CHAPTER 6

THE CONTENT–RELATED VALIDITY OF THE MIDYIS ASSESSMENT

The main aim of the current research is to investigate the feasibility of implementing the MidYIS monitoring system developed in the United Kingdom in the context of South Africa. The discussion in this chapter relates to the specific research question of ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** *As discussed in Chapters 3, 4 and 5, validity is a unitary concept but comprises several facets namely content-related validity, predictive validity and construct-related validity. The present chapter describes the outcome of the content-related investigation of the assessment instrument, not only in terms of the South African curriculum but in the field of abilities assessments as well.*

6.1 Introduction

This chapter represents the first of the results chapters and elaborates on the outcome of validation strategies relating to the first main research question ***how appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context?*** More specifically the chapter addresses the specific research question 1.2 (as described in Chapter 5) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** The focus of this chapter is on the validity of MidYIS in terms of content-related validity while the findings for construct validity, predictive validity, and reliability analysis are presented in Chapter 7.

In Chapters 3, 4 and 5 issues pertaining to the first main research question and the specific research questions were discussed in terms of the criteria for evaluating the quality of measurements and how the research project is designed in order to make inferences related

to the quality of measurements. The figure presented in Chapter 3 (Figure 3.3) can be adapted to reflect the key issues addressed in this research (refer to Figure 6.1).

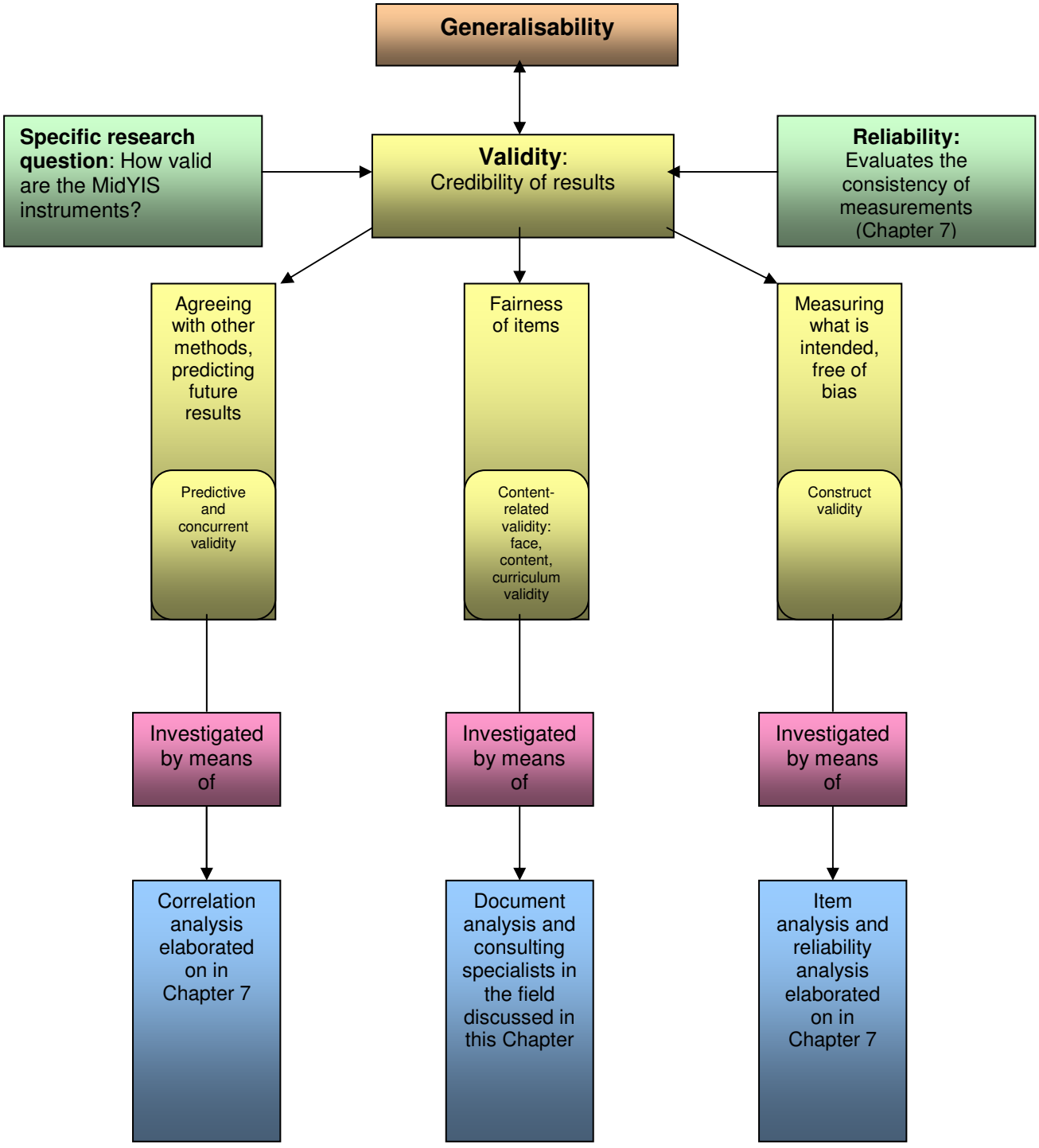


Figure 6.1 **Extension of the criteria for evaluating quality of measurement used in monitoring systems (adapted from Fitz-Gibbon, 1996)**

More specifically this chapter addresses two sub-questions related to ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa*** namely:

1.2.2 To what extent are the skills tested by MidYIS valid for the South African curriculum?

1.2.3 To what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa*?*

The discussion in this chapter also relates to the third specific research question (question 1.3 as described in Chapter 5) which is ***what adaptations are needed to transform MidYIS into SASSIS, a monitoring system for the South African context?*** The proposed adaptations, which were discussed in Chapter 4 (Section 4.6) and Chapter 5, present the first step in the transformation from the Middle Years Information System (MidYIS) to the South African Secondary School Information System (SASSIS). The adaptations addressed in this chapter are included in the sub-questions identified:

1.3.1 To what extent are the administration procedures appropriate and if not, how can these be adjusted?

1.3.2 To what extent is the content in MidYIS appropriate for second language learners?

1.3.3 To what extent is the format of the assessment appropriate and if not, how can it be changed?

1.3.4 To what extent are the time allocations appropriate and if not, what adjustments are needed?

The second section (6.2) of the Chapter addresses the sub-question (1.2.2) ***to what extent are the skills tested by MidYIS valid for the South African curriculum***, which is related to the curriculum validity of the assessment. This section is divided into four sub-sections. Background information drawing from the interviews undertaken with the National Department of Education officials and questionnaires completed by the Provincial Department of Education officials is provided in 6.2.1. The language learning area drawing on the curriculum documents and evaluations from language specialists is elaborated on in 6.2.2. The mathematics learning area is described in 6.2.3 drawing from information in the curriculum documents and the mathematics specialist. ***What adaptations are needed to transform MidYIS into SASSIS, a monitoring system for the South African context*** is explored in 6.2.4, by means of integrating the findings from 6.2.1, 6.2.2, and 6.2.3. Content-validity from a psychometric perspective is explored in 6.3, and addresses the sub-question (1.2.3) ***to what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa****.* The chapter concludes with Section 6.4 in which inferences are drawn based on the three sub-questions concerning the content-related validity of the MidYIS assessment addressed in this chapter.

6.2 Exploring the curriculum validity of the MidYIS assessment

In Chapter 5 issues relating to the validity of the MidYIS assessment were elaborated upon. Content-related validity was described from two perspectives namely from a curriculum and a psychological perspective. Section 6.2 elaborates on the curriculum perspective from an intended curriculum perspective (see Chapter 5).

> Policy is neither static nor does it occur in a vacuum. Instead, it is constantly subjected to various influences that impact upon it…As policy evolves towards practical application, distortions and obstacles to its successful execution become apparent (Mahomed, 2001, p. 105).

South Africa has undergone extensive policy changes in education since 1994. The issue of policy in terms of monitoring education and the curriculum has at times been difficult to navigate as was described at the end of Chapter 1. However, the aims and objectives of the curriculum as set out in the curriculum policy documents do have an inherent logic. For example, the aim of the South African intended curriculum, i.e. the vision or philosophy underlying the curriculum as expressed in curriculum policy documents (Travers & Westbury, 1989; Van den Akker, 2003), is to provide learners with generic skills and knowledge which can be applied to different contexts (Gultig, 2003). The need for a combination of skills and knowledge can be ascribed to the ever changing world of work where "greater skills are required" (Mohamed, 2001, p. 125) as a result of technological advances and globalisation (Kraak, 1998). Essentially the concepts of skills and curriculum are important. The connection between skills and curriculum is related to the sub-question (1.2.2) ***to what extent are the skills tested by MidYIS valid for the South African curriculum***.

Mahomed (2001, p. 133) states that the government adopted an outcomes-based education system because of its promise to "integrate content, skills and outcomes", however, he goes on to say that a "major cause of poor quality education in South Africa can be attributed to the pedagogical approach of education institutions especially curricular content and processes". The aim of this section is to provide an analysis of the extent to which the generic or basic skills tested in the MidYIS assessment are present in the curriculum. However, before presenting the results of the language and mathematics curriculum document analyses and evaluations from experts in the language and mathematics learning areas, background information from the National and Provincial Departments of Education is provided.

6.2.1 Perspectives from National and Provincial Departments of Education

The aim of the interviews with the National Department of Education was to elicit views pertaining to curriculum, assessment, and monitoring of learning. Although the current research is on a small scale, the ultimate aim is to have a monitoring system that could be implemented nationally. As a result, it was important to understand what would be acceptable for and what would be endorsed by the National Department of Education. For example would the Department promote a Tylerian approach in which the focus is on defined outcomes (du Toit & du Toit, 2003) and in which the quality of the curriculum is monitored by means of collecting data relating specifically to the outcomes (Burks, 1998). Alternatively, would the Department be in favour of a more holistic approach as advocated by Stake (1967) in which background information of learners, educators and schools, interactions between the school and community, educator and learner as well as outcomes are considered (Stufflebeam & Shinkfield, 1984).

Two interviews were undertaken with officials in key positions in the Department, a Chief Director, and Director. Although the sentiments expressed during the interviews were very similar in nature, one of the interviewees was particularly articulate; as a result many of the quotations included in the discussion to follow are taken from that interview (refer to Appendix E for the audit trail documents).

From the interviews (see below) emerge a suggestion that a more holistic approach to monitoring would be preferred. This is perhaps not surprising as the Systemic Evaluation Framework (National Department of Education, 2003a) draws heavily on an input-process-outcome model in which background information on learner, educator, and school-level is a key aspect. Furthermore, the *Whole School Evaluation* model implemented by the Department of Education is meant to be comprehensive by including information collected from all levels within the school, from management to classroom observations and learner performance. As illustrated below, the issue of quality is of importance and learner performance may be used as an indicator to determine the quality of education:

> …learner performance … can [be] used as indicator for quality or determining quality of the system (Interviewee 2, personal communication, June, 2005).

> I think we need to move beyond assessment you know especially learner assessment as the only tool of monitoring performance…[rather] a system that will operate at all levels of education, all levels starting from the classroom (Interviewee 2, personal communication, June, 2005).

What emerges from the interview is the idea that whichever monitoring system is used, it needs to be multilayered, and able to provide information at a number of levels namely school, district, province, and national. With this in mind perhaps a similar model to those used in the Quality Learning Project (QLP), the Integrated Education Project and the Khanyisa Education Support Programme could be adapted, a system including both the district and provincial level. The school improvement models that are used in these projects adopt a more systemic approach (Taylor & Prinsloo, 2005, p. 7):

> …schools and teachers respond best when support is accompanied by accountability demands, and that capacity therefore needs to be built at district, school and classroom levels so as to strengthen systems for both monitoring and supporting learning.

Another important consideration is the types of schools across the country. The monitoring system would have to be valid for the variety and diversity of school contexts. In Chapter 1 background information on education in South Africa was given. In one province there could be schools ranging from those with adequate facilities, trained teaching staff, and efficient management to those with less than adequate facilities, teaching staff who are barely qualified and no management to speak of. Thus a monitoring system would have to be applicable to the whole spectrum of contexts. This need was expressed as follows:

> A system that needs to talk to different contexts in our country (Interviewee 2, personal communication, June, 2005).

The implications for the current research are that if the MidYIS is to be accepted on a national-level, it should include in its framework a number of levels, namely classroom, school and provincial-levels, and be appropriate in a variety of contexts, taking into account the diversity of resources and people. Thus an approach in line with that of Stake (1967), mentioned earlier in the section, may be more appropriate in that background information on learner, classroom, school and provincial-level must be considered in conjunction with outcome or performance of learners. Apart from the monitoring system having to be

applicable to different contexts, it should also make use of assessment practices that are in line with the continuous assessment model advocated by the Department of Education.

> Different ways of collecting evidence are encouraged and assessment, which is linked to outcomes within the curriculum (Interviewee 1, personal communication, June, 2005).

> Assessment should be used formatively.... If you even ask the teacher, what do you do with the results of the assessment? Nothing I just record them and that's it (Interviewee 1, personal communication, June, 2005).

What was important for the research was the reiteration of the importance of skills in conjunction with knowledge or, as referred to here, as content. This emphasis is perhaps not surprising as it is rooted in the philosophy underpinning education documents, namely a competency-based approach to education (Kraak, 1998).

> I think there needs to be a relationship between what is taught and what is assessed. But this relationship goes beyond the content. It has to also focus on …skills … content … the two definitely need to go together (Interviewee 2, personal communication, June, 2005)

> You need to understand that there is a relationship between the teaching process and the assessment process (Interviewee 1, personal communication, June, 2005).

Judging from the interviews as illustrated in the quotations above, there is the tendency by the interviewees to delineate between what is taught and what is assessed. It appears that even though the interviewees refer to the relationship between assessment and the curriculum, they tend to separate the two without reflecting upon how assessment is embedded in the curriculum. In terms of conceptualisation, curriculum and assessment have traditionally been kept separate, but increasingly there is a specific focus on assessment of learning to assessment for learning (Gardner, 2006). In addition, the skills component of the curriculum is also kept as a separate issue, not embedded in the curriculum, but rather something additional to the curriculum. Kraak (1998) has suggested from a competency-based perspective that the integration of curriculum and skills is essential.

The assessment used must be aligned with the curriculum. This presented some challenges for this PhD research. Firstly, because the assessment used in the monitoring system being explored in this study was not designed as a curriculum-based measurement but rather as a developed abilities assessment. Secondly because the extent in which the skills being tested, although present in the curriculum (see 6.2.2. and 6.2.3), has to be ascertained. However, in any assessment used in a research project the challenge is always to provide for sufficient curriculum coverage while also considering practicalities such as time and length of the assessment. The discussion on MidYIS and curriculum overlap is elaborated on further in 6.2.2 and 6.2.3.

While it would appear that the National Department officials might accept the use of already developed assessments, as long as the assessment is clearly aligned with the curriculum, the three Provincial Education officials who completed the questionnaire (see Chapter 5) were not in favour of using already developed assessments or assessments that were not developed by the educator him/herself. The official who works for the Gauteng Department of Education Office for Standard in Education (OFSTED) was contacted telephonically to clarify some issues that emerged from the questionnaire he completed. When asked why he was not in favour of developed assessment, the respondent indicated that continuous assessment practices are new. Furthermore, the respondent revealed that he had not seen assessments that were closely related to the curriculum, as the curriculum was open to interpretation and customisation by schools. If however, the school had a programme or curriculum in place and the assessment was related to the programme he felt that then it might work. The OFSTED official also indicated that it might be good to have a standardised assessment in place, as some schools might want to have a benchmark from which to evaluate their performance, specifically against similar schools, as well as against international standards.

The statements from the OFSTED official reinforce the idea, which emerged from the interviews with the National Department officials, that the assessment should be curriculum-based. Therefore, if a monitoring system is to be acceptable to government, the tools used in the monitoring system should be valid for the school's curriculum and if learner progress is to be followed, the assessment should take place at intervals. This could be related to the curriculum-based measurement.

