinW3S is a within-in school sampling package developed by the Data Processing Centre of the International Association for the Evaluation of Educational Achievement (IEA). Special permission was obtained to use the program as the program is normally only used in IEA studies. In total 794 learners from the two classes in each school participated. The characteristics of the realised sample are discussed in detail in Chapter 7.

In addition, all eleven principals from participating schools were asked to complete a questionnaire as well as the mathematics and language educators of the selected classes (44 educators if each class had a different educator). Ten principals' questionnaires were received while 36 (out of the possible 44) questionnaires of mathematics and language educators were returned.

Apart from the eleven participating schools, two General Education and Training officials specialising in the areas of mathematics and language were asked to participate as well as a representative from the Gauteng Department of Education Office for Standards in Education (OFSTED). In addition, two national government officials in the fields of curriculum and assessment participated. Members of the provincial and national government were purposefully selected for their specialisation in the fields of curriculum, specifically language and mathematics and assessment.

5.3.2 Instruments

The instruments discussed in the section to follow pertain to both research questions. The first main research question is ***to what extent is the Middle Years Information System (MidYIS) appropriate as a monitoring system in the South African context?*** MidYIS consists of an assessment instrument and learner questionnaire as was discussed in Chapter 4. The learner questionnaire as well as the newly developed educator and principal questionnaires are of relevance for the second main research question, namely ***which factors could have an effect on learner performance and therefore inform the design of the monitoring system*** as the factors identified will be taken out of the questionnaires.

5.3.2.1 Assessment instrument

The assessment instrument consists of seven sub-sections, which were collapsed into four different scales namely the vocabulary scale, the mathematics scale, the skills scale, and the non-verbal scale each of which was designed to measure certain skills and abilities as discussed in Chapter 4. The seven sub-tests were timed and consist of multiple-choice items with the exception of the mathematics sub-test, which included both constructed response items and multiple-choice items (please refer to Appendix A for a description of sub-tests and number of items included). The assessment itself was a combination of a speed assessment and power assessment. Speed assessments measure not only the achievement, but also the speed with which participants perform tasks and the difficulty of tasks are manipulated through timing. A power assessment on the other hand has no time limit and difficulty is manipulated by increasing or decreasing the complexity of items. As the assessment is a combination of a speed assessment and a power assessment, the time limits typically allow the majority of participants to attempt most or all of the items (Urbina, 2004).

5.3.2.2 Questionnaires

Over and above the assessment instrument, various background questionnaires formed part of this project, including a learner questionnaire, educator questionnaire and principal questionnaire.

Learner questionnaire. The CEM centre designed the questionnaire for learners as part of another project called Student Attitudes Information System (SATIS) in which background information and attitudes of learners towards school were collected. The questionnaire included items pertaining to the demographic characteristics of the learner as well as attitudes towards school life, school class, future aspirations, home and family life, use of substances such as alcohol, personal or traumatic events that could have affected school work, school climate, particularly safety and finally motivation to achieve, motivation to continue learning and peer attitudes. The questionnaire was comprehensive in nature, but additional items on aspects pertaining to instructional practices of educators were included for triangulation purposes. The additional items which were included were based on school effectiveness literature (Newmann, 1991; Scheerens, 1992; Scheerens & Bosker, 1997; Mortimore, 1998; Sammons, 1999; Grobler et al., 2001; Howie, 2002; Harris & Chapman, 2004) or taken from developed questionnaires, such as the Third International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) learner questionnaires (refer to Appendix B for a description of subsections and number of items included). The items included were operationalised concepts as discussed in Chapters 2 and 3. They were mostly closed-ended items with the exception of three items that were open-ended.

Educator questionnaire. Literature (Scheerens, 1990; Newmann, 1991; Marsh, 1992; Teddlie, 1994c; Scheerens & Bosker, 1997; Sammoms et al., 1998; Sammons, 1999; Scheerens, 2001a; Willm & Somers, 2001) was used to identify factors on a classroom-level that have shown to effect learner performance. The educator questionnaire was developed not only to collect background information but also to ascertain educator attitudes in accordance with literature. The questionnaire included items pertaining to the age of the educator, qualifications and teaching experience, quality of instruction and instructional methods, revised national curriculum, assessments practices, opportunities to learn, challenges experienced, feedback and reinforcement resources, professional development, school climate, monitoring at classroom-level and attitudes towards the school and work. Developed questionnaires, which included the identified factors, were consulted in order to find exemplar items. Items pertaining to identified factors which were not included in already developed questionnaires were constructed by using school effectiveness literature as a departure point (refer to Appendix C for a description of subsections and number of items included). The following questionnaires were used in the development of the educator questionnaire; however, it is pertinent to mention that the items used in this research were adapted from the items in the questionnaires mentioned below:

1) School Achievement Indicators Program educator questionnaire.
2) Education Quality and Accountability Office educator questionnaire.
3) The Third International Mathematics and Science Study-1999 educator questionnaire.
4) Education Quality and Accountability Office Grade 3 and 6 Assessment of reading, writing, and mathematics.

The items included were mostly closed-ended items with the exception of two items, which were open-ended.

Principal questionnaire. The school or principal questionnaire was developed in order to collect background information as well as information pertaining to attitudes of principals. The principal questionnaire was based on school effectiveness literature (Scheerens, 1990; Bliss, 1991; Newmann, 1991; Marsh, 1992; Scheerens, 1992; Teddlie, 1994a; Teddlie, 1994c; Scheerens & Bosker, 1997; Sammons, Thomas, Mortimore, Walker, Cairns & Bausor, 1998; Grey et al., 1999; Heck, 2000; Scheerens, 2001a; Wills & Somers, 2001; Hill, 2001; Howie, 2002) or made use of exemplar items from already developed questionnaires. Items relevant to factors identified from literature which were not included in already developed questionnaires were constructed by using school effectiveness literature as a departure point (refer to Appendix D for a description of subsections and number of items included). The following questionnaires were used in the development of the principal questionnaire; the items, however, were adapted:

1) The Third International Mathematics and Science Study-1999 principal questionnaire.
2) Education Quality and Accountability Office principal questionnaire.

The principal questionnaire included items pertaining to the school's attitude toward achievement and approach towards assessment, leadership style, school climate, curriculum development and design, professional development, monitoring at school-level, resources, parental involvement and the impact of intended policies such as *Whole School Evaluation* and *Systemic Evaluation* (refer to Appendix D for a description of subsections and number of items included). The items included were mostly closed-ended items with the exception of two items that were open-ended.

***Provincial-level questionnaire*.** The provincial-level questionnaire was a short questionnaire developed for curriculum specialists within the Provincial Department of Education in order to collect information pertaining to the intended curriculum as discussed in 5.2. The questionnaire included items pertaining to assessment practices, and use of developed assessments, issues related to curriculum validity, items related to skills development in terms of the curriculum and background information such as age, gender, qualifications, previous as well as current employment. The questionnaire consisted of both closed-ended items as well as open-ended items.

5.3.2.3 Interview schedules

National Department of Education. The aim of the interview schedule for the National Department of Education was to collect information on assessment and curriculum issues including policy, making use of developed assessments, strategies advocated to ensure curricular or curriculum validity and issues pertaining to monitoring. The schedule was semi-structured in that although the questions had been formulated and the order determined the order as well as the questions were modified during the interview as deemed appropriate. The questions included were open-ended with the responses recorded and taped by the interviewer (Gay & Airasian, 2003).

Provincial Department of Education. The questionnaire developed for Provincial Department of Education officials was used in a telephonic interview. This was done to obtain clarity on the answers provided in questionnaire. The questionnaire lent itself to be used as an interview schedule due to the open-ended questions.

5.3.3 Validity issues related to the instruments

Validity ascertains the extent to which the interpretations of results are appropriate as well as meaningful. Validation is the process whereby validity evidence is gathered (Urbina, 2004). Issues pertaining to the validity of research are an important aspect, more so now that there are a variety of methodological choices available. According to Newman et al. (2003, p. 167):

> …researchers strengthen validity …when they can show the consistency among research purposes, the questions, and the methods they use. Strong consistency grounds the credibility of research findings and helps to ensure that audiences have confidence in the findings and implications of research studies.

Validity per se has the following characteristics (Gronlund, 1998; Linn & Gronlund, 2000):

- It is inferred from evidence and ultimately depends on many different types of evidence from which inferences are drawn.
- It is expressed by degree, in terms of high, moderate and low and is specific to a particular use.
- It is a unitary concept that is based on various forms of evidence, with construct-related validity being the central concept, and ultimately it is concerned with the consequences of using the assessment or questionnaire.

Among the factors adversely affecting validity, are:

- Tasks included that inadequately sample the domain to be assessed or do not function as they were intended. This could be due to lack of relevance, inappropriate difficulty, or bias (Gronlund, 1998).
- Inadequate administration causing directions to be unclear or inadequate time allowed, resulting in the collection of skewed data, thus lowering the validity of the results (Gronlund, 1998).
- Validity can be influenced if the items are poorly constructed, if there is an overemphasis of certain aspects, or if there is an identifiable pattern of answers (Linn & Gronlund, 2000).

In the section to follow, the aspects of validity are discussed separately for the assessment instrument (5.3.3.1), interview schedules (5.3.3.2), and contextual questionnaires (5.3.3.3).

5.3.3.1 Validation of assessment instrument

The concept of validity is multifaceted in terms of forms of evidence and the interpretation of validity differs depending on the context in which it is used. To illustrate the point, if one was looking at the degree to which items are spread over a particular domain (*content-related validity*) one would not interpret the information from the perspective of whether or not the results could be used to predict future achievement (*criterion-related evidence, predictive validity*). Validity in this research is seen as a unitary concept, even though it comprises several facets. The facets are highlighted according to the aims of the study and may be content-related validity (which includes face and curriculum validity), predictive validity, or construct validity (construct-related validity). Validity here is seen as a property of interpretation (Gronlund, 1998; Linn & Gronlund, 2000).

In the context of this research, the different types of validity were viewed as categories for accumulating evidence to support validity claims. Various forms of evidence are discussed namely in terms of content-related validity from a psychometric perspective, in terms of face validity and content validity and from an educational perspective in terms of curricular validity. Other forms of evidence included construct-related validity, and criterion-related evidence using a predictive study or predictive validity.

Traditionally, *face validity* is the extent to which an assessment looks as if it measures properties of constructs (Anastasi & Urbina, 1997). Face validity is considered a non-professional's assessment of validity. It refers to whether the assessment or questionnaire looks valid and is a desirable feature in assessments and questionnaires since it can increase the motivation of participants and high motivation is essential for valid testing (Kline, 1993). Face validity or the superficial appearance of what the test measures from the perspective of the participant is subsumed under content-related validity (Urbina, 2004).

C*ontent-related validity* is generally understood as the extent to which the questions in the assessment match the field within which the assessment can be located (Coolican, 1999). Thus the sampling of items from the broader domain is important (Gronlund, 1998) in terms of relevance as well as of representativeness (Urbina, 2004). Factors of importance when evaluating content-related validity are the emphasis of content areas and objectives, types of items included, number of items included and the appropriate difficulty level of items (Thorndike, 1997).

The MidYIS instrument is a developed abilities assessment and specialists are consulted to map the overlap of developed abilities tests domain and items included in the assessment. However, when administering a developed abilities test within an educational setting the curriculum and the abilities that are supposed to be taught (according to the curriculum) are important. Content-related validity refers specifically to the match of items and content domains and a different facet of content-related validity is necessary in addition to examining the extent to which items are adequately represented from the domain of abilities. This type of validity is called curricular or curriculum validity and refers to the extent to which the abilities or competencies assessed match the curriculum (Thorndike, 1997). Curriculum validity is of particular importance in this research as it attempts to map the tested competencies or abilities onto what the learners have been exposed to in the curriculum, especially in light of South Africa using an outcomes-based or rather competency-based curriculum. Content-related validity, which includes face validity and curriculum validity, is evaluated by means of drawing up tables of specifications or by consulting content specialists (Suen, 1990).

Criterion-related validity, in terms of traditionally *predictive validity*, examines the degree of relationship between the assessment scores and the criteria to be predicted (Gronlund, 1998) in order to estimate future levels of performance (Urbina, 2004). Therefore, predictive validity is the degree to which assessment scores can predict future scores (Coolican, 1999). Predictive validity is investigated by means of correlation analysis. When a correlation coefficient is used to ascertain validity, it is referred to as a validity coefficient. In addition to correlation analysis, expectancy tables can also be used to illustrate the relationship between two measures. The relationship can be represented in a twofold chart with the assessment scores arranged in categories down the left side of the table and the measure to be predicated arranged in categories across the top (Gronlund, 1998).

Construct validity or construct-related evidence on the other hand refers to whether the assessment possesses certain psychological characteristics (Gronlund, 1998) that are indicators of the theoretical construct of interest (Suen, 1990). These characteristics are hypothetical qualities or constructs which include a description of the theoretical framework that specifies the nature of constructs, description of development of the assessment and aspects of measurement and the pattern of relationships. Thus construct-related validity is the extent to which the results support the theory behind the research (McBurney, 1994). Construct validation includes all categories of evidence drawing upon the comparison of the sample of assessment tasks to the domain of tasks. This includes examining test features and their influence on the meaning of scores, determining the internal consistency of the assessment and correlating assessment scores with other assessments that measure the same construct or making use of other academic results (Gronlund, 1998).

Expert opinion as a validation strategy. One way in which to investigate content-related validity is by judgements regarding the degree to which the assessment adequately samples a particular content domain from a psychometric perspective or curricular domain from an educational perspective (Murphy & Davidshofer, 1994; Thorndike, 1997). For this research, it was important to examine how well the content of the assessment was aligned with both the abilities domain (Linn & Gronlund, 2000) and curricular domain (Thorndike, 1997). The judgement process can be formal or informal in nature. The least formal process used was a casual overall impression as to whether the assessment appears to measure what it was suppose to measure, thus looking at the face validity. A formal process entailed a systematic procedure of consulting content area specialists (Suen, 1990). While the judgements concerning face and content-related validity are neither final nor absolute, these judgements are not arbitrary (Murphy & Davidshofer, 1994). For this research, to ensure content-related validity the assessment was given to three education specialists, an educational psychologist and two research psychologists to assess.

The use of test-curriculum overlap as a validation strategy. Test-curriculum overlap (TCO) is the extent to which the content of the assessment is present in the curriculum (De Haan, 1992). De Haan distinguishes five (5) different approaches to TCO:

❖ One could *directly observe* what is being taught and whether the content is covered by the assessment.

❖ Alternatively, one could follow a *taxonometric approach* where both assessment content and curriculum content are mapped on a taxonomy, which consists of topics and intellectual processes or skills. The taxonometric approach lends itself well to ascertain the overlap of the assessment content and the intended curriculum.

- If either of the two approaches already mentioned are inappropriate, another alternative is to undertake *text analysis*. In text analysis, textbooks used by the sample learners are analysed and similar items as those in the assessment are identified. However, as it is not possible to ascertain whether all items in the textbook are exercised, this type of analysis can only give an indication of the assessment content and formal or intended curriculum.
- The fourth type of analysis is *instructional analysis*. By undertaking instructional analysis one tries to obtain estimations from educators, learners and curriculum experts as to whether the content in the test was or should be taught.
- Another approach is to make use of *data analytic strategies* in which one tries to assess the content coverage of tests by identifying unusual response patterns of individual learners.

Regardless of which approach one selects to undertake an analysis of the test-curriculum overlap, the reason behind such an analysis remains the same. The reason is always to ascertain the validity of such assessments: does the assessment measure what it is supposed to measure and is the assessment instrument suited to the study and fair to participants?

When reflecting upon the five approaches to TCO, it would seem that an either or situation is not preferable if curricular validity is to be assessed. Only a combination of approaches would strengthen any claim made. Two approaches were deemed best suited to this research, namely a taxonometric approach and instructional analysis. A taxonometric approach lends itself to document analysis and the development of a framework in which to match skills and curriculum, whilst instructional analysis lends itself to specialist evaluation. Specialists were asked to assess the assessment and also to assess the extent to which what had been assessed should have been taught. Therefore, the extent to which the content of the assessment was aligned with curriculum goals were examined (Linn & Gronlund, 2000). Both approaches focused on the intended curriculum.

The documents included for analysis were the national curriculum policy documents. These documents provide a snapshot of what the intended curriculum is according to the national government. The intended curriculum, namely what ideally should be taught (Van der Akker, 2003), was explored in this study as curricular validity refers to the extent to which the content of the assessment is aligned with curriculum goals. Only relevant policy documents were analysed, specifically the language and mathematics policy documents. These were important because the content of the assessment, although a developed abilities test assessing basic skills, is related to the areas of language and mathematics and because the National Department of Education has identified literacy, mathematics, and science as key areas of intervention (South African Yearbook, 2003). Elements of literacy and mathematics are directly covered in the assessment. By providing insight into the skills and competences that should be addressed per grade level, specifically for language and mathematics, the curriculum documents serve as a source of information on what skills and competences should be taught.