Curriculum-based measurement is a standardised measurement system in which key areas of the curriculum are identified and monitored in order to ascertain whether learners have reached a level of mastery in relation to the identified level within the curriculum (Fuchs &

Fuchs, 1991). Curriculum-based measurement systems are primarily used in special needs education but also used in mainstream education where basic skill areas such as vocabulary, reading, and mathematics are the focus (Espin, Shin & Busch, 2005). The point that Espin et al. (2005, p. 353) make is that "one of the most difficult components of education is the measurement of change. By measuring change in performance, teachers can reliably evaluate student learning and the effects of instructional interventions on that learning". Change in the context of the quotation refers to progress being made based on assessment results before and after interventions in instructions. The point here is that if the MidYIS monitoring system is to be used by schools to measure change, then the assessment should provide guidance as to what instructional interventions are needed. However, the assessment cannot provide the necessary guidance if the skills assessed cannot be linked to the curriculum, which is taught.

In an attempt to ascertain whether the skills assessed in the MidYIS (refer to column 1 of Table 6.1) are present in the intended curriculum, the Provincial Department of Education officials were asked to indicate whether the skills were indeed present. The results are depicted in Table 6.1. The list of skills was compiled based on the skills that are assessed in the MidYIS instrument (while the question of whether the skills mentioned in the curriculum are sufficiently covered by MidYIS is addressed in 6.2.2 and 6.2.3). The results indicate that the skills present in the instrument were present in the curriculum and that many of the skills were introduced to learners during primary school and therefore could be considered basic skills underpinning the secondary school curricula such as number sense in mathematics.

Table 6.1 ***Skills as indicated by the Provincial Department Education officials***

Skill assessed in the MidYIS assessment	Skills taught in Primary School	Skills taught in Grade 8	Skills taught in Grade 9
Recognising words	X		
Measurement	X		
Identifying synonyms	X		
Numbers, Operations and Relationships	X		
Proof reading	X	X	
Spotting mistakes quickly	X	X	
Identifying differences in information when comparisons are made	X	X	
2D and 3D ability	X		
Spatial ability	X		
Pattern Recognition	X		
Sequence Recognition	X		
Logical thinking	X		
Reasoning	X	X	
Critical thinking	X	X	
Skimming	X	X	
Scanning	X		X
Problem solving	X		

The clear message from both the National and Provincial Departments of Education is that monitoring is desirable but that the measure used in monitoring should be aligned with the curriculum. As can be seen in the table above (Table 6.1) the fundamental skills assessed in MidYIS seem to be present in the primary school curriculum and should be established on entry to secondary school. This provides some legitimacy and motivation for the investigation of curriculum aspects (whether the skills assessed in MidYIS are in the curriculum and whether MidYIS adequately covers the skills included in the curriculum), in addition to the traditional psychometric properties of the assessment. To enhance the discussion on the link between the intended curriculum and the MidYIS assessment, it was deemed appropriate to scrutinise the curriculum documents. The discussion that follows details the analysis of the South African curriculum documents. The aim of the sections to follow is to provide insight into the issue of curriculum validity of MidYIS. What is of importance, therefore, is the extent

to which the skills in the MidYIS assessment are taught in the language and mathematics learning areas. This will be addressed in the sections to follow.

6.2.2 The language learning area

There are six learning outcomes for the languages learning area, as presented in Table 6.2. The South African curriculum works on the principle of scaffolding where basic information is taught and learnt at the lower-levels while the sophistication of knowledge to be mastered increases with every grade.

Table 6.2 ***Outcomes in the languages learning area***

Learning outcome		Aim of the outcome
Learning outcome 1	Listening	To enable the learner to listen for information and enjoyment, and respond appropriately and critically in a wide range of situations.
Learning outcome 2	Speaking	To enable the learner to communicate confidently and effectively in a spoken language in a wide range of situations
Learning outcome 3	Reading and viewing	To enable learners to read and view information and respond critically to the aesthetic, cultural and emotional values in texts
Learning outcome 4	Writing	To enable the learners to write different kinds of factual and imaginative texts for a wide range of purposes
Learning outcome 5	Thinking and reasoning	To enable the learner to use language to think and reason, and access, process and use information for learning. Due to the nature of learning outcome 5, it does not form part of the additional languages curriculum
Learning outcome 6	Language structure and use	To enable learners to know and use the sounds, words and the grammar of a language to create and interpret texts

(Source: National Department of Education, 2002b)

Learning outcomes 1, 3, 4, 5 and 6 will be discussed briefly in the paragraphs to follow (for detailed discussion readers are referred to NRF Value-Added Technical Report, 2005). The outcomes are discussed with the intent to relate them to MidYIS as the learners are expected to listen, read, think, reason and know the structure of language. However, MidYIS does not

assess learners' ability to speak the language, in this case English, as described in learning outcome 2 and for this reason outcome 2 is not elaborated on.

Being able to ***listen*** and understand what is being said is an important skill that is used throughout life. In ***learning outcome 1 (listening)*** listening skills are focused on. Listening entails being attentively and actively paying attention to instructions, announcements, and being able to respond appropriately by means of carrying out instructions and follow directions. Learning outcome 1 also focuses on the development of phonic awareness so that the learner can distinguish between different phonemes, especially at the beginning of words (National Department of Education, 2002b, 2002c).

Learning outcome 3 (***reading and viewing***) can be broken down into certain skills namely ***viewing, reading, skimming, and scanning***. According to policy, ***viewing*** entails using visual cues to deduce meaning, in that the learner should be able to look at pictures and be able to recognise common objects and experiences. The learner should also be able to identify a picture or figure from the background, make sense of picture stories, match pictures and words (National Department of Education, 2002b, 2002c)

Reading on the other hand entails reading for meaning. The aim is to cultivate techniques and strategies that would help learners to read for meaning. Reading, in the policy documents, is seen as an essential element in the development of language, learning to write and learning about the world. Reading entails the ability to distinguish pictures from print and recognise the meaning being conveyed. The meaning then links up with learner experiences and the learner is enabled to describe and give opinions of characters in stories or television programmes (National Department of Education, 2002b; National Department of Education, 2002c). The aim of reading is to enable learners to read spontaneously and often, for pleasure and information, across a range of text types, to describe personal response and discuss the kinds of texts enjoyed and finally to use appropriate reading strategies such as skimming and scanning. ***Skimming*** according to policy entails glancing over texts in order to obtain a sense of the general ideas being conveyed. ***Scanning*** entails looking or searching for specific details (National Department of Education, 2002b, 2002c).

Learning outcome 4 (***writing***) can also be divided into a number of skills. Language does not only consist of spoken words but also of written words. The aim of learning outcome 4 is to develop ***writing skills*** that enables learners to write in such a way that others can understand. This entails enabling the learner to use appropriate ***grammatical structures*** and ***writing conventions*** and use writing frames that show different kinds of sentence and

text structures. In addition, the learner should be able to use basic ***punctuation*** and experiment with other punctuation marks. The learner should also be taught how to use punctuation appropriately and when to make use of spelling rules, strategies, and phonics to assist in spelling familiar and unfamiliar words correctly. Learners should be encouraged to use a thesaurus as well as identify synonyms and antonyms (National Department of Education, 2002b, 2002c). In addition to the writing skills mentioned above the learner should also be taught to be critical of their own work. The learner should be able to edit his/her own work by means of deleting or adding words to clarify meaning, re-ordering sentences. ***Proof reading*** forms a substantial part of an editing skill in that corrections are made to drafts of writing by applying knowledge of language in context, focusing on grammar and grammatical rules, punctuation, spelling and vocabulary (National Department of Education, 2002b, 2002c).

Thinking and reasoning (***learning outcome 5***) can be thought of in terms of three (3) components namely ***reason, critical skills, and processing information***. The curriculum is clear on the development of ***critical skills*** in the form of asking questions and searching for explanations, suggesting alternatives and offering solutions, solving puzzles and asking questions for clarification. In terms of the language learning area, critical thinking is articulated in asking critical questions where appropriate. Critical thinking is also displayed in responding critically to texts and being able to reflect on own work as well as that of one's peers. ***Reason*** on the other hand is characterised by inferring and deducing meaning. Reasoning entails identifying and describing similarities and differences with the aim to match things that go together and comparing things that are different. There is an element of classification and separating the parts from the whole. (National Department of Education, 2002b) While reason is characterised by inferring and deducing meaning, ***processing information*** is characterised by assimilating and using information for learning. This is done by means of picking out selected information from a description, organising the information and putting the information in the right order, summarising the information in various ways and categorising and classifying information (National Department of Education, 2002b).

The final learning outcome for languages (***learning outcome 6***) combines key skills touched upon in the other learning areas. Learning outcome 6 deals specifically with ***language structure and use*** where ***vocabulary, grammar, and punctuation*** are vital in creating and interpreting texts. Grammar and punctuation have been addressed in preceding paragraphs and will thus not be discussed again here. ***Vocabulary*** has not been addressed in any depth and will be discussed here. Vocabulary entails the understanding of the meaning of words, where words are letters used to form units, which in turn are used in sentences. Learners

should be able to explain and use word families as well as words of the same field of knowledge to develop vocabulary. Learners are also expected to know how languages borrow words from one another and how words change meaning with time. The meanings of words should also be understood in terms of connotative meanings, denotative meanings, implied meanings and multiple meanings could be identified (National Department of Education, 2002b, 2002c).

It would appear from the analysis of the language policy document discussed above that there is overlap between the MidYIS assessment and the intended policy documents. For example, the instructions provide some overlap with the learning outcome 1, which is listening. The main aim of the listening outcome is to enable learners to listen to the spoken word and be able to respond appropriately. The instructions for each scale are read to the learners to ensure standardisation of procedures. Learners have to listen to and understand the instructions in order to complete MidYIS in the correct manner. By the time learners reach Grade 8 they should be proficient in listening. The instructions to learners are read to them but they have to read each question in order to provide an answer. Thus learning outcome 3 (reading and viewing) is represented in MidYIS. Learning outcome 4 (writing) is represented in terms of proof reading, where learners should be able to identify the mistakes included in passages on a Grade 8 level. Finally, learning outcome 6 is present because vocabulary and proof reading contain elements of the structures used in language. Vocabulary has been included because learners should be able to recognise the meaning of words and their synonyms and be able to match words on this basis. Proof reading requires learners to identify mistakes in terms of spelling, grammar, and punctuation.

In order to verify the document analysis undertaken, specialists in the field of language were asked to evaluate MidYIS (refer to Chapter 5). From the specialists perspective the instructions, vocabulary sub-test, and proof reading sub-test were of relevance for the language learning area. Skills needed in order to succeed in these areas are taught in the curriculum, specifically learning outcome 1 (listening), learning outcome 3 (reading and viewing) and learning outcome 6 (language structure and use). Furthermore, one of the specialists indicated that the items in the MidYIS assessment were not biased in terms of gender or race and that the language used is age appropriate. However, the other specialist indicated that although the basic skills were present in the curriculum, certain items would prove difficult for second language learners and that these items should either be modified or replaced (refer to Appendix G for the detailed reports)

Table 6.3 provides a summary of the discussion between the researcher and the specialists on the overlap between skills assessed in MidYIS and the content and skills taught according to the language learning area. During the discussion, both the results of the document analysis and the evaluation reports were considered.

Table 6.3 ***Proposed overlap between the language learning area and MidYIS***

Outcome according to the curriculum documents	Sub-test in the MidYIS assessment	Result of the document analysis and expert appraisal
1) Listening: The learner is able to listen for information and enjoyment, and respond appropriately and critically in a wide range of situations.	All the instructions	The main aim of the listening outcome is to enable learners to listen to the spoken word and be able to respond appropriately. The *instructions* for each sub-test are read to the learners in order to ensure standardisation of procedures. Learners need to pay attention in order to complete MidYIS in the correct manner. By the time learners reach Grade 8, they should be proficient in listening.
3) Reading and viewing: The learner is able to read and view for information and enjoyment, and respond critically to the aesthetic, cultural, and emotional values in texts.	All the instructions Proof reading	Not only are the *instructions* read to the learners, the instructions are also printed on the first page of each sub-test as well as throughout the sub-test. This implies that the learner can read with the administrator or can read independently for meaning. In order to complete the *proof reading* section of MidYIS, learners would have to read the passage in order to make sense of the passage and rectify mistakes in terms of spelling, grammar, and punctuation. Spelling, grammar, and punctuation are skills in which learners should be proficient by the time they enter Grade 8 as emphasis is placed on these skills in preceding grade levels.
6) Language structure and use: The learner knows and is able to use the sounds, words and the grammar of a language to create and interpret texts.	Vocabulary Proof reading	Vocabulary and proof reading contains elements of the structures used in language. *Vocabulary* has been included because learners should be able to recognise the meaning of words and their synonyms and be able to match words on this basis. *Proof reading* requires learners to identify mistakes in terms of spelling, grammar and punctuation

The National Department of Education officials believe that the content and skills in assessments have to be linked to outcomes within the curriculum. From the discussion above it is proposed that there is overlap between the MidYIS assessment and the curriculum. However, of the six language learning outcomes only three language learning outcomes are represented. The MidYIS assessment does not include all of the skills represented in the language curriculum. What the MidYIS assessment does include are the basic proficiency skills needed to succeed in language, namely vocabulary, proof reading, and comprehension. These basic skills form the building blocks for the skills, such as reasoning, in the three learning outcomes not represented in the MidYIS assessment. It is important to note that all six learning outcomes are needed to succeed in language. The ability to speak a language (learning outcome 2), write in a language (learning outcome 4) and think and reason in a language (learning outcome 5) are important if the learner is to be proficient in the language. However, vocabulary needs to be learnt. Likewise, vocabulary and spelling are important for obtaining writing proficiency in a language. The three learning outcomes not assessed by MidYIS are important but proficiency in these three outcomes presupposes a basic knowledge of vocabulary, spelling, grammar, and punctuation. Furthermore, the limited curriculum validity for the language learning areas can be compensated for, if it can be shown that the MidYIS assessment is correlated with academic achievement (see Chapter 7) and thus have predictive validity.

Various facets of validity are investigated in this research and each of these provides information from which inferences can be drawn relating to the validity of MidYIS for the South African context. A distinction was made between the facets specifically between content-related validity and curriculum validity. Traditionally, content-related validity of an assessment ascertains the degree to which items included in MidYIS sample the domain of items for the specific construct under investigation. The MidYIS assessment is a developed abilities assessment and thus falls within the ambit of psychology, and intelligence theory more specifically. However, if the MidYIS assessment is to be used in school settings then it has to be shown that MidYIS is relevant for the context and the curriculum in which it is used. This means that MidYIS (or the South African version called SASSIS) has to provide information that educators can use to develop intervention programmes where necessary. The content of the programmes will undoubtedly be rooted in the curriculum. For this reason it is important to determine that MidYIS had curriculum relevance in terms of skills assessed. A skill, according to the Merriam-Webster dictionary (2006) is “the ability to use one's knowledge effectively” or a “learned power of doing something competently…a developed aptitude or ability”. According to Atherton (2003), a skill incorporates knowledge in terms of possession or accessibility. Drawing on the definition provided by the Merriam-Webster

Dictionary a skill is learnt and incorporates competency or proficiency. Proficiency was regarded as the level of knowledge or insight that learners have attained (Claassen, van Heerden, Vosloo & Wheeler, 2000). As MidYIS assessment is a developed abilities assessment, abilities have to be taught or included. In the context of the school environment, this implies that the skills or abilities should be rooted in the curriculum policy documents because the curriculum documents provide guidelines to educators as to what should be taught.

MidYIS does, however, have limited curriculum relevance for the language learning area and taking into account the concerns, perhaps additional scales should be added. However, this would substantially increase the time needed to administer MidYIS assessment in one sitting. The additional time needed may impact negatively on the school's timetable and schools may be less inclined to participate in the study. A possible solution to the lack of overlap between the assessment and the curriculum could be to develop a follow-up assessment that is more diagnostic in nature and more comprehensive in terms of the skills included in the language learning area. The diagnostic assessment could then be administered to learners who may benefit from an intervention programme, at a time convenient for the school. The intervention programme could then be tailored according to the results of the intervention programme.

6.2.3 The mathematics learning area

The aim of this section is to provide an answer to the question of whether the mathematical skills in the mathematics learning area curriculum document are sufficiently represented in MidYIS. Mathematics in terms of the South African mathematics curriculum is defined as a human activity that involves observing, representing and investigating patterns and relationships. Mathematics is seen as a product of investigation by different cultures – a purposeful activity in the context of social, political, and economic goals as well as constraints (National Department of Education, 2002d).