Specialists in the field evaluated the assessment, using the instructional analysis approach. Two language specialists and two mathematics specialists were approached. Due to unforeseen circumstances one of the mathematics specialists was unable to finish the review. One language specialist is a practitioner and head of curriculum development at the school where she works while the other is a member of the Faculty of Education, University of Pretoria who has specialised in languages and has a wealth of teaching experience at both school and university level. The mathematics specialist is a research consultant, who has worked in the field of mathematics for a number of years, and has authored a number of mathematics textbooks.

In addition to document analysis and specialist evaluation, two National Department of Education officials were interviewed in order to obtain additional information on issues pertaining to curricular validity and monitoring systems. Moreover, two Provincial Department of Education officials specialising in language and mathematics were asked to complete questionnaires. The questionnaires were developed based the results from the document analysis and specialist evaluation. Finally, one representative from OFSTED was interviewed. Both the questionnaire and the interview focused specifically on skills and competences in the curriculum.

Validity strategies for the qualitative data as part of test-curriculum overlap. Validity in qualitative research is described in terms of the trustworthiness, relevance, plausibility, credibility, or representativeness of the research (Babbie & Mouton, 1998; Lincoln & Guba, 1985; Trochim, 2001). The validity of the research is located with the representation of the participants, the purpose of the research and the appropriateness of the processes employed (Winter, 2000).

Validity for the qualitative component of this research has to do with the adequacy of the researcher to understand as well as represent the participants' meaning. Thus validity becomes a quality of the knower in his/her relation to the data, enhancing different vantage points and forms of knowing (Tindall, 1990). It raises questions about the validity of the results (Trochim, 2001). Validity in qualitative research is personal, relational, as well as contextual in nature. How the research was conducted was of importance in terms of whether the researcher was aware of her own perspective, processes, and the influence of these on the research (Marshall, 1986). When considering issues of validity in qualitative research it is accepted that one's impression of what truth is will determine how one views the trustworthiness, accuracy and reliability or dependability of the research (Winter, 2000). The examination of how one's own truth influences the research process is known as reflexivity (Tindall, 1990).

The notion of reflexivity in this research is important. It is not only consistent with the underlying paradigm, pragmatism, but is also an important component of qualitative research. Reflexivity is consistent with pragmatism, as pragmatism is concerned with the value-ladeness of research and calls for the researcher to select a research approach that would reflect what s/he deems important. Also of importance here is personal reflexivity, which refers to aspects of the researcher's identity and the fact that research undertaken is very often an expression of personal interests and values (Wilkinson, 1988, Tindall, 1990*).* In terms of mixed methods, an important aspect is that the assumptions of the methodology used must not be violated (Morse, 2003). Reflexivity is a vital part of the qualitative research process. Likewise the credibility, dependability, and conformability of the research results are of importance as indicators of the trustworthiness of the research (Babbie & Mouton, 1998).

Credibility is similar to the concept of internal validity (Lincoln & Guba, 1985). It refers to procedures aimed at ascertaining whether the interpretations of the data are compatible with the constructed realities of the participants (Babbie & Mouton, 1998). Although many procedures exist to ascertain credibility of interpretations such as peer debriefing or member checking, triangulation was used in this research.

Triangulation is the use of two or more methods of data collection in order to gather information on an aspect of behaviour (Cohen, Manion & Morrison, 2004). It reduces the risk of chance associations and bias as well as assists in formulating better explanations (Maxwell, 1996). The aim of triangulation was to explain the complexity of behaviour or a phenomenon in a comprehensive manner by studying it from more than one standpoint. For the purposes of this research, method triangulation (Tindall, 1990) was applied by using multiple instruments and by using different respondent categories. As triangulation allows for the illumination of different vantage points, investigator triangulation, using more than one researcher, was used in order to reflect on multiple viewpoints (Tindall, 1990).

Triangulation is a procedure also used to establish the credibility of interpretations. Procedures similar to those used to establish credibility could also be used to establish the dependability of results. Dependability is very similar to reliability in quantitative studies. It refers to the provision of evidence that if the inquiry were to be repeated with the same or similar people in the same or similar circumstances, the findings would be the same. Assertions pertaining to dependability are strengthened by means of an inquiry audit in which an “auditor” examines documentation on which the findings are based, in order to attest to the dependability and confirmability of results. Such documentation include interview and process notes (Babbie & Mouton, 1998).

An audit trail is a record of decisions made and processes followed during the data analysis process. The aim of the audit trail is to enable the auditor to establish whether the interpretations, conclusions, and recommendations are rooted in the data (Babbie & Mouton, 1998). Different kinds of documentation are required to undertake an audit trail ranging from raw data to process and data reconstructions. The audit trail (refer to Appendix E) for this research comprised (Lincoln & Guba, 1985):

- ***The raw data*** that included the transcripts of the interviews with National Education Department officials as well as the policy documents used.
- ***Data reduction and analysis products*** that included the theoretical notes taken from literature as well as the coding system used (presented in the form of a concept map).
- ***Data reconstruction and synthesis products*** that included the categories of themes and relationships as well as the conclusions drawn.
- ***Instrument development information*** that included the interview schedule and provincial education questionnaire.

5.3.3.2 Validation of the questionnaires

Content-related validity was used in validating the questionnaires. Content-related validity was discussed in detail under the validation of the assessment. Of importance is the extent to which the items in the questionnaires match the field within which the questionnaires can be located (Coolican, 1999), in this case school effectiveness and school improvement research. In order to establish whether the items match the field and whether the items looked valid, specialists in the field were consulted. Two researchers working in the fields of school effectiveness and school improvement reviewed all the questionnaires used in this research. Both researchers have extensive knowledge of survey research, having worked on cross-national studies such as the Third International Mathematics and Science Study (TIMSS). One of the researchers is the Director of the Centre for Evaluation and Assessment and National Coordinator for the Progress in Reading Studies (PIRLS). While the other researcher has been chairman of the IEA for a number of years and is currently coordinating the Second International Technology in Education Study (SITES).

5.3.4 Data collection

The data collection in this study comprises various forms of evidence ranging from document analysis, assessments and questionnaires to interviews (refer to Table 5.2). In the section to follow, the data collection process for each of the different types of evidence is described (refer to Chapters 2, 3 and 4).

Table 5.2 *Summary of research questions, sources and data instruments*

Research questions	Specific research question	Sources and participants	Data strategies and instruments				
			Document analysis	Evaluation reports	Partially structured	questionnaire	assessment
1. How appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context?	1.2 How valid and reliable are the data generated by the MidYIS monitoring system for South Africa?	Policy documents	X				
		-Specialists in the area of psychology and education		X			
		National Department of Education officials			X		
		Provincial Department of Education officials			X	X	
		Learners				X	X
2. Which factors could have an effect on learner performance and therefore inform the design of the monitoring system?		Principal				X	
		Educator				X	
		Learner				X	X

5.3.4.1 Document analysis

Document analysis was undertaken in order to establish the test-curriculum overlap of the skills tested in the assessment and the skills that were taught in the curriculum. The analysis pertained to the investigation of curriculum validity as highlighted by the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** The documents included for the analyses were the South African language learning area curriculum policy document as well as the mathematics learning area curriculum policy document. The documents were imported into Atlas *ti,* and analysed by means of identifying themes of the skills learners were meant to be taught, the results were used in conjunction with the evaluation reports (see below) in order to make inferences with regard to the test-curriculum overlap.

5.3.4.2 Evaluation reports

The evaluation report forms part of the validity strategy employed in this research, specifically focusing on issues of content-related validity as highlighted by the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** In order to investigate the different aspects of validity, specialists in the field of psychology and education were approached. Two research psychologists as well as an educational psychologist evaluated the assessment instrument for content-related validity. The psychologists were asked to complete an evaluation form relating to issues of language, bias and content covered (refer to Appendix F). A meeting was scheduled to discuss the results of the evaluation and process notes taken.

Specialists in the field of education, specifically in mathematics and language, were also approached and the assessment was evaluated from a curriculum perspective. The specialists were asked to complete an evaluation form similar to the form used for the psychologists. Issues of curriculum relevance were covered and a table of specification drawn up in order to identify the difficulty of items. Grade level introduction of the content were included from a theoretical perspective (refer Appendix G). Once the evaluation task was completed, a meeting was scheduled with each specialist to discuss the results of the evaluation and process notes taken.

5.3.4.3 National-level data collection

Two national government officials were visited for face-to-face interviews with the purpose of obtaining additional information on issues pertaining to curriculum validity and monitoring systems more generally. The interviews with National Education Department officials is related to curriculum validity issues as highlighted by the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** The participants were contacted telephonically to request that they participate in the research project. Details of the project were provided and the aim and background of the interview itself were explained. Forty-five minutes were requested; however, the interviews were completed in 30 minutes. The interviews were conducted using a semi-structured interview schedule, which was emailed to the participants before the interview took place. The interview was recorded with the permission of the participants.

5.3.4.4 Provincial-level data collection

The purpose of this data collection was to provide additional information pertaining to curriculum validity and it relates to the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** The provincial-level education officials were first contacted telephonically to request their participation in this research. Background information was provided on the project as a whole and the purpose of the questionnaire explained. Once the officials had agreed the questionnaires were emailed and faxed to them. Upon completion, the questionnaires were emailed and faxed back to the researcher (refer to 5.3.2 for a discussion on topics covered in the questionnaire).

The OFSTED representative was contacted telephonically to ascertain whether if the official would be willing to participate in this research. During the telephonic conversation, information on the project was provided - the purpose of the questionnaire as well as the follow-up interview. Once the official had agreed to participate, the questionnaire was emailed to the OFSTED official. That was followed-up with a telephone interview using the questionnaire as an interview schedule. The telephone interview lasted 15 minutes (refer to 5.3.2 for a discussion of topics covered in the questionnaire and follow-up interview).

5.3.4.5 School-level data collection

Each school was visited on a separate day and fieldworkers administered the instruments. The instruments included the learner assessment, learner questionnaire, principal questionnaire and questionnaires for the mathematics and language educator of the two classes selected (refer to 5.3.2 for a discussion on topics covered in the instruments). Each classroom had a fieldworker overseeing the standardised administration procedure. The fieldworker read a script explaining the assessment and questionnaire as well as the time limits for each sub-section. This ensured that the administration procedures were standardised across the schools and that each learner received exactly the same information. The assessment as well as the questionnaire took approximately two and a half hours to complete. The English script was translated into Sepedi and Afrikaans (the two additional languages of instruction for the sampled schools) to ensure that each learner would understand what was expected. Two groups of translators were used for the translation of the administration script. The first group translated the English script into Sepedi and Afrikaans, while the second group of translators checked the Sepedi and Afrikaans translations against the English version. Any changes or corrections were made in consultation with the specialists before the scripts were finalised. Thus administration of the instructions for assessments took place in English, Sepedi, and Afrikaans depending on the school that was visited. The assessment itself, however, was in English.

In order to capture the administration process the fieldworkers completed an administration questionnaire detailing the administration process, including problems experienced, comments made by learners and general impressions and time taken for the majority of learners to complete the sub-sections.

While learner assessments were conducted, the principal and educators were asked to complete background questionnaires. However, in certain instances the principal and educators could not completed the questionnaires in the time taken for the learner data collection. In these cases, the questionnaires were either collected later or faxed to the researcher.

5.3.5 Data analysis

Data analysis is the vehicle used to generate and validate interpretations, formulate inferences, and draw conclusions. In this study, parallel mixed methods analysis was used. In parallel mixed methods analysis, interpretation and writing up of the qualitative and quantitative data are undertaken separately (Onwuegbuzie & Teddlie, 2003). For the qualitative data, such as curriculum documents and semi-structured interviews, document analysis and thematic content analysis were undertaken (5.3.5.1 and 5.3.5.2). The assessment data was analysed by means of descriptive statistics, Rasch analysis, reliability analysis and correlation analysis (5.3.5.3) while the contextual data was analysed by means of descriptive statistics, reliability analysis (5.3.5.4). The results from the assessment as well as variables from the contextual data were used to build a multilevel model (5.3.5.5).

The data analysis for the quantitative component was undertaken by means of the Statistical Package for the Social Sciences (SPSS) unless otherwise indicated, while the qualitative component was undertaken using Atlas *ti*. Atlas *ti* is a qualitative data analysis tool that is classified as coding and theory building software (Barry, 1998). It is therefore acceptable for this research as thematic content analysis is used. The only other software package that is classified as a coding and theory building software is Nudist. Atlas *ti* was chosen for this research, however, and for the following reasons

- ❖ Its visual and spatial qualities, creativity, and the ability to interlink ideas (Barry, 1998).
- ❖ The researcher is able to visualise relationships between different parts of the data

and theoretical ideas, enabling pattern recognition (Barry, 1998).

- It is easy to work with and easier to learn, especially at a basic level of operation, much easier than Nudist (Barry, 1998).
- It is ideal for less complex projects (Barry, 1998) and thus ideal for the purposes of this research. It provides easy access to documents, quotations, codes, and memos and allows the researcher to work with data in the form of text, graphics, or sound (Henning, van Rensburg & Smit, 2004).
- It provides researchers with a code-retrieve function and essentially provides support for theory building (Henning et al., 2004).
- It enables researchers to connect codes in order to facilitate higher-order classifications and categories (Henning et al., 2004).
- It provides a platform for the facilitation of cross-referencing of data and enables the researcher to develop networks to describe relationships (Henning et al., 2004).

In the section to follow the analysis techniques used for the document analysis, qualitative data and the assessment data as well as the questionnaire data are discussed in detail.

5.3.5.1 Analysis of documents

Document analysis was undertaken to provide information pertaining to the curriculum validity of the assessment. Curriculum validity is a crucial component of the specific research question 1.2 ***(how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?)****.* The curriculum policy documents were imported into Atlas *ti* and a thematic content analysis undertaken. The procedure followed for the thematic content analysis is discussed in detail in the section to follow, analysis of qualitative data (5.3.5.2). Both an inductive and a deductive approach to analysis were followed. A deductive approach was followed because the structure of the document, in terms of learning areas and learning outcomes, was used as an overarching framework. An inductive approach was used because skills were identified and directed by the text of the curriculum documents.

However, themes that were coded in text generally referred to the type of skills the curriculum tries to develop and the way in which the skills are developed. As mentioned earlier, the developed abilities assessment is essentially a psychological assessment, yet the relevance to the curriculum also had to be established. Thus the various skills the curriculum aims to develop constitute an important link that needed to be explored. Specific skills, such as problem solving, skimming and scanning capabilities and proof reading, were focused on.

5.3.5.2 Analysis of qualitative data

Semi-structured interview schedules were used in the research in order to obtain data relevant to the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** The interviews were recorded so that transcriptions could be generated. As in any analysis, data reduction must take place. The procedure includes organising data so that emerging themes or patterns can be identified (McRoy, n.d.). Data reduction however was not separate from the analysis, it is inherently part of the analysis, first in terms of editing, segmenting, and summarising and secondly in terms of coding, memoing (notes on codes and themes), and developing themes to conceptualising and explaining. The aim of the reduction of data was to reduce the data without substantial loss of information or context.

Although a semi-structured interview schedule was used, both an inductive and a deductive approach were followed. The interview schedule served as a guide to ensure that information relevant to certain aspects was covered, however, this did not bind the participants to adhere strictly to what was asked. Thus participants had the freedom to elaborate on issues that were not necessarily covered in the schedule but which they felt were important.

The next step in the qualitative data analysis was data display in which information was organised, compressed, and assembled by means of graphs, charts or networks and models. The final stage of the data analysis is drawing and verifying conclusions (Punch, 1995). Figure 5.2 provides a visual representation of the process followed in the research when undertaking the data analysis.