Within this framework certain features and/or skills can be identified, all of which are encapsulated in the curriculum. The features and/or skills include working with numbers, data, space, and shape, visualising, measuring, ordering, calculating, estimating, interpreting, making informed choices, comparing, contrasting, classifying, and representing. Furthermore, the learner should be able to display ***critical and insightful reasoning and interpretative and communicative skills*** when dealing with mathematical and contextualised problems (National Department of Education, 2002d).

Five learning outcomes can be distinguished in the mathematics learning area, as presented in Table 6.4. For the purposes of this discussion, only the first four learning outcomes are discussed as the fifth outcome (data handling) is not represented in the MidYIS assessment. The mathematics curriculum, as does the language curriculum, follows the principle of scaffolding where basic information is taught and learnt at the lower-levels while the level of sophistication of required knowledge being mastered increases with every grade.

Table 6.4 ***Outcomes in the mathematics learning area***

Learning outcome		Aim of the outcome
Learning outcome 1	Numbers, operations and relationships	To enable the learner to recognise, describe numbers and represent numbers and their relationships. In addition, the learner is also enabled to count, estimate, calculate, and check with competence as well as confidence when solving a range of problems.
Learning outcome 2	Patterns, functions and algebra	To enable the learner to recognise, describe and represent patterns and relationships as well as use algebraic language and skills in solving problems.
Learning outcome 3	Space and shape	To enable learners to describe as well as represent characteristics of and relationships between 2-D shapes and 3-D objects in terms of different orientations and positions.
Learning outcome 4	Measurement	To enable learners to use appropriate measuring units, instruments, and formulae in a variety of contexts
Learning outcome 5	Data handling	To enable learners to collect, summarise, display, and critically analyse data in order to draw conclusions and make predictions

(Source: National Department of Education, 2002d)

Highlighted in the policy is numbers sense, as this entails knowledge of basic number facts and also of accurate methods for calculation and measurement by means of a range of strategies for estimating and checking results. Learners with a ***good sense of number and operations*** have the mathematical confidence to make sense of problems in various contexts. ***Learning outcome 1* (*numbers, operations, and relationships)*** entails being

able to describe and recognise numbers. This includes knowing what numbers mean and being able to identify how numbers relate to one another, knowing the relative size of numbers and how to order and compare numbers in terms of more, less or equal. In addition, the learner should be able to manipulate numbers by adding, subtracting, multiplying, dividing, building up numbers, breaking down numbers, rounding off and compensating. The learner should have an understanding of whole numbers, place value, fractions and decimal fractions, percentages, decimals, ratio, rate and be able to convert numbers from one form to another. The learner, according to policy, should be able to use a range of techniques and tools at his/her disposal to perform calculations efficiently and to the required degree of accuracy (National Department of Education, 2002d).

Learning outcome 2 (***patterns, functions and algebra***) focuses on ***patterns and relationships*** and on making use of ***algebraic skills to solve problems***. A key element and focus area of this learning outcome is the ability to describe patterns and relationships**,** using symbolic expressions, graphs, and tables. Also of importance is the ability to identify and ***analyse regularities and changes in patterns and relationships*** to be able to make predictions and solve problems. ***Numeric and geometric patterns*** are investigated and extended in order to establish relationships between variables or express rules governing patterns in algebraic language or symbols. The patterns and relationships should be explained so that the rules used could be justified. Patterns and relationships are important elements in algebra. A central part of learning outcome 2 is for the learner to achieve efficient manipulative skills using algebra. The study of algebra begins with writing number sentences to describe a problem situation, solving or completing number sentences by inspection or by trial-and-improvement and checking the solutions by substitution. Learners will also be able to write algebraic expressions, formulae, or equations in simpler or more useful equivalent forms in context and to interpret and use algebraic vocabulary in context (National Department of Education, 2002d).

Learning outcome 3 is the study of space and shape. According to policy, the study of space and shape improves understanding and appreciation of the pattern, precision, achievement, and beauty found in natural and cultural forms. The focus of this outcome is on the ***properties, relationships, orientations, positions and transformations of two-dimensional shapes as well as three-dimensional objects***. The aim of the learning outcome is to enable the learner to describe and represent characteristics of and relationships between two-dimensional shapes and three-dimensional objects. The learner should be able to ***recognise, identify, sort, and compare*** two-dimensional as well as three-dimensional objects. The learner should also be able to identify three-dimensional objects

from different positions and orientations. As in every outcome, there is a progression from simpler forms to more complex forms. In the case of learning outcome 3, the learner first starts with two-dimensional shapes and progresses to three-dimensional objects, geometric objects, and shapes. The outcome culminates in making use of transformations, congruence, and similarity in order to investigate, describe, and justify properties of geometric figures and solids (National Department of Education, 2002d).

Measurement (***learning outcome 4***) focuses on the ***selection and use of appropriate units, instruments, and formulae to quantify characteristics*** of events, shapes, objects, and the environment. It is suggested in policy that the study of measurement should be introduced by means of using everyday occurrences such as describing time of day in terms of day and night and concretely comparing objects using appropriate language to describe mass, capacity, and length. The learner should be able to use time-measuring instruments to appropriate levels of precision in order to describe and illustrate ways of representing time. Furthermore, learners should be able to estimate, measure, record, compare, and arrange two-dimensional shapes and three-dimensional objects. S.I. units should be used with appropriate precision for mass (grams and kilograms), capacity (millilitres and litres), length (millimetres, centimetres, metres, and km), and temperature using degree Celsius. Learning outcome 4 aims to ***expand knowledge of measurement through various investigative activities*** such as time, distance, speed as well as derive rules for calculating measurements relating to geometric figures and solids (National Department of Education, 2002d).

Initial indications are that it would appear from the policy documents that there is some agreement between the MidYIS assessment and the mathematics curriculum document. Out of all the sub-tests included in the MidYIS assessment, the mathematics sub-test is the most curriculum-bound, as internationally, there is convergence in terms of the mathematics curricula, especially at the Grade 8 level (TIMSS 1999, 2003 are examples of international studies where this was found). In the mathematics scale, various items are included which can be linked to learning outcome 1 (numbers, operations and relationships) in terms of various grade levels from basic number manipulations to more complex calculations all of which are in line with the curriculum until Grade 8. In the mathematics sub-test, various measuring units and formulae are used (learning outcome 4: measurement). The type of items included is grade appropriate in that learners should have been exposed to the skill in preceding grade levels. Furthermore, learning outcome 2 (patterns, functions and algebra) is represented in both perceptual speed and pictures sub-tests as these sub-tests include items where learners need to find or complete the pattern given while the mathematics section includes algebraic equations, all of which are reasonable for Grade 8. Block counting and

cross-sections are measures of spatial ability, thus these two sub-tests are representative of learning outcome 3 (space and shape). Spatial ability requires certain skills in 2D and 3D manipulation. These two sub-tests are in line with the basic skills that are taught in this learning area in order to prepare learners to be successful in geometry.

In order to verify the results of the mathematics document analysis, a mathematics specialist was consulted. The mathematics specialist was asked to develop assessment specifications (refer to Appendix J) that match items to learning outcomes. Mathematics has set laws, principles, and operations that are universal in nature and the level of complexity is easier to ascertain as compared to the language learning area. For example, adding and subtracting are taught first and are less complicated than multiplication and division. Multiplication and division make use of the principles taught in adding and subtracting. A similar table was not constructed for the language learning area, as the language learning area provides the challenge of characterising the tasks in the proof reading sub-test as easy, moderate, or difficult on an item basis as a passage is presented. The vocabulary that learners should be exposed to is not set out in the same manner, nor is it clear from the policy documents in terms of the complexity or sophistication of words introduced at each grade level. A summary of the mathematics framework is provided in Table 6.5.

Table 6.5 ***Accessibility of mathematics items***

Mathematics Learning outcome	Number of items N=154	%	Accessibility with regard to the grade level (Grade 7 (end) and/or Grade 8 (beginning)			
			Very easy	Easy	Moderate	Difficult
Learning outcome 1: Numbers, operations and relationships	45	29%	39%	4%	31%	27%
Learning outcome 2: Patterns, functions and Algebra	51	33%				
Learning outcome 3: Shape and space	52	34%				
Learning outcome 4: Measurement	6	4%				
Learning outcome 5: Data handling	0	0				

The specialist, however, raised a concern that certain items were excessively easy. However, the MidYIS assessment is a combination of a speed and power assessment as was discussed in Chapter 5. The difficulty therefore does not necessarily stem from the item but the fact that a number of items have to be completed within a time limit. The mathematics specialist felt that the time allocation was not sufficient and suggested that the time allocations be revisited. As can be seen from Table 6.5, 43% of the mathematics section was considered easy by the specialist. What makes the section more difficult is the time allocated to the section.

The items' degree of difficulty, according to the specialist, ranged from very easy to difficult (refer to Table 6.5), which is consistent with sound assessment practices (the item difficulties will be elaborated on further in Chapter 7). The inclusion of easier items is in line with the type of assessment where time limits and speed are factors. As the MidYIS assessment is a combination of a speed and power assessment, as indicated in Chapter 5, more difficult items have been included. The mathematics specialist also indicated that even though certain items were not present in the mathematics curriculum they would still be accessible to an average Grade 8 learner due to general knowledge, experience, and problem solving strategies.

During the discussion with the mathematics specialist, the results of the document analysis were presented. The evaluation of the mathematics specialist concurred with the results of the document analysis (refer to Table 6.6). The mathematics specialist indicated that the skills needed for four out of the five learning outcomes were represented in MidYIS, namely learning outcome 1 (numbers, operations and relationships), learning outcome 2 (patterns, functions and algebra), learning outcome 3 (space and shape) and learning outcome 4 (measurement) (refer to Table 6.6), with no items representative of learning outcome 5 (data handling).

Table 6.6 ***Proposed overlap between the mathematics learning area and MidYIS***

Outcome in accordance with curriculum documents	Sub-test in the MidYIS assessment	Result of the document analysis and expert appraisal
1) Numbers, operations and relationships: The learner is able to recognise, describe and represent numbers and their relationships and can count, estimate, calculate and check with competence and confidence in solving problems.	Mathematics	Out of all the sub-tests included in the MidYIS instrument mathematics is the most curriculum-bound. Mathematics and all the elements that go with it are skills that have to be taught. In the mathematics sub-test various items are included which can be linked to learning outcome 1 at various grade levels from basic number manipulations to more complex calculations all of which are in line with the curriculum until Grade 8.
2) Patterns, functions and algebra: The learner is able to recognise, describe and represent patterns and relationships, and solve problems using algebraic language and skills.	Perceptual speed and accuracy Pictures Mathematics	Both perceptual speed and pictures include elements of finding or completing the pattern given, while the mathematics section includes algebraic equations all of which are reasonable for Grade 8.
3) Space and shape: The learner is able to describe and represent characteristics of and relationships between 2-D shapes and 3-D objects in a variety of orientations and positions.	Block counting Cross- sections	Block counting and cross-sections are a measure of spatial ability. Spatial ability requires certain skills in 2D and 3D manipulation. These two sub-tests are in line with the basic skills that are taught in this learning area in order to prepare learners to be successful in geometry.
4) Measurement: The learner is able to use appropriate measuring units, instruments, and formulae in a variety of contexts.	Mathematics	In the mathematics sub-test, various measuring units and formulae are used. The type of items included is grade appropriate in that learners should have been exposed to the skill in preceding grade levels.

From the preceding section, it is clear that there is overlap between the MidYIS assessment and the curriculum. However, learning outcome 5 (data handling) is not represented at all. This raises some doubt about the extent of the curriculum validity of MidYIS in terms of the mathematics learning area. However, validity cannot be thought of in absolute terms. Instead, validity is best thought of in terms of a continuum ranging from high to low (see Chapter 5). In any assessment, the challenge remains to cover a range of skills, given practical considerations, such as time. Even though learning outcome 5 is not represented in MidYIS, the basic skills needed to succeed in data handling are found in the other four learning outcomes. Furthermore, inferences drawn in terms of the curriculum validity of the MidYIS assessment are strengthened if the correlations between the MidYIS assessment and academic performance in mathematics are high (the predictive validity of the MidYIS assessment will be addressed in Chapter 7).

Based on the results of the document analysis and evaluation report from the mathematics specialist, it would appear that MidYIS does have a degree of curriculum validity. Items that are easy, moderate, and difficult should however cover all learning outcomes. If MidYIS is to include items completely representative of the skills included in the curriculum, then items related to data handling (learning outcome 5) should be included in future versions of the South African adaptation of MidYIS.

6.2.4 Exploring possible suggestions for the revision of MidYIS

In 6.2.2 and 6.2.3, the curriculum validity of MidYIS was explored. Clearly, there is overlap between the skills taught in the curriculum and the skills assessed in MidYIS. Suggestions can be put forward to adapt MidYIS to the South African context. These suggestions draw on the document analysis and evaluations from the specialists (both language and mathematics). The adaptations relate to items, administration procedures, and format. Items for example could be either rewritten or added in accordance with learning outcomes not covered in MidYIS. The adaptations discussed in this session are related to the discussion in Chapter 4, Section 4.6 as well as the specific research question 1.3 (as described in Chapter 5) ***what adaptations are needed to transform MidYIS into SASSIS, a monitoring system for the South African context.*** The specific adaptations suggested are represented under the sub-research questions identified. Each sub-question focuses on an important adaptation. These include administration procedures, level of language, format of the assessment and time allocations. Under each of the sub-questions the adaptations are described for each of the areas identified.

1.3.1 To what extent are the administration procedures appropriate and if not, how can these be adjusted?

The expert evaluation reports from the language specialists indicated that the instructions could be ambiguous and difficult to follow. The instructions were therefore rewritten so that learners could better understand what was expected of them. The modifications were based on the suggestions given by the language specialists.

The mathematics specialist indicated that the instructions for some sections were ambiguous, specifically citing cross-sections and block counting.

The instructions were modified in collaboration with the mathematics and language specialists. The modified instructions were translated and back translated to ensure accuracy and congruence between translated versions.

The second area of adaptation refers specifically to the level of language used in the assessment.

1.3.2 To what extent is the content in MidYIS appropriate for second language learners?

The language specialists indicated that the more complex words would not be accessible to second language learners, especially items in the vocabulary section. The specialist felt that certain words in the vocabulary section were ambiguous and that the way in which the words were presented was not in line with how vocabulary was taught. As a result, the vocabulary sub-test was revised on the basis of the suggestions provided by the language specialists. The specialists provided options for replacing problematic or ambiguous words. The core word for which a synonym had to found was placed within the context of a sentence. One of the language specialists provided the context sentences. The sentences provided were then reviewed to ensure that they were valid, specifically whether the sentences included any gender and cultural bias.

It is suspected that the as a result, the items may be easier but more accessible to second language learners (the difficulty of the items is reported on in Chapter 7). Furthermore, when data from the pre-pilot is reviewed in conjunction with performance on the reviewed vocabulary items it was found that the mean score was 40% as compared to 47%.

The mathematics specialist felt that some items might be inaccessible for some second language speakers because of the length and level of written language included. The items flagged were discussed and the suggested changes effected in collaboration with the specialist.

The third area of adaptation refers to whether the format of the assessment was appropriate.

1.3.3 To what extent is the format of the assessment appropriate and if not, how can it be changed?

The language specialists pointed out that learners, when unsure of what to do, would have to page to the beginning of the sub-test in order to reread the instructions. This would waste time. For this reason, the instructions were included at the top of the page, at the suggestion of the specialists, throughout MidYIS so that learners if uncertain could reread the instructions.

The fourth area of adaptation refers to whether the time allocations were appropriate.

1.3.4 To what extent are the time allocations appropriate and if not, what adjustments are needed?

The language specialists were not satisfied with the time limits allocated for various sections of MidYIS. The time allocated to each sub-test was therefore increased so as to allow the majority of the learners to complete or almost complete the sub-test. Time limits were decided upon in collaboration with the language specialists. A key consideration was the nature of MidYIS, which is a combination of a speed and power assessment. The assessment was administered during a formal school day. This meant that the allocated time had to fit into the school's timetable so as to not overly impose on teaching time.

Time was a major issue for the mathematics specialist. He felt that some learners would not be able to finish (or nearly finish) certain sections, among them mathematics, cross-sections and block counting. The time allocations were discussed with the mathematics specialist and the same procedure applied as with the language specialists. It was agreed that the time allocations would be adjusted so that learners would have at least 30 seconds to complete an item.

6.3 Exploring the content validity of the MidYIS assessment

In Chapter 5, the concept of validity was addressed as opposed to different types of validity. The facets are the traditional terms of content validity, face validity and construct validity. Acknowledging the view that validity is a unitary concept, it is for conceptual and analytical reasons easier to separate it into facets and to address these individually. Content-related validity issues are addressed from a curriculum perspective and explored from a psychometric perspective. It is thought that this approach would add depth to inferences drawn because the exploration would not only be from a curriculum perspective but would also draw on the theory base related to ability testing. Content-related validity issues from a psychometric perspective addresses the sub-research question (question 1.2.3 as described in Chapter 5) ***to what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa.***

MidYIS is a developed abilities assessment. Developed abilities are the common ground between intelligence, aptitude, and achievement and reflect the effects of experience and learning (Reschly, 1990). Developed abilities can also be thought of in terms of skills or competencies (Merriam-Webster, 2006). Competence, according to Kouwenhoven (2003, p. 43,), is the "state of being competent", "the capability [ability] to choose and use (apply) an integrated combination of knowledge, skills and attitudes" (Kouwenhoven, 2003, p. 71). Ability may refer to cognitive traits used when solving problems where cognitive refers to information processing. If it is said that an assessment is the measuring of developed abilities then aspects of developed abilities are covered (Kline, 2000). From a curriculum perspective, it means that the skills taught in the curriculum are included in MidYIS. From a psychometric perspective, this means that the abilities or skills to be assessed are covered in the field of ability testing.

The systems developed by the CEM centre all have these characteristics in common (see Chapter 4) and stem from the need to have an assessment that could predict future performance but which was not curriculum-based. At the time when the first system, the Advanced Level Information System (Alis), was being developed, the publishing of league-tables had started to take effect (Fitz-Gibbon, 1996). There was a need to have an alternative assessment apart from the Key Stage examinations on which the league-tables were based. The Key Stage examinations are curriculum-driven, thus a developed abilities assessment was used. Developed abilities, although not strictly curriculum-based, do provide a measure of proficiency in basic skills needed to succeed academically (refer to Chapter 2).

In a developed abilities assessment, both generic competencies as well as domain specific competencies are assessed. Generic competencies are skills, which are transferable to other situations whereas domain-specific skills are skills associated with a specific content domain (Kouwenhoven, 2003). For example, in MidYIS the mathematics sub-test is specific to the mathematics domain, i.e. domain specific, while perceptual speed and accuracy may be used in mathematics to find a mistake in an equation and in geography to find a location on a map.

It is important to explore whether the sub-tests in MidYIS are comparable to sub-tests of other ability assessments (refer to Chapter 5). Researchers using factor analysis in an effort to understand the "nature of human abilities" (Kline, 2000, p. 69) have identified key ability factors. Table 6.6 provides a summary of the various ability factors (Cooper, 1999; Hunt, 1985; Kline, 1993, 2000; Sternberg, 1985). For the purposes of this discussion, only the factors which are assessed in current ability, aptitude assessments, and MidYIS are included in Table 6.7 (see Appendix K for a comprehensive list of ability factors).

Table 6.7 ***Summary of ability factors associated with abilities or aptitude assessments***

Ability	Definition of the ability	Assessment in which ability is found
Verbal ability, verbal comprehension and verbal relations	Denotes the understanding of words (Kline, 2000) as measured by tests of vocabulary and reading comprehension (Sternberg, 1985), using words in context: understanding proverbs, verbal analogies and vocabulary (Cooper, 1999).	General Scholastic Aptitude Test Battery (GSAT) Senior South African Individual Scale (SSAIS) South African Wechsler Adult Intelligence Scale (WAIS) Junior Aptitude Test (JAT) Senior Aptitude Test (SAT) Washington-Pre-College Test Battery Wechsler Intelligence Scale for Children (WISC) Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Grammar or language usage	Measured by means of identifying poor grammar and correcting errors (Hunt, 1985)	Washington-Pre-College Test Battery Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Spelling	Denotes the recognition of misspelled words (Kline, 1993).	Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Numerical ability	Facility in the manipulation of numbers but does not include arithmetic reasoning (Kline, 2000).	General Scholastic Aptitude Test Battery (GSAT) Senior South African Individual Scale (SSAIS) Junior Aptitude Test (JAT) Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Numerical facility	Denotes the ability to use algebra and other forms of mathematical operation (Cooper, 1999).	South African Wechsler Adult Intelligence Scale (WAIS) Junior Aptitude Test (JAT) Senior Aptitude Test (SAT) Washington-Pre-College Test Battery Wechsler Intelligence Scale for Children (WISC) Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Spatial ability	Ability to recognise figures in different orientations (Sternberg, 1985; Kline, 2000).	Junior Aptitude Test (JAT) Senior Aptitude Test (SAT) Washington-Pre-College Test Battery Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Perceptual speed and accuracy	Denotes the ability to rapidly assess differences between stimuli (Kline, 2000) and measured by the rapid recognition of symbols (Sternberg, 1985)	Junior Aptitude Test (JAT) Senior Aptitude Test (SAT) Wechsler Intelligence Scale for Children (WISC) Differential Aptitude Test (DAT) Middle Years Information System (MidYIS)
Speed of closure	The ability to complete a pattern with a part missing (Kline, 2000).	General Scholastic Aptitude Test Battery (GSAT) Senior South African Individual Scale (SSAIS) South African Wechsler Adult Intelligence Scale (WAIS) Wechsler Intelligence Scale for Children (WISC) Middle Years Information System (MidYIS)

Table 6.7 provides a summary of the various types of ability factors prominent in ability or aptitude assessments. In order to make inferences of the content-validity from a psychometric perspective, specialists in the field of psychology were asked to evaluate MidYIS. An educational psychologist as well as two research psychologists formally reviewed the MidYIS instrument. The brief was to review MidYIS for content-related validity specifically in terms of intelligence or ability theory. Whether MidYIS was similar to other developed abilities assessment such the as the Wechsler Intelligence Scale for Children (WISC) or Differential Aptitude Test (DAT) had to be evaluated, also whether the language was appropriate and any biases were obvious in terms of gender or race. The outcome of the reviews indicated that the sections represented in the MidYIS assessment do correspond with the domain of items found in ability assessments; specifically the ability factors of verbal ability, comprehension and relations, spatial ability, grammar or language usage, perceptual speed and accuracy and numerical ability and facility (see Table 6.7). The psychologists indicated that the items were not biased in terms of language or gender. However, the psychologists pointed out that they could not comment on the difficulty of the vocabulary and mathematics sections specifically as they were not content specialists.

6.4 Conclusion

The aim of this chapter was to address issues associated with the content-related validity of the MidYIS assessment. The content-related validity of MidYIS can be evaluated from two perspectives namely a curriculum perspective and a psychometric perspective. Although these two perspectives are addressed separately, there is an apparent link between the two. From a psychometric perspective, MidYIS is a developed abilities assessment. Ability is a competence in, a skill or an aptitude. The current curriculum has its roots in competency-based education (Kraak, 1998). Competence can refer to general intelligence or aptitude, as motivation or as a set of key competencies or skills (Kouwenhoven, 2003). Due to the nature of the relationship between MidYIS as a developed abilities assessment and the curriculum with its roots in competency-based education, both aspects had to be explored. The curriculum perspective is reflected in the sub-question ***to what extent are the skills tested by MidYIS valid for the South African curriculum*** while the psychometric perspective is reflected in the sub-question ***to what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa***.

The sub-question ***to what extent are the skills tested by MidYIS valid for the South African curriculum*** was explored by means of curriculum document analysis and specialist evaluations, while background information was provided by the National and Provincial

Department of Education. The clear message from the National and Provincial Departments of Education was that assessment used in a school setting must be aligned with the curriculum. In order to explore the alignment of the MidYIS assessment with the South African curriculum, document analysis was undertaken and specialists consulted. Two learning areas were selected namely language and mathematics as the fundamental skills assessed in MidYIS corresponded with these two learning areas (refer to Chapter 5).

For the language learning area three of the six outcomes were represented in the MidYIS assessment indicating a moderate alignment between MidYIS and the curriculum. However, the skills assessed in the MidYIS assessment which can be found in the curriculum refer to the basic skills needed for example vocabulary. Teal (2003) is of the opinion that vocabulary knowledge is one of the best predictors of reading comprehension. Vocabulary knowledge provides a source of prior knowledge and word meaning that can be used to enhance reading comprehension. In addition, word recognition is considered an essential goal (Artley, 1996), as well as reading comprehension, decoding and language comprehension (Aarnoutse & Brand-Gruwel, 1997). Word recognition and comprehension are important because if a learner becomes better at reading, s/he will be able to read more difficult texts resulting in a larger vocabulary and syntactic knowledge that in turn positively affects language ability (Aarnoutse & Brand-Gruwel, 1997). It is clear that even though the MidYIS assessment does not directly include three of the six learning outcomes, what it does include is the basic skill that is needed to succeed in the other learning outcomes. However, it is possible to construct additional scales that do directly relate to the other learning outcomes, such as reading a passage and answering questions relating to the passage. The act of reading helps to increase the learner's vocabulary and also his/her awareness of language and structure of text (McFarlane, 1997). By including an additional section, learner reading skills and comprehension can be directly assessed.

Inferences in terms of curriculum validity for the mathematics learning area are substantially stronger because four of the five learning outcomes are represented in MidYIS. The acquisition of mathematical problem-solving and reasoning skills in addition to the ability to apply the skills to mathematical situations and real-life situations constitutes a major goal or objective of mathematics education (Verschaffel, 1999). A primary goal of mathematics education is to enable learners to apply their knowledge of facts, concepts, formulas, and procedures in order to solve problems in a variety of learning situations (Muth, 1997). Solving mathematics problems requires learning of domain-specific knowledge that is well structured and flexible, including content, procedures and reflective knowledge, in order to be able to

solve the given problem (Nelissen, 1999). In order to solve problems, learners need to have basic mathematical skills and be able to observe, relate, question, and infer. To solve mathematical problems learners must be able to reason about ideas, see the relationships and connections, and be able to make sense of mathematics. Learners should be able to draw conclusions, induce patterns, and deduce ideas resulting in learners having the ability to use models and mathematical ideas to explain thinking (Holmes, 1995). In order to be able to explain thinking learners should have basic mathematical skills that can be built upon (Cathcart, Pothier, Vance & Bezuk, 2003). It would appear from the document analysis and specialist evaluation that MidYIS has a reasonable degree of curriculum validity. However, the proposal for additional items pertaining to the outcome currently not represented would make inferences drawn that much stronger.

The second sub-question addressed relating to content-related validity is ***to what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa***. The evaluations from the psychologists indicate that the items in the MidYIS are in agreement with the ability domain. Furthermore, MidYIS is comparable to other ability assessments currently used in South Africa such as the Differential Aptitude Test (DAT) and is not biased in terms of gender or race.

Finally, it is clear that adaptations had to be made to MidYIS to make it relevant for South Africa. Some of the adaptations are easier to effect than others. Adaptations that are needed, range from allocating more time per sub-test to possibly including additional sub-tests. To answer the sub-question of ***what adaptations are needed to transform MidYIS into SASSIS, a monitoring system for the South African context*** the following suggestions have been made:

1.3.1 To what extent are administration procedures appropriate and if not how can they be adjusted?

The expert evaluation reports indicated that the instructions could be ambiguous and difficult to follow. Thus the instructions were revised, on the basis of the suggestions provided by the specialists, so that learners would understand what was expected of them but that the rewritten version would still be comparable to the original.

1.3.2 To what extent is the content in MidYIS appropriate for second language learners?

The specialists indicated that a number of items would not be accessible to second language learners. The specialists identified items and provided feasible alternatives. The changes suggested by the specialists were effected.

1.3.3 To what extent is the format of the assessment appropriate and if not, how can it be changed?

Overall the format of MidYIS was acceptable However, the specialists indicated that when unsure of what to do, learners would have to page to the beginning of the sub-test in order to reread the instructions. This would waste time. The instructions were therefore included at the top of the page throughout MidYIS, as suggested by the specialists, so that learners could reread the instructions without wasting time.

1.3.4 To what extent are the time allocations appropriate and if not, what adjustments are needed?

The specialists were not satisfied with the time limits allocated to various sections of MidYIS. Therefore, the time allocated to each sub-test was increased, using the recommendations of the specialists so that the majority of the learners would be able to complete or almost complete sub-test. This is also in accordance with the type of assessment, as MidYIS is a combination of a speed and power test as was discussed in Chapter 5.

CHAPTER 7

THE CONSTRUCT-RELATED VALIDITY AND RELIABILITY OF THE MIDYIS ASSESSMENT

This chapter details the investigation into the construct-related validity and reliability of the MidYIS assessment. Several analytical strategies are included that address sub-questions related to the specific research question ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** *These strategies were designed to answer whether the data support the MidYIS scales, whether the results of the assessment are reliable and whether the results could be used to predict future performance. In order to explore the data structure Rasch analysis were used. Reliability analysis was undertaken to investigate the consistency of results while correlation analysis was used as a preliminary step in investigating the possibility of using the results of the MidYIS assessment to predict academic performance.*

7.1 Introduction

This chapter represents the second of the results chapters and portrays the outcome of the reliability and validation strategies relating to the first main research question ***how appropriate is the Middle Years Information System (MidYIS) as a monitoring system in the South African context.*** More specifically the chapter addresses the specific research question 1.2 (as discussed in Chapter 5) ***how valid and reliable are the data generated by the MidYIS monitoring system for South Africa?*** In Chapter 3, the main research question and the specific research questions were discussed in terms of criteria for evaluating the quality of measurements and how one would collect information in order to make inferences related to the quality of measurements (specifically that of validity and reliability; see Figure 3.3). In Chapter 5, the figure presented in Chapter 3 (3.5) was elaborated upon.

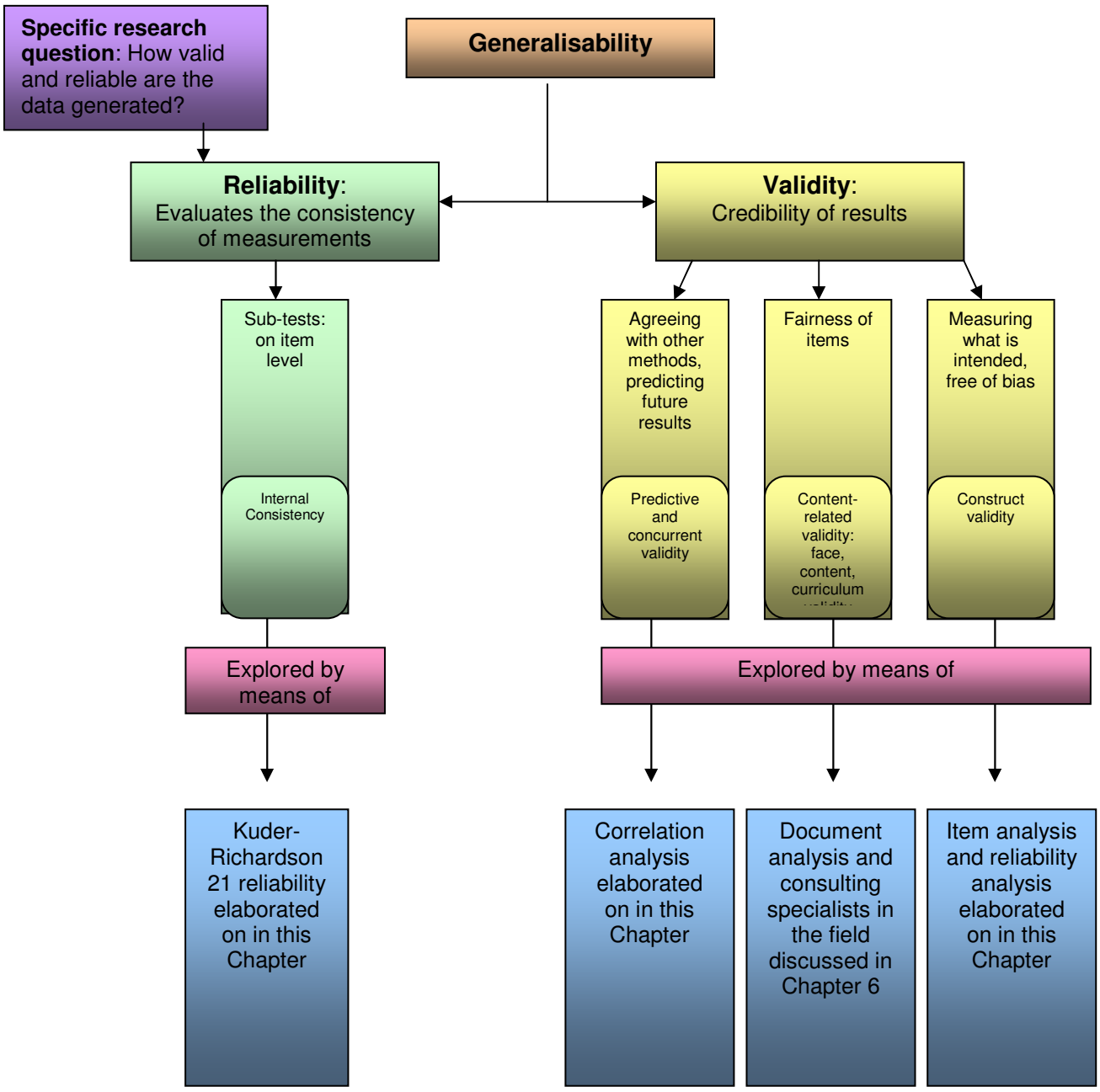


Figure 7.1 **Extension of the criteria for evaluating quality of measurement used in monitoring systems (adapted from Fitz-Gibbon, 1996)**

In order to address specific research question 1.2 adequately, five sub-questions were identified (see Chapter 5):

1.2.1 To what extent are the results obtained on MidYIS reliable?

1.2.2 To what extent are the skills tested by MidYIS valid for the South African curriculum?

1.2.3 To what extent are the items in MidYIS in agreement with the domain of ability testing and applicable for South Africa?