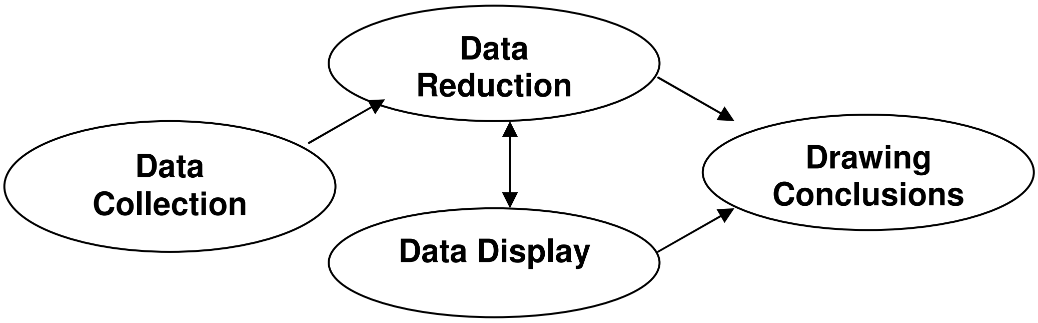

Figure 5.2 **Interactive model of the different components of qualitative data analysis (adapted from Punch, 1995)**

Thematic content analysis is an analytical method that makes use of a set of procedures to draw valid inferences from text (Weber, 1985) or to analyse the content of text where the content refers to words, meanings and themes and where text refers to anything written, visual or spoken (Neuman, 1997). In this research, thematic content analysis was chosen for the analysis of curriculum documents and interviews because it provides the tools necessary for the chunking and synthesising of data for the creation of a new whole. Through this process interviews that had been captured verbatim were coded according to different units of meaning (Henning et al., 2004). Codes are the tags or labels that refer to pieces of data. The pieces of data could be words or paragraphs. The aim of assigning these tags or labels was to attach meaning, to index the data. In the initial stages of the analysis, open coding was used for breaking up the data in order to generate theoretical possibilities within the data, some of which were targeted by the interview schedule and out of which categories and eventually themes could be developed (Punch, 1995). The following guidelines were used when coding (Berg, 1998):

1) Asking the data specific and consistent questions such as how the information is relevant to the research problem, or to what extent the data contributes to the objectives of the research. For the document analysis undertaken, the questions asked were related to the skills which should, according to the curriculum, be fostered and the grade level at which these skills are introduced. Analysis questions for the interview data were (i) what does the data mean for assessment practices, and (ii) what do the data mean for monitoring systems as well as the alignment of curriculum and assessment?
2) Thorough analysis of the data as this is the initial coding procedure. The exhaustive analysis of the data results in the saturation of the curriculum documents and interview transcriptions with repetitive codes, which allows one to move faster through

the document.

3) Frequent interruptions of the coding process in order to write theoretical notes in order to keep a record of comments and concepts that are similar, that seem to convey the same idea, and that are in line with the original purpose.

Once the open coding phase of the analysis was completed, a coding frame or scheme was developed in order to organise the data and identify findings (Berg, 1998). The coding frame defined the recoding units, which provided a framework of what aspects of the texts were classified. These aspects of text were then grouped together to form categories. These categories began to show the themes that were constructed from the data (Henning et al., 2004). The categories and themes were used to draw conclusions, which are elaborated on in Chapter 6.

5.3.5.3 Analysis of assessment data

The assessment data was analysed in terms of descriptive statistics, including item and scale analysis as well as a variety of inferential statistical procedures. Item analysis using item response theory (IRT) was undertaken which was followed by reliability analysis and correlation analysis all of which are discussed in the section to follow. The analyses described in the section to follow pertain to the first main research question ***how appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context?*** However, more specifically it pertains to ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa****?*

Descriptive statistics. Descriptive statistics were used to summarise data. Measures of central tendency such as the mean, mode and median as well as measures of dispersion, including the range of scores, minimum, maximum, standard deviation, were analysed in terms of individual items, scales of items and types of items, i.e. multiple choice and free response items (Coolican, 1999).

Item response theory. The development of item response theory (IRT) took place in the reaction to the perceived weaknesses of classical test theory (Henson, 1999). As with classical test theory, IRT examines item functioning (Crocker & Algina, 1986), but, uses probabilistic models focusing on the interplay between items and the respondents (Henson, 1999). IRT is the process, which relates certain characteristics of items (item parameters) to characteristics of individuals (termed latent traits) to identify the probability of a positive response (Hambleton, Swaminathan & Rogers, 1991). It estimates, by means of a mathematical model, how participants of different ability levels for a specific trait should respond to an item (Crocker & Algina, 1986). IRT is preferred to classical test theory because the knowledge gained by means of IRT can be used to compare performance on different tests and allows one to apply the results of an item analysis to groups with ability levels different from those of the group used for the analysis (Crocker & Algina, 1986).

Concepts central to IRT include latent traits, item characteristics curves (ICC) and the assumption of local independence. A latent trait refers to the characteristics of an individual (Hambleton et al.1991) which are unobservable and cannot be measured directly, for instance reading ability. Latent traits are referred to as abilities or theta (θ) (Baker, 2001), and can be plotted on a continuum of ability (Henson, 1999). An item characteristics curve is the visual representation of the probability of responding correctly to an item as a function of a latent trait that underlies the performance on the test (Crocker & Algina, 1986). Thus the ICC is a visual representation of the relationship between a latent trait and an item (Henson, 1999). The ICC takes the shape of a smooth "S", as shown in Figure 5.3, which indicates that the probability of a correct response is near zero at the lowest level of ability and increases to the highest level where the probability of a correct response approaches one. Thus the S-shape indicates the relationship between the probability of a correct response to an item and the ability scale (Baker, 2001). Furthermore, the S-shape goes from left to right rising continually and every person has some ability even if it is very little and no person has perfect ability (Henson, 1999). Thus the curved line approaches but never reaches zero while the upper asymptote approaches one.

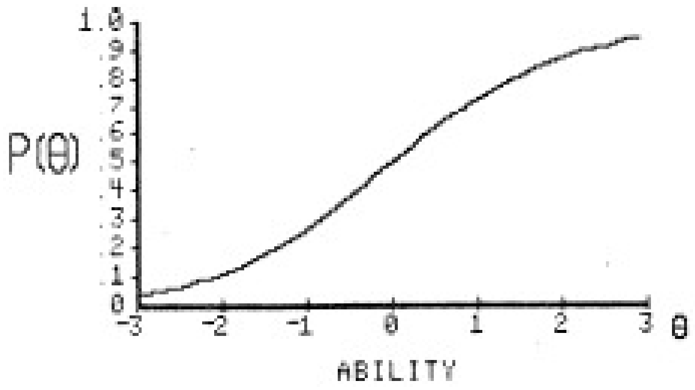

Figure 5.3 **Example of an item characteristic curve (Baker, 2001)**

The final central concept in understanding IRT is the assumption of local independence, is related to the term statistical independence, and refers to estimating response patterns by means of using the correct and incorrect responses (Crocker & Algina, 1986). The idea is that if items are to have statistical properties across samples then the items must be answered independently of one another. Items should contain no information that could be used to answer other items (Cantrell, 1997; Henson, 1999). Unidimensionality on the other hand refers to the statistical dependence of items, which can be accounted for by a single latent trait (Cantrell, 1997; Crocker & Algina, 1986) or rather that the items represent only one latent trait or dominant factor (Henson, 1999; McCamey, 2002).

Many approaches or models can be used under the umbrella term of IRT namely, one, two, three, and four-parameter models including elements of difficulty, discrimination and guessing (Crocker & Algina, 1986). In this research, Rasch modelling is used (which is a one-parameter model). The Rasch model not only contributes to inferences made about construct validity but also indicates how well the item fits within the underlying construct (Bond & Fox, 2001). Rasch was used as interval measures and are constructed by means of a stochastic process that creates inferential stability and locates a person on the latent continuum. This analysis technique is ideal for exploratory data analysis where one wants to understand the structure of items or identify items functioning well, as in the case of this research (McCamey, 2002). Furthermore, assessments based on Rasch are item and person free in that the person's response is the dependent variable while the independent variables are the person's trait score and item difficulty (McCamey, 2002). Rasch enables researchers to estimate person abilities independently of the sample used and provides statistics that indicate the precision at which abilities are estimated (Henson, 1999). Items which contribute to the sub-test are identified and poor items are eliminated. Items which are regarded as poor are items that do not contribute to the sub-test or possibly measure another construct contrary to the construct under exploration (Barnard, 2004). This is an essential first step and forms the building blocks in which the sub-tests are combined into the theoretical scales as identified by the CEM centre.

The Rasch model uses the parameter "item difficulty" (*b*). Item difficulty is defined as the position on a latent trait variable in which a person has a fifty percent probability of a correct response (McCamey, 2002). The more the participant's latent trait (θ) exceeds the item difficulty the more likely it is that a person will answer the item correctly. If the item difficulty exceeds ability (θ), according to the Rasch model, the participant will not answer the item correctly (McCamey, 2002).

For the purposes of the analysis, a dichotomous Rasch model was used where 1 denotes a correct response and 0 an incorrect responses. Missing data were kept and not recoded into incorrect as a missing value could indicate that the participant never reached the item. It is probable that by allocating an incorrect response in the model, incorrect assumptions could be made. Furthermore, retaining missing data is not problematic for the WINSTEPS program (Linacre, 2005). For the purposes of analysis any item which all participants answered incorrectly or correctly, were removed from the analysis as this does not provide additional information. After this was done, calibrations were undertaken for both persons and items in order to place both statistics on the same metric scale (Henson, 1999). The data were transformed into measures that are linear so that meaningful comparisons could be made. Logits were used to achieve approximate linearity (Cantrell, 1997).

The mean was used to centre item difficulty estimates at zero, with a standard deviation of 1. Once the item difficulties were calibrated, the initial person abilities were derived. The real person and real item separation was evaluated to the estimated standard errors of measurement that were adjusted for any misfit in the data. In addition, the real person and real item separation reliabilities were scrutinised (Smith, 2003). The separation reliabilities are similar to measures of internal consistency in that a value between 0 and 1 is obtained. The interpretation of the separation reliabilities is the same as when evaluation internal consistency reliability, in that a higher value is advantageous (Andrich, 1982).

The INFIT and OUTFIT statistics were considered. The INFIT means square (MNSQ) is associated with the response patterns, and the OUTFIT mean square (MNSQ) is associated with response patterns that are not expected. Both pick up aberrant response patterns with the former not as influenced by the outliers as the OUTFIT statistic. The question of fit is related to discrimination or how well the item discriminates between persons of high and low ability. Traditionally, high discrimination values are a “desirable characteristic” (Masters, 1988, p. 15). However in Rasch analysis, items with unusually high discriminations are eliminated from the analysis (Masters, 1988), as this over-discrimination does not provide any additional information and the fit is considered too good to be true (Andrich, 2006). Both these statistics have an expected value of 1, values lower than 1 indicate a lack of fit, while values higher than 1 indicate what is referred to as “noise” (Smith, 2003). For the purposes of this analysis values of 0.7 to 1.3 for the mean squares were considered adequate (Bond & Fox, 2001; Barnard, 2004). This is more stringent than the values of 0.5 – 1.5 recommended by Linacre (2005). The corresponding Z values were also evaluated in order to provide a complete picture. However, more weight was attached to the mean square (MNSQ) interpretation as Z-values derived from more than 300 observations tend to be very sensitive and items which should not misfit tends to misfit (Linacre, 2005). However, they are important to consider. Items with an absolute Z-value greater than 2.0 were identified. Generally, a Z-value of greater than 2.0 indicates irregular response patterns across items (lack of unidimensionality) while a Z value of less than -2.0 would indicate possible redundancy which indicates a violation of local item independence (Schumacker, 2004). Any person or item misfitting the above criteria was removed from the analysis.

The item number and the logit values were displayed on a continuum (Schumacker, 2004) in order to evaluate items and odd ratios. The odds in Rasch measurement refers to the probability of successfully answering an item correctly divided by the probability of answering the item incorrectly. The natural logarithm of the odds ratio is called natural log-odds, which in turn are referred to as logits. In terms of items, the item difficulty in logits is the natural log-odds of failure, where positive values indicate items that are more difficult and negative values indicate less difficult items. The logit for person measures, on the other hand, are the natural log-odds of success on items included in the scale or variable. A positive value here would indicate more ability on the scale, while a negative value indicates less ability on the scale. If however, both an item and a person share the same logit location on the scale, then the person has a 50% chance of answering the item correctly (Schumacker, 2004). In this research, a program called WINSTEPS was used to undertake the analysis.

WINSTEPS as a data analysis tool WINSTEPS was designed for practitioners in the field who, due to the nature of their job, have to make practical decisions while developing effective tests and assessments (WINSTEPS, n.d.). WINSTEPS constructs Rasch models by using participants' responses to a set of items; these responses could take the form of letters or integers of varying characters (Linacre, 2005). The advantage of using WINSTEPS is that, once one has familiarised oneself with the program, it is easy to use in combination with other programs such as SPSS or EXCEL. Furthermore, one is able to analyse data stemming from dichotomous, multiple-choice, rating scale or partial credit items as with other programs such as RUMM and Quest (Bond & Fox, 2001). Another advantage of WINSTEPS is that, along with other Rasch programs, it handles missing data well (Bond & Fox, 2001). It was designed specifically for the facilitation of data exploration by providing the researcher with tools to analyse items and participants in depth. The diagnostic procedures used in WINSTEPS provide information on outliers, unexpected data points, and multidimensionality (WINSTEPS, n.d.). WINSTEPS makes use of the joint maximum likelihood estimation method, which is an unconditional estimation (Schumacker, 2004) unlike RUMM that makes use of conditional estimations (Bond & Fox, 2001).

WINSTEPS analysis starts with a central estimate for each person measure, item calibration, and response structure calibration. Furthermore, the output generated by WINSTEPS consists of graphic representations in the form of graphs, plots, and tables that can be incorporated into reports (Linacre, 2005). The item and person outputs include measures, standard errors, fit statistics, reports on item/person responses that cause person/item misfit, as well as DOS files for additional analysis. In addition to complete output files of observations, residuals and their errors for additional analyses of differential item function and other residual analyses (WINSTEPS, n.d.).

Limitations of using WINSTEPS as a data analysis tool. The WINSTEPS program was designed using the Rasch measurement model as a departure point. However, the Rasch model is a one-parameter model as explained in the beginning of the section. If one wanted to include additional parameters such as discrimination (i.e. two-parameter model), another program such as PARSCALE would have to be used. Another limitation is that WINSTEPS does not permit a Bayesian maximum likelihood estimation method in order to infer logit values for individuals with extreme scores (Schumacker, 2004). WINSTEPS is however an ideal program for exploratory purposes, where the aim is to analyse items with the purpose of developing sound constructs and even though it does not permit Bayesian maximum likelihood estimation method it does provide diagnostic information on outliers.

Reliability analysis. Reliability analysis was undertaken in order to provide assertions in response to the specific research question (1.2) ***how valid and reliable, for South African schools, are the results of the instrument used in the MidYIS monitoring system on which feedback is based?*** Although the analysis is primarily undertaken to investigate the reliability component of the specific research question, inferences can be made from the content-related validity of the assessment (Suen, 1990).

Reliability addresses the extent to which the results are free from error (Gronlund, 1998). Generally, reliability refers to the consistency of scores, which are obtained by the same individuals when they are requested to complete the assessment on different occasions (Anastasi & Urbina, 1997). Furthermore, reliability is important, because unless results are stable one cannot expect the results to be valid. Additionally, consistency of results indicates smaller measurement errors - and thus is more dependable (Gronlund, 1998). This also gives an indication of how constant the scores were which were obtained in different administrations (Owen & Taljaard, 1996).

Internal consistency is a pre-requisite for construct validity, where one would expect a high item-total correlation since items measuring the same construct contributes to the total score of a test (Kline, 1993). As the assessment data was recoded into a dichotomous items, Kuder-Richardson 21 (KR-21) was used, which is a special form of Cronbach's alpha

(Coolican, 1999). KR-21 and Cronbach's alpha estimate internal consistency by determining how well all the items on the assessment relate to one another as well as to the total test (Gay & Airasian, 2003). Reliabilities for assessment data should be high, preferably around 0.9 but should never drop below 0.7 (Kline, 1993).

Apart from overall reliability coefficients, the standard error of measurement was calculated. The standard error of measurement is a useful way of expressing test reliability as it gives an indication of the amount of error allowed for interpreting individual results (Thorndike, 1997). The standard error of measurement is an index of instability of performance on the assessment using the reliability to estimate how much an individual score might change from one testing to another (Thorndike, 1997).

Internal consistency reliability is not recommended in assessments where speed is a factor, as the results may be artificially inflated (Frisbie, 1988). The decision to make use of internal consistency in this research instead of using another method to estimate the reliability was based on the fact that it is the method preferred by CEM and the assessment is not a pure speeded test but rather a combination of a speed test and a power test. Time allocations were adjusted so that the majority of learners would be able to attempt the majority of the items if not all the items.

Correlation analysis. The aim of the correlation analysis was to establish whether a relationship existed between the ability assessment and academic achievement in specifically language and mathematics, i.e. whether the ability assessment could be used to predict future achievement. The analysis pertained to the specific research question (1.2) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa.*** More specifically the sub-research question is focused on, namely ***to what extent does the data predict future performance?*** Thus predictive validity was investigated by means of correlation analysis where a correlation (r) is the measure of association between two variables (Blaikie, 2003). As the variables used in this analysis are continuous in nature, Pearson product moment correlation was used, which measures the extent to which a variable covaries with another or rather the degree to which variables are related (Yaffee, 2003).