1.2.4 How well do the items per sub-test function and do they form well-defined constructs?

1.2.5 To what extent could the data predict future achievement?

The two sub-research questions 1.2.2 and 1.2.3 were addressed in Chapter 6. The focus of this chapter is on the empirical analysis associated with the validation strategies and reliability analysis (sub-questions 1.2.1, 1.2.4, and 1.2.5).

Validity is seen as a unitary concept as described in Chapter 5. Validity is, in the words of Messick (1989, p. 5), "…an integrated evaluative judgment of the degree to which empirical evidence and rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment".

All instruments or measures are faced with the challenge of establishing validity. This is reflected in the "theoretical value, empirical value or applied value" as stated by Greenwald, Nosek and Sriram (2006, p. 56). It seldom occurs that an instrument will have no validity or 100% validity. Rather, the idea here is to make inferences based on a continuum. Thus an instrument or measure may provide some evidence about a person's level on a construct but may not necessarily represent everything included in the definition of the construct (Reckase, 1998).

In order to make sound judgments regarding validity more generally, both statistical and judgmental analyses are required (Sireci, 1998). For this reason, the current research included investigations into the content-related validity of the assessment, as was discussed in Chapter 6, as well as statistical or more "empirical" analyses that are presented in this chapter. Specifically, construct validity in terms of empirical evidence and predictive validity are discussed in this chapter. These are discussed separately, as a distinction can be drawn between the facets of validity as was explained in Chapter 5, under the banner construct-related validity (5.3.3). This also provides a way of addressing conceptual aspects of validity (Messick, 1989). It has to be understood, however, that (Messick, 1998, p. 37):

> All validity is of one kind…Other so-called separate types of validity – whether labeled content validity, criterion-related validity, consequential validity, or whatever – cannot stand alone in validity arguments. Rather, these so-called validity types refer to complementary forms of evidence to be integrated into an overall judgment…What needs to be valid are the inferences made about score meaning, namely, the score interpretation and its action implications for test use.

For the purposes of this research, construct validity is viewed as the extent to which an assessment measures a particular construct which is inferred from theory (Huysamen, 1996). An assessment intends to measure predefined constructs. However, it has to be established whether the items are functioning as they should. MidYIS was designed to measure seven constructs, each forming the sub-tests of the assessment, namely vocabulary, mathematics, proof reading, perceptual speed and accuracy, cross-sections, block counting and pictures (see Chapter 4). Whether the items in each of the sub-tests measure the same trait in South Africa has to be established. Thus Rasch analysis was used for this purpose.

In addition to construct validity, predictive validity was explored, specifically whether the South African data can be used to predict future academic performance. In the United Kingdom, MidYIS is used to predict future achievement, in addition to calculating the value the school has added to learners (see Chapter 4). Statistical procedures such as correlation analyses and ordinary least squares analyses (also referred to as regression analysis), have been undertaken by the CEM centre. The same procedures have to be undertaken in the South African context, if the assessment is to be used in the same way.

The first step was to explore whether there are any correlations between the MidYIS scores and academic performance (Kline, 1993; Huysamen, 1996). According to Huysamen (1996, p. 33) “this correlation tells us how accurately ultimate success” can be predicted. The second step was to draw a nationally representative sample, administer the assessment and then correlate the data with academic performance as defined by national written examinations. The South African data is not used to predict performance as this is beyond the scope of this thesis. However, initial groundwork is presented here in an effort to establish whether relationships exist between the MidYIS scores and academic performance as obtained from the schools.

The concept of reliability is also addressed in this chapter and is viewed in terms of the consistency of results, and was detailed in Chapter 5. Reliability analysis can also be used the strengthen inferences pertaining to construct validity (Gronlund, 1998), as the analysis identifies items which appear to be measuring a different trait. Many factors may improve the reliability of an assessment, such as the test length, item type, assessment administration procedures and time limits (Traub & Rowley, 1991). However, before issues of validity and reliability are addressed, the participants are described in terms of age, gender and population group (7.2). How well the items are functioning for each sub-test is addressed in 7.3, while reliability is explored in 7.4. Whether the MidYIS scores can be used for prediction

purposes is detailed in 7.5. Finally, in the conclusion section (7.6) main inferences drawn from the analyses are described.

7.2 Participant characteristics

Seven hundred and ninety-four learners of the same cohort participated in this study. Fifty-one percent of the learners were female. Ninety-three percent of the learners were between the ages of 13-15. It is of interest to note that the older learners tend to be male. Table 7.1 provides details of the age distribution of participating learners.

Table 7.1 ***Age and gender distribution of participating learners***

Age	Number of learners	Percentage of the sample per age group	Percentage male	Percentage female
12	15	2	60	40
13	299	38	43	57
14	320	41	47	53
15	109	14	55	42
16	29	4	62	38
17	9	1	78	22
Overall	**781**	**100%**	**47%**	**51%**

The majority of the learners in the sample were not first language English speakers (see Table 7.2). Only 21% of the learners who responded to the question of home language were first language English speakers. Fourteen percent of learners who responded to the question indicated that their home language was Afrikaans while 12% of learners who responded to the question speak Sepedi in the home (see Table 7.2 for details). Perhaps surprising is the large percentage of learners who did not respond to the question (29%). A possibility is that learners speak more than one language in the home, that they did not want to supply the information or that they preferred not to comment.

Table 7.2 ***Home language of learners who participated***

Home language	Number of learners	Percentage of learners that predominantly speak the language in the home
Afrikaans	107	14
English	167	21
IsiNdbele	8	1
IsiXhosa	3	.4
IsiZulu	33	4
Kirundi	1	.1
Portuguese	1	.1
Sepedi	95	12
Sesotho	56	7
Setwana	72	9
Siswati	4	.5
Tshivenda	13	2
Xitsonga	4	.5
Did not respond	230	29

The majority of the learners in the study were African (69%) while there were fewer learners from other population groups. Fourteen percent of the learners were Coloured, 12% were White and 6% were Indian.

Table 7.3 ***Population group of learners who participated***

Population Group	Number of learners	Percentage
African	545	69
Coloured	110	14
White	91	12
Indian	48	6

7.3 Elaborating on construct validity

Construct validity focuses on identifying an underlying construct inherent in data structures. The constructs are defined by researchers and are based on literature. Theoretical constructs are made explicit by the researcher in an attempt to capture the construct by developing items (Bond & Fox, 2001). This section explores (from a construct validity angle)

how well the items included in the assessment are functioning. This was done by means of Rasch modeling. The Rasch model not only contributes to inferences made about construct validity but also provides "indicators of how well each item fits within the underlying construct" (Bond & Fox, 2001, p. 26). 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. MidYIS has seven sub-tests, which were described in Chapter 4. The seven sub-tests are combined to form four scales and an overall score. In the section to follow, the items are first explored per sub-test as explained previously. Rasch analysis can be used to explore the extent to which items form defined constructs. The sub-tests can then be combined into scales based on the theoretical definitions identified in literature as well as the common skills assessed. The theoretical combination of the sub-tests is akin to the idea on content-related validity where the idea of test quality defined by content-related validity refers to some kind of "domain definition, domain relevance, domain representation, and appropriate test construction procedures" (Sireci, 1998, p. 101).

7.3.1 Investigating construct validity by means of Rasch analysis

The approach to the Rasch analysis was discussed in Chapter 5. For the purposes of the analysis, a dichotomous Rasch model was used. 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 reliabilities is the same, in that a higher value is advantageous (Andrich, 1982).

As described in Chapter 5, the INFIT and OUTFIT statistics were considered. 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). The aim was to identify and retain the best core items, thus, criteria that are more stringent were used. Also, Z-values derived from more than 300 observations tend to be very sensitive in which items that should not misfit do (Linacre, 2005).

The item number and the logit values were displayed on a continuum (Schumacker, 2004) in order to evaluate items and odd ratios (also named odds). 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 odd 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, is the natural log-odds of success on items included in the scale or variable. A positive value here indicates more ability, while a negative value indicates less ability. 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).

The main purpose of undertaking Rasch analysis was to explore the performance of items. Hence, the aim is to identify good items which contribute to the sub-test and poor items in the sense that they do not contribute to the sub-test or possibly measure another trait contrary to the trait under exploration (Barnard, 2004). The way in which good and poor items are identified is by means of fit or misfit. An explanatory note of the fitting or misfitting of items or persons is needed in order to provide background information on how to interpret fit and misfit. In Rasch analysis, fit is not interpreted in the same way as in the world of measurement where one would state that the model fits the data. Rather, fit statistics are used to detect discrepancies between the Rasch model prescriptions and the data (Bond & Fox, 2001). Thus when one speaks of misfitting persons this is the degree to which the response pattern of the individual is more haphazard than the Rasch model would have expected and therefore would be unexpected. The unexpected response pattern could indicate more or less variation than expected. The aim is to ensure the Rasch model expectations are met in the data, especially as it is only possible to add the equal intervals measures together if the specifications have been met (Bond & Fox, 2001).

7.3.1.1 Vocabulary sub-test

Forty items were included in the initial analysis (for details on item level refer to Appendix L). For both the persons and items the INFIT and OUTFIT mean squares (MNSQ) are close to 1 (refer to table 7.4 for details). The mean square statistics are used to check the compatibility of the data with the model (Bond & Fox, 2001). The person separation reliability is .83, which indicates that the scale does discriminate between persons while the item separation reliability is 0.99 indicating that the items do create a well-defined variable.

Table 7.4 ***Initial statistics for vocabulary sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	19.2	8.5	0.99	0.98	0.83
Item	378.6	165.2	1.01	1.02	0.99

Of the forty items included in the analysis, nine items misfitted (almost 25%), viz. items 1, 3, 4, 17, 20, 27, 36, 38 and 39. This evaluation is based on cut-off points for OUTFIT or INFIT mean squares (MNSQ) of 0.7 to 1.3 as stated earlier (see Appendix L for the WINSTEPS output). The items could be misfitting due to unusual response patterns across all persons. Thus the items could be flawed; they may not tap the same ability as the other items in the sub-test or they may be biased in terms of gender or subgroups (Barnard, 2004). Misfitting persons were also identified. Of the 794 persons, seventy-two persons were identified as misfitting. Thus these persons did not meet the specifications of the Rasch model as explained in the beginning of the section and were removed (Bond & Fox, 2001). The items were removed due to unexpected responses or irregular test taking behaviour (Barnard, 2004) that could be attributed to guessing. Furthermore, it could be that this inconsistency with an otherwise well-fitting model may indicate a failure to provide an appropriate measure for the ability of the person (Barnard, 2004).

The analysis was undertaken again, this time without the seventy-two persons identified as misfitting (refer to Table 7.5). Once again the INFIT and OUTFIT mean squares (MNSQ) are close to 1 for both persons and items, which indicated that the data does fit the model relatively well. The separation reliabilities for persons and items are 0.83 and 0.99 respectively indicating both adequate discrimination between persons and a well-defined construct.

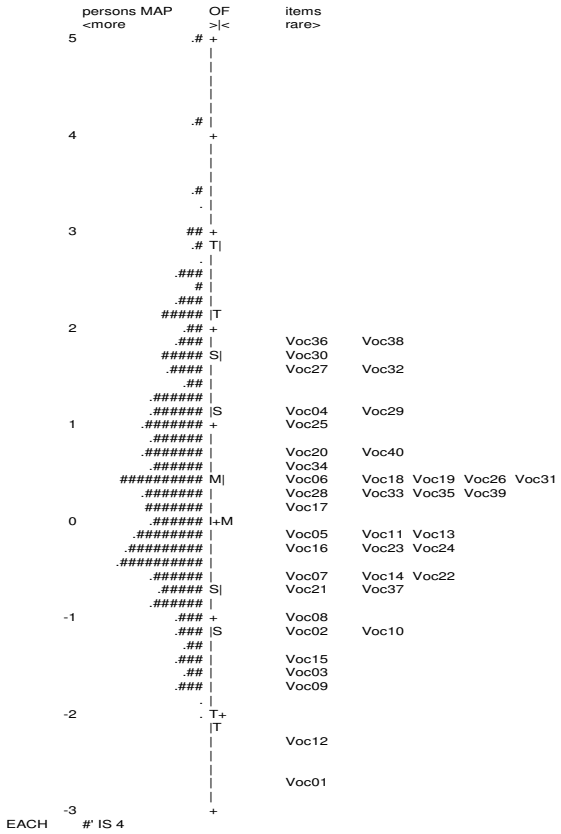
Table 7.5 ***Final statistics for vocabulary sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	19.6	8.5	1.00	0.96	0.83
Item	350.7	154.8	1.01	0.99	0.99

Once persons have been removed

Of the forty items included, only five items remained problematic namely items 17, 20, 27, 38 and 39. The reasons for this could be due to some form of bias in the items in terms of

gender or subgroups, an inability to tap the same ability level as the other items or the item may be flawed in some way (Barnard, 2004). Upon inspection of the items, it was found that the item stem, or list from which to identify the synonym, contained words which are not used in everyday English, such as pester, caress, and resent. In addition, it is possible that due to the words being placed in context, participants tried to make sense of the sentence by substituting alternatives that meant the same or nearly the same according to them. A number of alternatives seemed plausible. For example, item 39 states "I **resent** my parents for not allowing me to stay out after 10 o'clock at night". Participants had to identify a word meaning the same or nearly the same as the word **resent**. The options included fear, hate, jealousy or grudge. Here two options seem plausible, namely fear or hate. Based on the results it was suggested that the misfitting items be excluded from further analysis and that they should be revised or replaced for further versions of the assessment.



Each # indicates participating persons or learners; "M" marker represents the location of the mean; "S" marker represents one sample standard deviation away from the mean; "T" marker indicates two sample deviations away from the mean.

Figure 7.2 Item and person map for the vocabulary sub-test

Figure 7.2 provides an overview of the items included in the analysis and the learners participating on the same scale. On the right-hand side of the continuum are the items, with the persons displayed on the left. Ideally, the persons should form a standard normal curve, as one would expect persons of high and low ability to be at the ends, but the majority of the

persons in the middle. Clearly, in Figure 7.2 clusters of persons at the top, middle and bottom end of the scale can be identified. The item map does not include the misfitting persons but does include all of the items. This provides a visual display of items and persons with the most able persons and more difficult items located at the top of the map e.g. Voc 36 and 38, while items toward the end (negative logits) of the scale indicate that the item is easy (e.g. Voc 01) and persons or participating learners toward the bottom of the scale have less estimated ability. The figure illustrates that the items cluster well and range from easy to moderately difficult. What is of concern is that the sub-test seems too easy for a group of participants (approximately 80) and thus there may be a ceiling effect. It is suggested that the five items that do not fit should be rewritten to target participants with greater ability. Furthermore, two items, viz. item 1 and item 12, were very easy. Item 1 is "The teacher was **cross** with the class for not doing their homework". Although a good test design should have items which range through easy, moderate and difficult, items that are too easy should be avoided. These two easy items are not well targeted, as they are too easy. It is suggested that perhaps these two items should be replaced. However, even though it is suggested that the misfitting items be replaced, cognisance is given to the content-related validity and specifically the curriculum validity of the sub-test. Any item that is to be replaced should be replaced with the specifications of content-related validity in mind. This will be elaborated on further in Chapter 9.

7.3.1.2 Mathematics sub-test

For the analysis pertaining to the mathematics, sub-test seventy-four items were included. For both person and item the INFIT and OUTFIT mean squares (MNSQ) are close to 1 indicating a good fit and lack of noise (see Table 7.6). The separation reliabilities for both persons and items are high 0.89 and 1.00 respectively indicating there is sufficient discrimination between persons and that the items do form a well-defined construct.

Table 7.6 ***Initial statistics for mathematics sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	27.5	10.5	1.00	1.02	0.89
Item	295.1	239.5	1.01	1.13	1.00

The initial analysis undertaken indicated that of the seventy-four items included, twenty-four misfitted (this equates to approximately one third) possibly due to an inability to tap the same ability level as the other items or some form of bias (as was discussed earlier). The majority

of these items were located at the beginning and end of the sub-test (see Appendix L for details). One hundred and two persons included in the initial analysis misfitted (one out of seven persons), and were identified as misfitting due to the unexpected response patterns of these individuals. As the specifications of the Rasch model have to be adhered to, the misfitting persons were eliminated from the analysis (Bond & Fox, 2001). The misfit could also be attributed to an inability to provide an appropriate measure for the ability of the persons (Barnard, 2004). Once these persons were removed, the analysis was undertaken again (see Table 7.7).

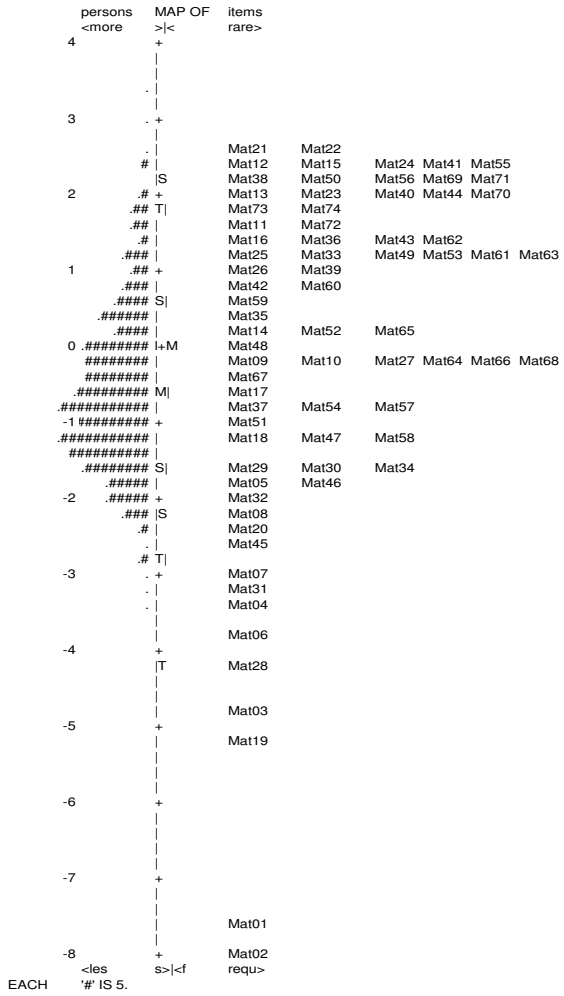
The INFIT and OUTFIT mean squares (MNSQ) are again close to 1, indicating relatively good fit between the theoretical model and the data. The separation reliability for persons and items is 0.89 and 0.99 respectively, indicating discrimination between persons and forming of a distinct construct.

Table 7.7 ***Final statistics for mathematics sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	26.6	10.4	1.00	0.97	0.89
Item	251.9	206.1	1.02	1.08	0.99

Once persons have been removed

Twenty-five items did not meet the stipulated criteria (OUTFIT or INFIT mean squares of 0.7 to 1.3). It was found even after the misfitting persons were removed, that the same items misfitted. The possibility exists that either the items are flawed in some way, unable to tap the same ability level of the other items or perhaps they are biased either in terms of gender or population group (Barnard, 2004). Upon inspection, it was found that the misfitting items included identification of the largest or smallest number, percentages, simple multiplication and division, fractions, area, co-ordinates and manipulation of three different sizes of cogs. These items were also located at the top of the item map (Figure 7.3) indicating that they were extremely hard for learners. Items which learners found easy contained simple addition sums, familiar shapes such as a star and sequences such as identifying which number was next (2, 4, 6, 8…). The sub-test items 21, 22 as well as 1, 2, 19, 3, 28, 6, 4, 31, 7 were not well targeted, as no person is located in the same position, on the item map, as these items. It is possible to replace these items with more appropriate items, ones covering a topic area that is underrepresented, perhaps a topic related to data handling. By including additional items for data handling, inferences related to the curriculum validity of the sub-test would be stronger. The issue of curriculum validity is addressed further in Chapter 9.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.3 Item and person map for the mathematics sub-test

7.3.1.3 Proof reading sub-test

The proof reading section has two components. The first component consists of a passage that participants had to read while identifying spelling or punctuation mistakes. The second component, however, asks participants to identify mistakes by comparing sentences. For the purposes of the analysis, these two components were kept separate and two analyses are presented in this section. Only the items containing actual mistakes that had to be identified were included. The reasoning behind including items with actual mistakes was that due to the coding procedure used when initially preparing for data capturing. All items selected by learners were given a “1” in the original coding procedure but would have been recoded as incorrect if no spelling or punctuation marks were present.

For the first section of the proof reading sub-test the INFIT and OUTFIT mean squares are 1 or close to 1. The OUTFIT mean square (MNSQ) for both persons and items are slightly over 1, indicating the possibility of slight "noise" (see Table 7.8). The separation reliabilities are high, 0.89 for persons and 0.99 for items.

Table 7.8 ***Initial statistics for proof reading 1 sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	19.7	10.0	1.00	1.14	0.89
Item	262.5	143.4	1.00	1.19	0.99

Fifty-eight items were included for analysis, eighteen misfitted due to inconsistent response patterns because of bias or inability to tap the same ability level as the other items (Barnard, 2004). Of the 794 persons, included in the initial analysis, 104 were identified as misfitting due to unexpected response patterns and were removed (Bond & Fox, 2001), also the misfit could be due to an inability to adequately attribute ability levels to individuals (Barnard, 2004). The analysis was undertaken again (refer to Table 7.9 and see Appendix L for details). The INFIT and OUTFIT mean squares (MNSQ) are close to 1, indicating relatively good fit between the data and the theoretical model. The fit statistics for the reanalysis is much the same as for the initial analysis (separation reliabilities for both items and persons are the same with 0.89 and 0.99 respectively).

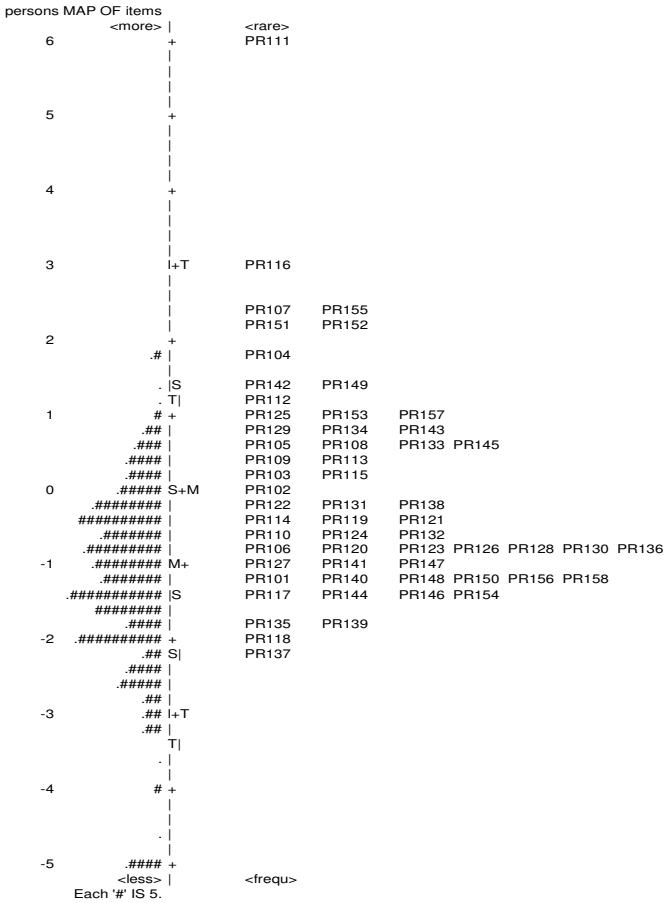
Table 7.9 ***Final statistics for proof reading 1 sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	19.7	10.1	1.00	1.04	0.89
Item	226.6	125.4	1.00	1.04	0.99

Once persons have been removed

Once the misfitting persons were removed, seventeen items misfitted (see Appendix L for details), possibly due to the reasons mentioned earlier (see Barnard, 2004). They included errors in punctuation such as a full stop and spelling errors e.g. "there" instead of "their", 'referr" instead of "refer", "lead" instead of "led". The items most difficult for learners (see Figure 7.4) were spelling errors, such as "than" and "then" and when to include commas. Learners found obvious spelling mistakes easier to identify. What is of concern is the large number of items that do not have persons located on the same logit (e.g. PR 116, PR 107, PR 155, PR 151, PR 152). This indicates that these items are not well targeted. It is

suggested that this section be shortened or the time allocated be extended. Perhaps the time factor is causing participants to overlook mistakes, although this in itself provides information that could be used for remedial purposes.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.4 **Item and person map for the proof reading 1 sub-test**

For the second section of the proof reading sub-test, participants had to identify mistakes by comparing a master list to a copy list. The INFIT mean square (MNSQ) for both persons and items is 1, while the OUTFIT mean square (MNSQ) for both persons and items is slightly lower than 1 (refer to Table 7.10). The separation reliability for persons and items are 0.90 and 0.98 respectively.

Table 7.10 ***Initial statistics for proof reading 2 sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	18.3	9.40	1.00	0.98	0.90
Item	361.6	73.9	1.00	0.98	0.98

Thirty-four items were included in the initial analysis, which resulted in fifteen items misfitting (almost 50%) possibly due to systemic inconsistencies in the form of bias or items that could have been flawed in some way (Barnard, 2004). Of the 794 persons included in the analysis, fifty-seven persons misfitted (see Appendix L for details) due to unexpected response patterns or an inability to attribute appropriate ability measures (Bond & Fox, 2001; Barnard, 2004). The analysis was repeated with the misfitting persons removed (refer to Table 7.11). The INFIT mean square (MNSQ) is similar to the initial analysis; however, the OUTFIT mean square (MNSQ) is slightly lower than the initial analysis with 0.97, this could indicate a slight lack of fit between the data and the theoretical model. The separation reliabilities for both persons and items are 0.91 and 0.98.

Table 7.11 ***Final statistics for proof reading 2 sub-test***

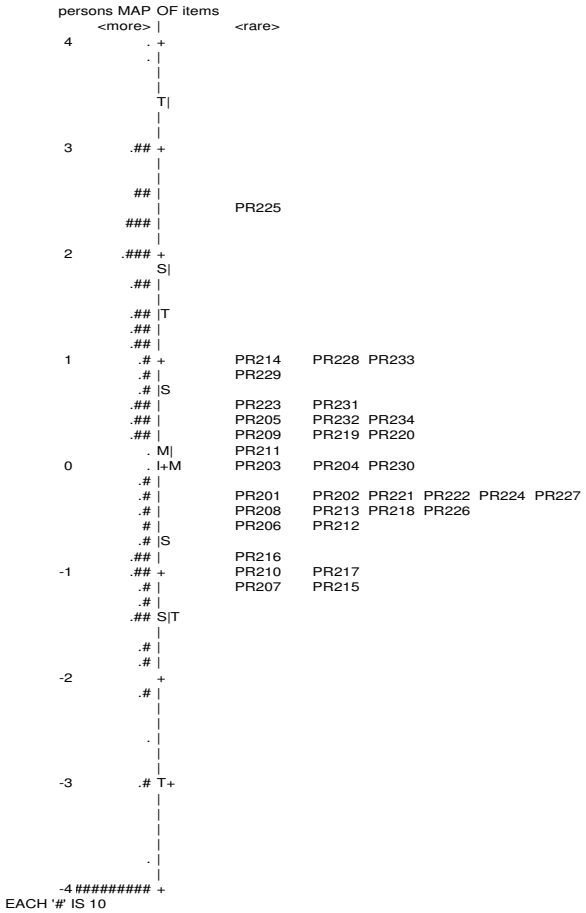
	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	18.2	9.40	1.00	0.97	0.91
Item	321.6	63.9	1.00	0.97	0.98

Once persons have been removed

Seventeen items misfitted in the reanalysis (see Appendix L for details), which was more than the original fifteen (exactly 50%). This misfit could be attributed to poor items or the item in itself may be good but does not form part of the set of items that collectively define the single measurement trait (Barnard, 2004). It is also important to note that every time the analysis is undertaken again, a new theoretical model is constructed and this could account for the discrepancy between the initial analysis and the reanalysis.

Upon inspection, it was found that the misfitting items included words in which letters were switched around or omitted when transferred from the master to the copy list, words like “Sandels” and “Sandles” or “Alexandra” and “Alexandria”. It appears, from the item map (see Figure 7.5), that a group of participants have ability measures that are higher than the most difficult item. These participants are located at the top of the item map. This could indicate a

ceiling effect. What is perhaps more disturbing is the small group of participants with ability levels which are lower than the items identified as easy. With more time allowed, fewer mistakes would perhaps be made or participants could attempt more items. For future versions of the assessment, more time should be allocated so that more persons can attempt the items. The Rasch model can make extrapolations to missing data based on performance on other items.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.5 Item and person map for the proof reading 2 sub-test

7.3.1.4 Perceptual speed and accuracy sub-test

The initial analysis for the perceptual speed and accuracy sub-test included twenty-six items. Both the INFIT and OUTFIT mean squares for persons and items are acceptable although INFIT and OUTFIT mean squares (MNSQ) for items (0.96 and 0.94 respectively) is slightly below 1, indicating slight lack of fit (see Table 7.12). What is cause for concern is the relatively low separation reliability for persons (0.67), an indication that discrimination between persons is not as desired. However, in this sub-test learners obtained more correct

responses than in any other (see the item map Figure 7.6). As a result of the similar learner abilities in this sub-test, it may prove difficult to identify distinct ability groups. The item separation reliability is 0.96 which indicates that the items do form a well-defined construct.

Table 7.12 ***Initial statistics for perceptual speed and accuracy sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	17.1	7.1	1.01	0.96	0.67
Item	406.7	79.1	0.96	0.94	0.96

The initial analysis revealed that eight of the twenty-six items misfitted (30% of the items); this could be due to these items measuring a different trait (Barnard, 2004). Of the 794 participants, fifty-seven misfitted (7% of the persons) due to unexpected response patterns (Bond & Fox, 2001). The analysis was undertaken again with the misfitting persons excluded (see Table 7.13). The INFIT mean square (MNSQ) for persons and items are 1.01 and 0.97 respectively, indicating fit between the data and the theoretical model. The OUTFIT mean square (MNSQ) for persons and items are 0.86 and 0.87, indicating a slight of lack of fit. The separation reliabilities are 0.66 and 0.96. That indicates lack of discrimination between learners but does suggest a distinct construct is present.