The correlation coefficient has direction, as it can be positive or negative, has magnitude as it can be large or small and is interpreted in terms of statistical significance (Yaffee, 2003). For this research, a positive correlation of 0.3 or 0.4 was considered sufficient for the ability-academic relationship (Kline, 1993). While the language and mathematics achievement obtained from the participating schools are used it is acknowledged that the correlation analysis is an indicator of possible predictive validity. However, will only be relevant to the school and class used in the analysis due to possible variations in composition of marks, different educators and assessments were involved.

Generally, large correlation coefficients are obtained when the characteristics of the assessments correlated are alike, the spread of scores is large, and the scores are stable (Linn & Gronlund, 2000). A 1.00 refers to a perfect correlation. However, in practice, this seldom happens.

5.3.5.4 Analysis of contextual data

The analysis described in the section to follow pertains to the first main research question ***how appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context,*** with specific reference to reliability issues. However, the analysis also forms the basis for the second main research question of ***which factors could have an effect on learner performance and therefore inform the design of the monitoring system****.* The questionnaire data was analysed in terms of descriptive statistics as well as a variety of inferential statistical procedures. Reliability analysis was undertaken. The procedures are briefly discussed below.

Descriptive statistics. Measures of central tendency such as the mean, mode, and median as well as measure of dispersion, including the range of scores, minimum, maximum, standard deviation, were analysed in terms of individual items as well as for scales of items (Coolican, 1999). The object of this exercise is to summarise and describe the data in order to make the reporting process easier.

Reliability analysis. Reliability analysis allows the researcher to study the properties of measurement scales in terms of relationships between individual items and the scale as a whole (SPSS, 2001) and individual items giving an indication of the stability over time and internal consistency of items. Problematic items can easily be identified. These items were omitted not only for the analysis but also from future versions of the questionnaires. Internal consistency is a pre-requisite for construct validity, where one would expect a high item-total correlation since items measuring the same construct contribute to the total score of a test (Kline, 1993). Internal consistency is measured by Cronbach's coefficient alpha when the score is not dichotomous and it reflects how well the different items complement each other in their measurement of different aspects of the same variable (Litwin, 1995).

The closer the alpha is to one, the greater the internal consistency of the items in the questionnaire being assessed (George & Mallery, 2001). According to Kline (1993), reliabilities should ideally be high for assessments, around 0.9 but should never drop below 0.7 (Kline, 2003). However, for questionnaire data DeVillis (1991) states that, the minimally acceptable reliability is between 0.65 and 0.7, although a coefficient as low as 0.5 was acceptable for exploration of the data (Howie, 2002).

5.3.5.5 Building an exploratory model using assessment and contextual data

The sections to follow pertain to the second main research question guiding this study namely ***which factors could have an effect on learner performance and therefore inform the design of the monitoring system***. Often in research processes at a higher level of analysis influence processes at a lower-level (Luke, 2004). Multilevel analysis is used for the analysis of complex data (Snijders & Bosker, 1999) such as data with hierarchical or nested structures (learners in classes, classes in schools for example) with the aim of explaining variability in a dependent variable(s) through a set of independent variables (Heck & Thomas, 2000). In the words of Snijders and Bosker (1993, p. 237):

> Hierarchical linear models incorporating both random and fixed effects provide a useful statistical paradigm for situations where nesting is an obvious and direct consequence of multistage sampling as well as situations with nested sources of variability.

Important concepts in multilevel analysis include intra-class correlations, random and fixed coefficients. The first concept intra-class correlation refers to the degree of dependence of individuals. At the core of the concept is the view that the more individuals share common experiences the more similar they are or rather the more homogeneous the groups are (Kreft & de Leeuw, 1998). The intra-class correlation is equal to the estimated proportion of group level variance as compared to total variance (Hox, 1995). Random and fixed coefficients refer to different sections or parameters of the multilevel equation. Random coefficients are values assumed to be distributed as a probability function or the residual error terms. Fixed coefficients are the intercept and slope, which are estimated from data. Other important concepts include cross-level interactions, and estimation methods. In multilevel modelling, cross-level interactions refer to interactions between variables at different levels of the data structure (Kreft & de Leeuw, 1998). Finally, estimation methods refer to the techniques used to estimate parameters.

When investigating the effect of factors on performance, literature suggests that some form of regression analysis be undertaken (Newton & Rudestam, 1999). In school effectiveness research, multilevel analysis is undertaken to investigate which factors are associated with performance of learners (Riddell, 1997; Sammons, 1999; Scheerens, 1990, 2001a; Scheerens & Bosker, 1997). Multilevel analysis is specifically suitable for this purpose because of the nested structure of data collected in education where learners are nested in classes that are nested in schools, as described in Chapter 3. When identifying factors that influence performance, the nested structure that is present cannot be ignored because inferences may be skewed. Multilevel analysis was therefore used in order to ascertain which factors, identified from literature and correlating with achievement, affect learner performance on the assessment. Multilevel analysis allows for the identification of factors on the different levels and makes it possible to determine to what extent these factors affected the outcome of the assessment (Hox, 1995). As the research is exploratory in nature, the ideal would be to explore direct, indirect and interaction effects where possible. A prerequisite is to have information on three levels at least (Scheerens, 1997). Thus the research attempted to build a model using three levels, namely school, classroom, and learner-level.

Sampling considerations. Generally when undertaking multilevel analysis, the larger the sample the smaller the effect size and greater the power of the analysis (Snijders & Bosker, 1999). Literature on multilevel analysis has tended to focus on estimation and interpretation and not necessarily on sample design questions (Cohen, 1998). As the use of multilevel analysis increases, however, and the approach is used in a variety of contexts, issues of sampling do become important. Cohen (1998) is of the opinion that where it is important to estimate the variance components one must sample more learners and fewer schools. Furthermore, Maas and Hox (2004) state that if one is only interested in the fixed effects of the model as few as ten groups can lead to good estimates. According to Snijders and Bosker (1993) the optimality of sample sizes in a design means that minimal standard errors for the parameters of interest are calculated and if the sample size for either the macro or micro level is smaller than 10, the resulting standard errors should not be trusted. Snijders and Bosker suggest that the sample should be greater than 10. Conversely, Maas and Hox (2002) state that ten groups are too few, but if one is only interested in fixed regression coefficients, as few as ten seem reasonable. But they advise one to use bootstrapping or a simulation-based method to assess sampling variability (Maas & Hox, 2004). Mok (1995) is of the opinion that more schools and fewer learners per school are required in order to minimize bias. From the discussion above it is apparent that sample size has to be evaluated not only in terms of the power of the statistical test but also in terms of the effect that the variance component of the model has on the estimates (Hox, 1998) The sample sizes of the three levels were therefore evaluated (10 on the school-level, 36 on the classroom-level and 794 on the learner-level).

Approach to model building. This research is exploratory in nature and therefore a specific model was not specified in the beginning. Rather, the analysis started with the most basic model, which is the intercept-only model, and parameters were systematically included. The most basic model is specified by:

$Y_{ijk} = \gamma 000 + v_{0k} + u_{0jk} + e_{ijk}$

Where (Hox, 2002):

Y_{ijk} = dependent variable, in this case the results on the assessment

$\gamma 00$ = intercept or regression coefficients

v_{0k} = residual error at the highest level

u_{0jk} = residual error at the second-level

e_{ijk} = residual error at the lower-level

This model does not include any explanatory variables. It estimates the intra-class correlation and provides a measure of the degree of misfit in the model (Hox, 1995; 2002). The model is built by adding the first-level or lower-level explanatory variables so that the contribution of each explanatory variable could be assessed. The lower-level explanatory variables are the learner-level variables of age, gender*,* learner attitudes (in terms of school, language and mathematics classes) and motivation to achieve (language and mathematics classes, pressure from parents and peers, future aspirations). This was followed by the analysis of the slope of the explanatory variables in order to evaluate whether there was a significant variance component between the groups.

The second-level explanatory variables were added, making it possible to examine whether these variables explained between-group-variation in the dependent variable (Hox, 1995; 2002; Luke, 2004). The second-level explanatory variables refer to educator variables such as gender, age, experience, qualifications, educator attitude towards achievement, quality of instruction, instructional method, and opportunities to learn. The third-level explanatory variables were added, making it possible to examine whether these variables explained between-group-variation in the dependent variable (Hox, 1995; 2002; Luke, 2004). The third-level of explanatory variables included type of school and school attitude to achievement.