Table 7.13 ***Final statistics for perceptual speed and accuracy sub-test***

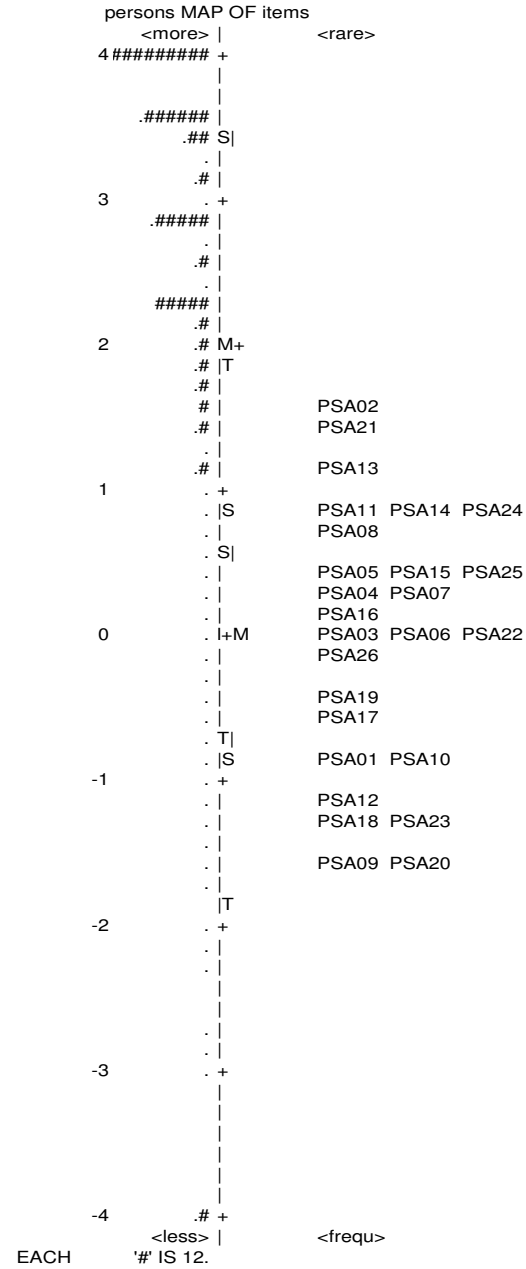
	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	17.5	6.9	1.01	0.86	0.66
Item	375.8	73.3	0.97	0.87	0.96

Once persons have been removed

Twenty-six items were included in this reanalysis (see Appendix L for details), of which nine misfitted (35% of the items). The items were identified as misfitting perhaps due to the items being flawed, or that they did not tap the same ability or perhaps systemic inconsistencies due to bias were present (Barnard, 2004). In this sub-test, participants visually compare and find matches between two columns. It is possible that some of the symbols included were unfamiliar to participants or were confusing, for example □©õù, <v^v, or ¢£ß.

The item map (Figure 7.6), indicated that the ability of most participants is higher than the most difficult item, denoting a ceiling effect. From this result, it appears that generally learners were able to access the items. There is however, a very small group of participants

with low ability. It is suggested that either more items are added or that the time allocations be adjusted so that learners have less time. This would increase the difficulty of this sub-test.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.6 **Item and person map for the perceptual speed and accuracy sub-test**

7.3.1.5 Cross-sections sub-test

Sixteen items are included in the cross-sections sub-test of the assessment. As with the other sub-tests, the INFIT and OUTFIT mean squares (MNSQ) are close to 1, indicating good fit and lack of noise (see Table 7.14). The separation reliability for items is excellent, 0.99, indicating a well-defined construct. The separation reliability for persons, however, is relatively low at 0.54, especially in comparison with other sub-tests, which indicates that the discrimination between persons is not what it should be. Learners did not fare well in this particular sub-test and it is likely that clearly defined ability groups would be difficult to identify.

Table 7.14 ***Initial statistics for cross-sections sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	7.5	2.8	0.99	1.09	0.54
Item	360.2	157.4	0.99	1.09	0.99

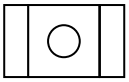
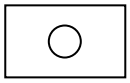
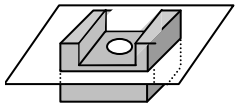
In the cross-sections sub-test, participants are requested to identify the 2D shape that would result if a 3D shape were cut through. Of the sixteen items, only two misfitted (see Appendix L for details) which could indicate a flaw in the items (Barnard, 2004). Eighty-one persons misfitted, due to unexpected response patterns because of either too little variation or too much variation in responses (Bond & Fox, 2001). The analysis was undertaken again without these persons (refer to Table 7.15). The INFIT and OUTFIT mean squares (MNSQ) are close to 1, indicating good fit between the theoretical model and the data. The separation reliabilities for persons and items are 0.50 and 0.99 respectively, indicating lack of discrimination between participants even though the construct itself appears sound. It is possible that the person separation is affected, as 10% of the total sample was removed due to unexpected response patterns. However, these persons had to be removed, as they did not adhere to the specifications of the model.

Table 7.15 ***Final statistics for cross-sections sub-test***

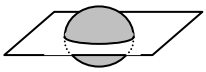
	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	7.5	2.87	1.01	0.98	0.50
Item	323.6	147.5	0.99	0.98	0.99

Once persons have been removed

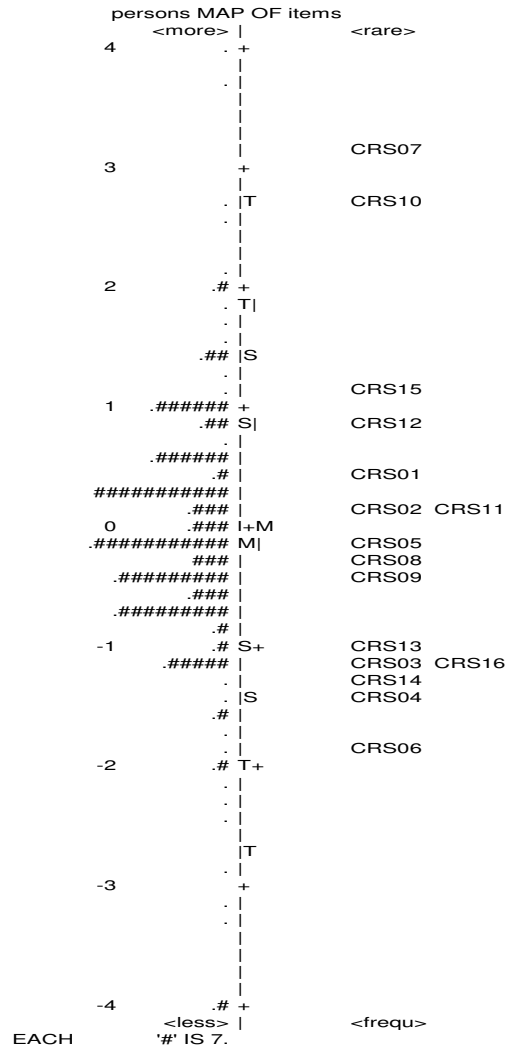
The results of the reanalysis indicated that three items misfitted, namely 7, 10 and 12. Items 10 and 12 are very similar in nature and could have been easily confused as the instruction was to find the shape which was used to create the cross section:

Item 10  Item 12  Associated with 

The 2D shape for item 7 has no match but learners could have selected another option that is slightly smaller or slightly bigger than the shape in item 7:

Item 7  Associated with 

A very small group of participants' ability measures exceeded items 15 and 12 (see Figure 7.7). However, it would appear as if some learners could not access seven of the sixteen items.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.7 Item and person map for the cross-sections sub-test

7.3.1.6 Block counting sub-test

The block counting sub-test consists of twenty items in which participants have to identify the number of small blocks and the number of large blocks in the figure presented. Participants were also requested to identify the minimum number and maximum number of small blocks in the figure presented. The INFIT mean square (MNSQ) for both persons and items are below 1, indicating a slight lack of fit. The OUTFIT means square (MNSQ) values for both persons and items are well above 1, indicating noise within the data (see Table 7.16). The person separation reliability is 0.74 and the items reparation reliability is 1.00. Both values are acceptable.

Table 7.16 ***Initial statistics for block counting sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	9.2	3.5	0.93	1.59	0.74
Item	355.2	246.8	0.94	2.70	1.00

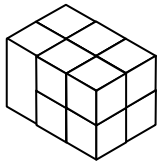
Of the twenty items, fifteen items initially misfitted (75% of the items, which is very high). This misfit could be due to the items being flawed in some way, that the items do not tap the ability as the other items, that a different trait is measured by theses items or that there might be bias in the items. Misfitting persons were identified (see Appendix L for details) and were removed from the analysis, as these individuals did not comply with the specifications of the Rasch model (Bond & Fox, 2001). It is also possible that though the model in itself seems to be functioning relatively well, an appropriate measure of the relevant ability could not be provided (Barnard, 2004). The analysis was undertaken again (refer to Table 7.17). With the reanalysis, the INFIT mean squares (MNSQ) and OUTFIT mean squares (MNSQ) were around 1, indicating fit between the data and the theoretical model.

Table 7.17 ***Final statistics for block counting sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	9.6	3.2	0.99	1.08	0.73
Item	288.4	205.0	0.98	1.10	1.00

Once persons have been removed

Of the 20 items, eight misfitted, which is 40% of the sub-test, and this is substantially better than the initial 75% of the items. Four of the items referred to the minimum (two items) and maximum (two items) number of small blocks possible.



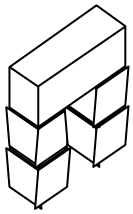
MINIMUM number of small blocks possible:

0	1	2	3	4	5	6	7	8	9	10

MAXIMUM number of small blocks possible:

0	1	2	3	4	5	6	7	8	9	10

The remaining four items referred to the number of small blocks (2 items) and number of larger blocks (2 items). For example:



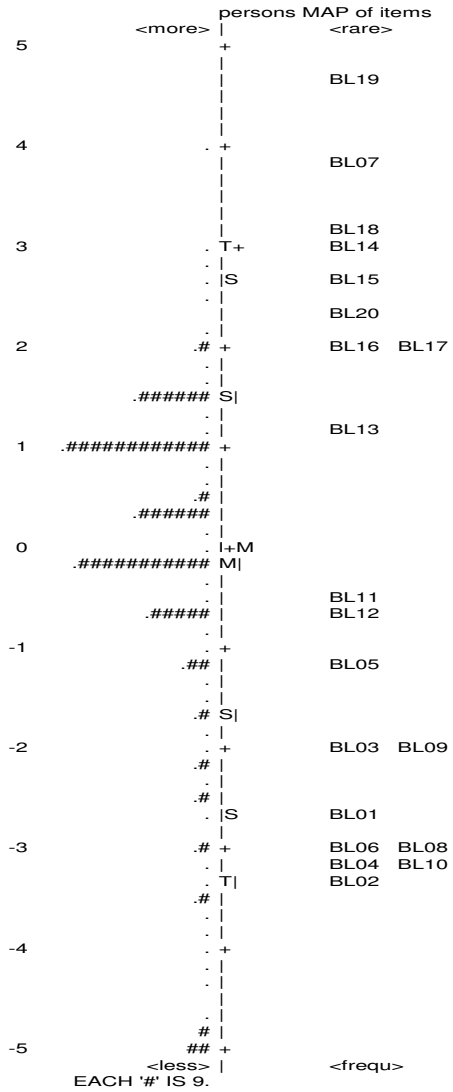
Number of SMALL blocks

0	1	2	3	4	5	6	7	8	9

Number of LARGE blocks

0	1	2	3	4	5	6	7	8	9

From the item map, (see Figure 7.8) it is clear that there is a group of items (6 in total) which participants were unable to access due to the difficultly of the items – an example is item 19. A small group of participants had a fifty-fifty chance of answering the next two difficult items correctly. Of the 20 items, only 12 items were accessible to participants (see Appendix L for details). On the map, a clear cluster of exceptionally difficult items can be identified at the top. These items are not well targeted. It is suggested that the items be re-evaluated. It is also possible that participants were fatigued at this stage of the assessment or did not understand the instructions clearly.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.8 **Item and person map for the block counting sub-test**

7.3.1.7 Pictures sub-test

The pictures sub-test consists of three sections, namely adding pictures, subtracting pictures and picture sequences. There are 18 items in total, 6 items per section. The INFIT mean square for both persons and items is 0.99, is very close to 1 (See Table 7.18). The OUTFIT mean square is 1.21 and 1.28 for persons and items respectively is slightly elevated, indicating some noise in the data. The person separation reliability is 0.73 and the item separation reliability is 1.00. For both the person and item, the separation reliability is acceptable, as discussed by way of introduction in the beginning of 7.3.1. The value of 0.73 is an indication of discrimination between persons, although lower than some of the other

sub-tests. The separation reliability for items (of 1.00) is an indication of a well-defined construct.

Table 7.18 ***Initial statistics for pictures sub-test***

	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	8.1	3.7	0.99	1.21	0.73
Item	349.6	201.1	0.99	1.28	1.00

Of the 18 items included in the initial analysis, eight misfitted (see Appendix L for details), this could be due to these items measuring another trait, the items themselves may be flawed or there may be bias in some way (Barnard, 2004). Four of the eight items are in the subtracting pictures section. One hundred and forty one persons misfitted because of unexpected response patterns, as explained earlier (Bond & Fox, 2001). It is possible that the learners did not listen to the instructions given, as a result did not answer the items correctly. The analysis was undertaken again after the misfitting persons had been removed. The INFIT mean square (MNSQ) for both persons and items is 0.99 while the OUTFIT mean square (MNSQ) for both persons and items is 1.06 (see Table 7.19). The separation reliability for persons and items is 0.76 and 1.00 respectively. That indicates discrimination between participants and a clearly defined construct.

Table 7.19 ***Final statistics for pictures sub-test***

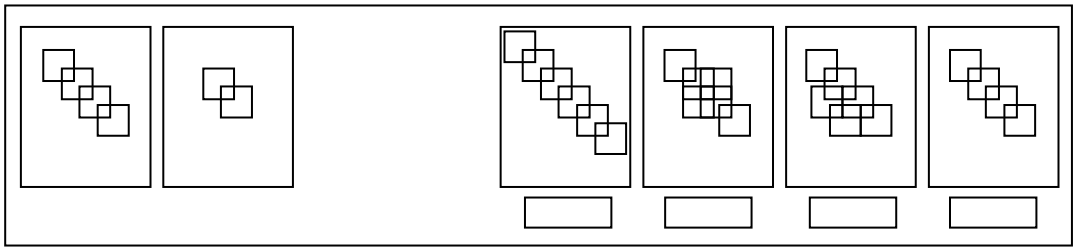
	Mean	S.D.	INFIT MNSQ	OUTFIT MNSQ	Separation reliability
Person	8.4	3.7	0.99	1.06	0.76
Item	305.1	176.8	0.99	1.06	1.00

Once persons have been removed

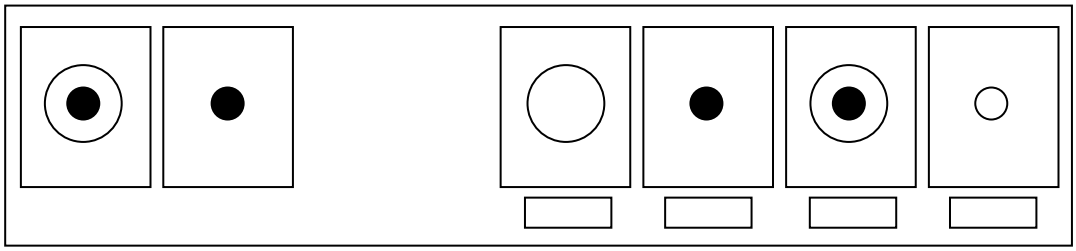
Nine items misfitted amounting to 50% of the sub-test (see Appendix L for details). The source of the misfit could be poor or flawed items, measuring another trait or bias in terms of gender or subgroups, in this case population (Barnard, 2004). Four items are located in the subtracting pictures sections; three items are located in the adding pictures section while the remaining two items are in the sequences section. The pictures sub-test was designed with the adding pictures first, followed by subtracting pictures and pictures sequences. It is possible that learners did not read the instructions at the top of the subtracting pictures, thus treating the section as adding instead of subtracting. Furthermore, this is the last sub-test in

the assessment and participant fatigue could have been a contributing factor. Examples of misfitting items are adding pictures, subtracting pictures and picture sequences (see below):