Finally, the cross-level interactions between explanatory group level variables and the individual level explanatory variables were added (Hox, 1995; 2002; Luke, 2004). After each step, additional parameters were added. The results were inspected to see whether the parameters were significant and to examine the residual error left (Hox, 1995) until a final model was constructed including cross-level interactions, explanatory group level variables and individual level explanatory variables as given by:

$$Y_{ijk} = \gamma_{000} + \gamma_{p0}X_{pijk} + \gamma_{0q}W_{qjk} + \gamma_{0r}Z_{rk} + \gamma_{pqr}Z_{rk}W_{qjk}X_{pijk} + u_{pjk}X_{pijk} + v_{0k} + u_{0jk} + e_{ijk}$$

Where:
$\gamma_{p0}X_{pijk}$ = lower explanatory variables
Subscript p = explanatory variables at the lower-level
$u_{pjk}X_{pijk}$ = variance of slopes of the explanatory variables
$\gamma_{0q}W_{qjk}$ = second-level explanatory variables
Subscript q = explanatory variables at the second-level
$\gamma_{0r}Z_{rk}$ = third-level explanatory variables
Subscript r = explanatory variables at the third-level
$\gamma_{pqr}W_{rk}Z_{qjk}X_{pijk}$ = cross-level interaction term

To conclude, in each of the steps indicated above a decision was taken on which regression coefficients to include based on significance tests, the change in deviance and change in variance components (Hox, 1995).

MLwiN as data analysis tool. *MLwiN* version 2.02 was used in this analysis as opposed to HLM (Hierarchical Linear Models). The Centre for Multilevel Modelling, in the United Kingdom at the University of Bristol, developed *MLwiN*. HLM was developed by Scientific Software International (SSI) based in Lincolnwood, Illinois in the United States of America. Features of HLM are useful. Easy-to-use features include multinomial and ordinal models for two-level data, multivariate models for incomplete data, latent variable analysis, and log-linear model for heterogeneous level 1 variance. However, HLM has some restrictions with regards to data preparation in the database files, which are not present in *MLwiN*. HLM also does not have any facility to carry out data manipulation. Users have to resort to other software to prepare data for the format required by HLM and then import it into the program. Data input is therefore difficult (Yang, 2006).

MLwiN provides researchers with a system that meets the demands of specification and analysis of multilevel models because of its following characteristics (Hox, 1995; (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Lanford & Lewis, 2001) :

- The program has a graphic user interface for specifying and fitting models. Additional features include plotting, diagnostic and data manipulation facilities, and a user-friendly help system.
- It includes a spreadsheet with columns denoting variables, frequency data, or parameter estimates while the rows denoting the lowest level units in the hierarchy used to structure data.
- The software allows researchers to analyse data with arbitrary levels and to use FLM estimation as well as RLM estimation.
- *MLwiN* allows all regression coefficients to be random at all levels. It is therefore able to analyse non-standard as well as standard multilevel models.
- *MLwiN* allows for the repetition of computations and for the use of residuals from one analysis as an input in another model.
- It is very interactive in nature and provides researchers with control over computations, making it easy to try out different sub-sets of variables and error structures.

However, as with many programs used for statistical analysis, *MLwiN* does not handle missing data well. Before importing data into *MLwiN* missing values were replaced with plausible alternatives such as the mean, median or mode depending on the type of variable. If the number of missing values was more than 5%, dummy variables were used (Luyten, personal communication, November, 2005).

5.3.6 Research procedures

The initial design was refined in terms of the resources such as time, personnel, and finances. This took place in 2003 and 2004. Initial adaptations of the assessments started in 2004. This comprised format changes, as the assessment used by CEM is electronically marked and therefore set out in order to facilitate this process. The sample was drawn in late 2003. It consisted of several schools in the Pretoria area. The Gauteng provincial government was contacted in 2004 to obtain permission to undertake the study and permission was granted. Three schools were contacted in early 2004 to participate in a pre-pilot. Principals were contacted telephonically to describe the project. In addition consent forms were sent to principals and the parents of the learners participating in the study. National Department of Education officials as well as Provincial Department of Education officials were contacted telephonically and follow-up emails were sent. The pre-pilot took place in Mid-2004. The process of instrument development and adaptation took place in early 2005, which included sending the instruments to specialist and adapting the instruments accordingly. Principal and educator questionnaires were developed from a review of literature as well as already developed questionnaires towards the end of 2004 while the interview schedule for the national department officials and questionnaire for provincial department officials were developed in early 2005. Data collection took place from May 2005 to July 2005. The National Department of Education interviews took place during May 2005, while the Provincial Department of Education questionnaires were emailed in May 2005. The follow-up interview took place during June 2005. School-level data collection took place during May 2005 and June 2005. The instruments were coded and captured during June 2005 and July 2005. Finally, the results were written up and reports compiled for schools during August 2005 (refer to Appendix H for a diagrammatic representation of the procedures followed).

5.4 Ethical issues

Two aspects, regarding ethical issues are discussed for this research. Firstly, the ethics requirements as prescribed by the Faculty of Education of the University of Pretoria and secondly the researcher's integrity. Before the research could take place, permission had to be sought from the Faculty of Education in terms of the ethical considerations of the research project. The Faculty was satisfied with the procedures suggested and granted permission to continue with the research.

In terms of professional integrity, it was important to be transparent about the research. The Gauteng Department of Education as well as schools were contacted in order to obtain informed consent (refer to Appendix I for the letters to the participants). In addition, parents of every learner were sent a letter explaining the project and asking them to grant permission for their child to participate. The research was also explained to the learners who were selected to participate. The project was placed in context, namely that the project was undertaken in collaboration with the CEM centre at the University of Durham. Furthermore, schools and educators were approached and asked if they would be willing to participate by completing a background questionnaire. Confidentiality was promised to parents, learners, educators and schools as well as National and Provincial Department of Education officials. Furthermore, participants could withdraw at any time.

5.5 Methodological constraints

In this chapter, the research design and methodology was elaborated on in detail. The research is situated within a pragmatic paradigm and makes use of mixed methods. Several constraints were observed during the course of the research:

- The sample sizes on classroom-level as well as school-level were rather small. This resulted in a situation where certain analyses, for instance reliability analyses, could not be performed due to minimum sampling requirements that were not met. Small sample sizes also influenced the multilevel analyses (refer to sampling considerations under 5.3.3.5).
- The sample included on urban and peri-urban schools from one province. Thus schools from rural areas and schools from other provinces were not included.
- Schools from only one province were included for study. Although steps were taken to ensure a maximum variation sample, the results still reflect an urban/peri-urban setting and therefore do not transfer to the more rural areas. This is seen as a constraint as a large percentage of South African schools are situated in rural areas. As this research was exploratory in nature this constraint may however be an artificial one seeing that the aim was to investigate the feasibility of using MidYIS in the South African context.
- The measure used for academic achievement was provided by the schools. Therefore the inferences made in terms of predictive validity can only be investigated per class or teacher and are relevant only to specific schools and specific educators, as was discussed earlier in this chapter.

5.6 Conclusion

In this chapter, the information pertaining to the research design and methodology followed in this research was detailed. The intention of the current research was to investigate the feasibility of adapting an already existing and well-functioning monitoring system to the South African context. In order to accomplish the aims of the research, the research was grounded in the pragmatic paradigm. The use of pragmatism as a grounding paradigm was explored and justified. Pragmatism makes use of qualitative and quantitative methods of inquiry, which is called a mixed method approach, where the research question drives decisions about which method to use. A case was made for the appropriateness of mixed methods as well as a concurrent nested strategy approach, where such an approach implies that there is a dominant method that guides the research. The dominant method of this research is quantitative while the qualitative component is given a lesser priority though nested within the quantitative approach. The qualitative approach can be said to be nested or embedded because it addresses a different question to the quantitative approach while also seeking information from a different level.

The reader was also provided with a description of the sample included in the research, which comprised National Department of Education officials, Provincial Department of Education officials, specialists in key areas, schools, principals, educators and learners. Several instruments were used to collect data, ranging from semi-structured interview schedules to assessments and questionnaires. As with any instrument used for research purposes, issues of validity are of utmost importance, even more so in this research because the feasibility of adapting an already existing monitoring system was being explored. Validity issues pertaining to content-related validity, curricular validity, construct-related validity and criterion-related validity were discussed. Also described were the validation strategies employed for both the qualitative and quantitative component, including expert opinion, test-curriculum overlap, and triangulation.

Various data collection strategies as well as data analysis strategies were discussed. The data analyses included thematic content analysis, descriptive statistics, item response theory, reliability analysis, correlation analysis, and multilevel analysis. These discussions included an elaboration of what the analysis entails, steps undertaken and statistical considerations that were taken into account. The results of which can be found in Chapters 6, 7 and 8. Lastly, the chapter also included a discussion on the ethical considerations as well as the methodological constraints.