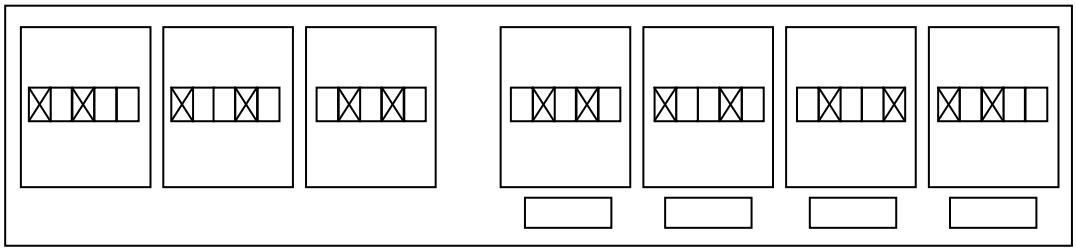
Adding Pictures



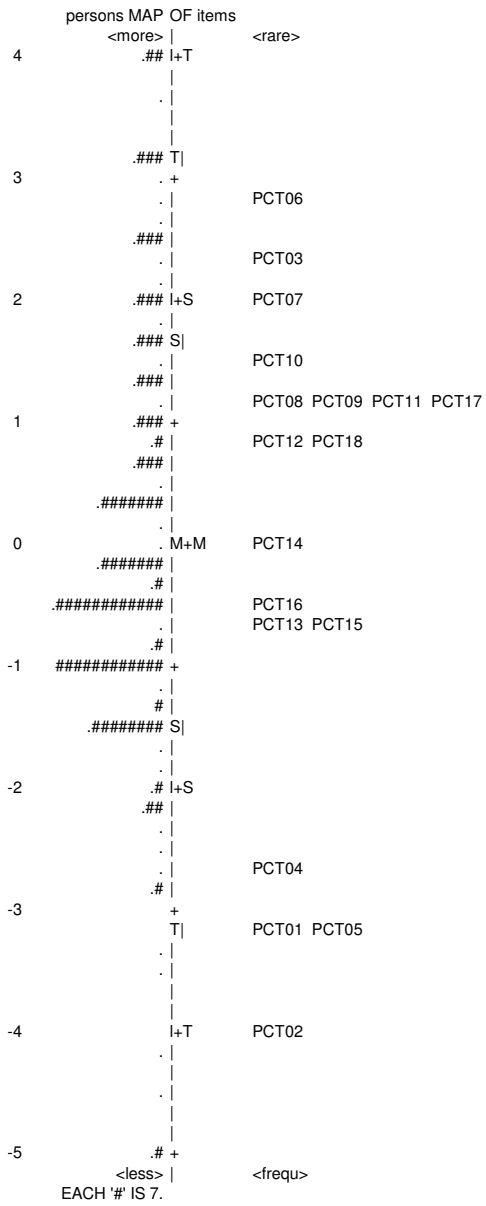
Subtracting Pictures



Picture Sequences



As with some of the other sub-tests, there is a group of participants with ability measures exceeding the difficulty of items (see Figure 7.9). This could possibly cause a ceiling effect. There is once again a small group of participants whose ability measure is very low but it appears as if the majority of the participants were able to access at least half of the items across the three sections.



Each # indicates participating persons or learners; “M” marker represents the location of the mean; “S” marker represents one sample standard deviation away from the mean; “T” marker indicates two sample deviations away from the mean.

Figure 7.9 **Item and person map for the pictures sub-test**

7.3.2 Conclusions drawn from the Rasch analyses

> In establishing the construct validity of a test the first step involves the definition and delineation of the meaning of the test variable (Kline, 2000, p. 37).

In Chapter 6 various definitions were provided for what the sub-tests common to aptitude and ability assessments measure. The result is that the sub-tests included in MidYIS do have an

empirical base. The question now remains to what extent the items included in the sub-tests measure the same concept.

The aim of the Rasch analysis was to identify the items included in the various sub-tests which were unidimensional or which measure the same construct or concept. This was the first step in finding answers for the sub-research question 1.2.4 ***how well do the items per sub-test function and do they form well-defined constructs.*** In the Rasch analysis the smallest unit was used, namely the items, which are included to form a set of items associated with the various sub-tests. The objective was to determine which core items best measure the theoretical concept underpinning the sub-test.

It is clear that there are a number of items per sub-test which are unidimensional and do measure the theoretical concept that they were intended to measure. However, there were items that misfitted. The source of the misfit could be attributed to (Barnard, 2004):

- Flawed or poorly written items;
- Items not measuring the same trait;
- Some form of bias in terms of gender or subgroups.

These items would need to be revised or if revision were not possible, additional items would have to be generated. However, items would have to be revised or rewritten with content-related validity in mind. The resulting misfitting items necessitate improving the sub-tests for the South African context not only from a curriculum perspective but also from a psychometric perspective. These “new” items would also need to complement the other items in the sub-test and for this reason it is suggested that any development work be undertaken in conjunction with a set assessment framework.

The question remains ***how well do the items per sub-test function and do they form well-defined constructs.*** What does emerge out of the Rasch analyses is that there are core items that can be included in sub-tests and these do form well-defined constructs. This result can be taken in conjunction with the results of evaluation of the items in terms of the domain they represent (as was described in Chapter 6). As a result of this analysis in conjunction with the results presented in Chapter 6, it is possible to suggest that not only do the sets of items cohere to form the constructs measured in each of the sub-tests but that the sub-tests themselves could be combined to form the scales as developed by the CEM centre.

This is in line with Messick (1981) who suggests that the relevance of a construct should be evaluated in light of a particular applied purpose. Here issues of content are associated with judgments of relevance, where relevance is seen as whether the sample of items under investigation can be aligned to the content domain. Sireci (1998) elaborates on content relevance to include the "congruence between the test content and the purpose of testing" (p. 99).

If it is said, that for reporting purposes it may be easier to combine sub-tests into scales, then the next step would be to ascertain whether there is any congruence between the underlying skills assessed by the seven sub-tests. If this line of thought is followed, then it is possible to combine the perceptual speed and accuracy with the proof reading sections as both sub-tests are designed to measure fluency and speed in finding patterns as well as spotting mistakes. So, theoretically this would be a sound argument to make.

The same line of reasoning can be used when considering whether block counting, cross-sections and pictures should be combined. Once again a set of common skills can be identified, namely that these sub-tests attempt to measure 2-D and 3-D visualisation, spatial aptitude, pattern recognition, and logical thinking. According to Anastasi and Urbina (1997), non-verbal assessments typically do not include language that participants have to read in order to answer items. Rather, pictures are used for this purpose. If this definition were used as an underpinning rationale, then it would make sense to combine these three sub-tests as pictures are used instead of written items, which have to be read.

In the words of Messick (1981, p. 11), "we must go beyond judgments of content consistency to an assessment of response consistency". In the section to follow, the consistency of responses is explored. This is done by means of reliability analysis, in which the theoretical argument that sets of items associated with sub-tests can be incorporated into the scales as identified by the CEM centre is empirically tested.

7.4 Exploring the reliability of the MidYIS assessment

> A test cannot correlate with anything more highly than it does itself…it is a peculiar measuring instrument if different parts of it are measuring different variables, as must be the case with low reliability…low internal consistency implies considerable error of measurement (Kline, 2000, p. 29).

Internal consistency or reliability refers to the consistency of scores, obtained by the same individuals completing the assessment on different occasions (Anastasi & Urbina, 1997). According to Krathwohl (1998), internal consistency is the degree to which all the items measure the same thing. It is to be expected that the measure will be affected only by the construct of interest and that the participants should respond the same way to similar items. As internal consistency reliability "reflects the extent to which each item is measuring the same variable" (Kline, 2000, p. 28), inferences about content-related validity of the assessment are strengthened (Suen, 1990). This form of reliability is also a prerequisite for construct validity (Kline, 1993).

Internal consistency was used to make inferences pertaining to the reliability of scores as was discussed in Chapter 5. Kuder-Richardson (KR-21) was used, which is a special form of Cronbach's alpha (Coolican, 1999). Reliabilities for assessment data should be high, preferably around 0.9, and should never drop below 0.7 (Kline, 1993). In the section to follow, the reliability analysis is presented. Core items identified by the Rasch analysis were used in this analysis.

The reliability coefficients for the MidYIS scales are provided in Table 7.20 and are based on the South African data. The reliabilities for all four scales are high (see Appendix L for details). Three of the four scales had reliability coefficients of 0.90 or higher, while the non-verbal scale had a reliability coefficient of 0.84. This indicates that in the South African sample of schools, the items for the various scales do seem to be measuring the same construct. This also provides an empirical basis for the theoretical extrapolation put forward in 7.3.2.

Table 7.20 ***Reliability analysis and standard error of measurement per scale***

Scale	N	Reliability coefficient	Standard error of measurement	Number of items
Vocabulary	794	0.90	2.42	35
Mathematics	794	0.92	2.45	48
Skills	794	0.94	3.72	77
Non-verbal	794	0.84	2.37	34
Total	794	0.97	5.58	194

An analysis per population group was also undertaken (see Table 7.21), as the context of South African schools can be vastly different. A similar pattern emerges from across the

population groups (see Appendix L for details). Most of the reliabilities obtained were above the 0.7 cutoff point, the only exception being for Indian learners on the mathematics scale (0.69). It is important to note that the analysis was undertaken per population group and not according to school type, for example previously advantaged and previously disadvantaged schools. The reasoning behind this is that there would be "previously disadvantaged" learners in "previously advantaged schools".

Table 7.21 ***Reliability analysis per scale and population groups of learners***

Scale	African	Coloured	White	Indian
Vocabulary	0.88	0.88	0.91	0.89
Mathematics	0.77	0.79	0.74	0.69
Skills	0.94	0.92	0.94	0.92
Non-verbal	0.81	0.75	0.84	0.87
Total	0.88	0.85	0.89	0.88

From the reliability analysis it is clear that the results seem consistent and that any inferences made based on the items included in the analysis can be made with confidence. The exception is perhaps the mathematics scale for Indian learners. However, Indian learners constituted the smallest group (6%) as was mentioned in the beginning of the chapter. As sample size could be one of the causes for the result, it is recommended that future analysis be undertaken with a larger sample.

Thus in answer to the question ***to what extent are the results obtained on MidYIS reliable*** it would appear from the overall analysis that the results on the reduced number of items are consistent and that each scale reliably measures the underlying construct.

7.5 Exploring relationships between MidYIS scores and academic achievement

> Even academic success which would appear to be clearly related to intelligence is affected by other factors: the skill of the teachers, the peer group of the children, the family circumstances and the health of the child…Thus a modest but positive correlation would be acceptable as evidence of predictive validity (Kline, 2000, p. 33).

In Chapter 5, correlation analysis was discussed. In this case, the aim of the correlation analyses was to establish whether relationships exist between the MidYIS scores and academic achievement, specifically language and mathematics achievement (see Appendix

L for details). This is the first step toward determining whether MidYIS would be able to predict future achievement of South African learners. Correlation analyses was undertaken using the MidYIS scale scores, resulting from the Rasch analyses, and English and mathematics final marks as received from the schools. Of the 11 schools that participated in the study, nine schools provided information pertaining to the final year results in English and mathematics of the learners who participated in the study. Although repeated attempts were made to obtain results from all the participating schools, two schools did not feel comfortable providing the information. The final marks obtained from the schools comprised a combination of a continuous assessment mark and a final examination mark. The MidYIS scale scores and the English and mathematics marks were also explored in order to ascertain whether any assumptions underlying correlation analysis was not violated.

Table 7.22 details the results of the relationships between the various MidYIS scales and the mathematics results obtained from schools while Table 7.23 provides the results of the analysis for English (refer to Appendix L). All of the MidYIS scales were included in the analysis and not just scales directly relevant to mathematics and English. The reason behind including all the scales in both analyses is the interrelated nature of the skills assessed. In mathematics for example, language proficiency is an important criterion for success (see Howie, 2002).

Correlations of above 0.3 (Kline 1993) for the MidYIS scales and the mathematics and English marks are considered indicative of a positive relationship, but, in addition to the positive correlations, the variance explained also has to be considered. The variance explained is calculated by squaring the correlation (Kline, 2000).

Table 7.22 ***Correlations between the revised MidYIS scales and school mathematics***

School	Vocabulary		Mathematics		Skills		Non-verbal		Total	
	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained
School 1	0.587**	35	0.731**	53	0.608**	37	0.477**	23	0.726**	53
School 2	0.508**	26	0.447**	20	0.592**	35	0.450**	20	0.605**	37
School 3	0.574**	33	0.696**	48	0.347**	12	0.273*	7	0.620**	38
School 4	0.589**	35	0.724**	52	0.460**	21	0.458**	21	0.677**	46
School 5	0.201	4	0.476**	23	0.250*	6	0.303**	9	0.388**	15
School 6	0.294*	9	0.446**	20	0.162	3	0.258*	7	0.352**	12
School 7	0.561**	32	0.604**	36	0.634**	40	0.422**	18	0.695**	48
School 8	0.403**	16	0.449**	20	0.411**	17	0.375**	14	0.540**	29
School 9	0.262*	7	0.193	4	0.317**	10	0.515**	27	0.441**	20

*** Significant at the 0.01 level*
** Significant at the 0.05 level*
Grey = Former Department of Education and Training
White = Former Model C Schools
Yellow = Former House of Delegates
Green = Former House of Representatives

From the results, it seems as if positive relationships exist between the MidYIS scales and the mathematics marks (see Table 7.22). The exception would be for vocabulary in which weak relationships exist for schools 5, 6 and 9. This may be explained by the difference in mathematics and vocabulary (language) as well as that very often learners are more proficient in one than the other. It is possible that language could be a factor in addition to the nature of marks received from the school, as these are not standardised results. Interestingly the mathematics scale does not correlate with the mathematics mark for school 8, but as can be expected, high correlations can be found between mathematics and the MidYIS mathematics scale for the other schools. The non-verbal scale presents interesting results. The non-verbal scale includes 2D and 3D shapes that have to be manipulated. The ability to use 2D and 3D shapes cannot be underestimated as this forms the basis for geometry. However, in schools 3 and 6 the correlation between non-verbal and mathematics is less than 0.3. Two schools obtained results lower than 0.3 for the skills scale, namely school 5 and school 6.

What is noteworthy is the percentage of variance that MidYIS explains in terms of mathematics academic achievement. For the former Model C schools the percentage of variance explained ranges from 26% to 33% on the vocabulary scale. However, percentages as low as 4% (school 5), 7% (school 9) and 9% (school 6) are recorded. A similar result is obtained for the skills and the non-verbal scales as with the vocabulary scale. The percentage of variance explained in terms of academic success for mathematics is better than the other scales. However, in school 9 as little as 4% of the variance can be accounted for. This means that abilities alone explain little in school's variation in terms of performance, even though the scales can be related to the domain of mathematics. Thus other factors possibly on a learner, classroom or school-level must be considered, for instance, language spoken in the home of the learner, age of the learner, socio-economic status of the learner, gender of the learner or educator, language of teaching and learning, teaching style of the educator or principal management style.

Table 7.23 ***Correlations between the revised MidYIS scales and school English***

School	Vocabulary		Mathematics		Skills		Non-verbal		Total	
	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained	Correlation	% Variance explained
School 1	0.756**	57	0.734**	54	0.684**	47	0.505**	26	0.812**	66
School 2	0.754**	57	0.503**	25	0.714**	51	0.559**	31	0.766**	59
School 3	0.642**	41	0.665**	44	0.586**	34	0.234	5	0.754**	57
School 4	0.758**	57	0.685**	47	0.519**	27	0.429**	18	0.732**	54
School 5	0.313**	10	0.380**	14	0.188	4	0.302**	9	0.353**	13
School 6	0.564**	32	0.610**	37	0.312*	10	0.321*	10	0.561**	32
School 7	0.764**	58	0.596**	36	0.661**	44	0.445**	20	0.771**	59
School 8	0.525**	28	0.496**	25	0.482**	23	0.386**	15	0.625**	39
School 9	0.429**	18	0.563**	32	0.287**	8	0.010	0	0.449**	20

*** Significant at the 0.01 level*
** Significant at the 0.05 level*
Grey = Former Department of Education and Training
White = Former Model C Schools
Yellow = Former House of Delegates
Green = Former House of Representatives

As might have been expected, the correlations between vocabulary and the English mark in most of the schools exceeded 0.3 and were significant at the 0.01 level (see Table 7.23). Strong correlations were found between the mathematics scale and the English marks. Less substantial correlations were found between non-verbal and the English mark with weak correlations for school 3 and 9. One might have expected a slightly higher correlation between the skills scale and the English mark as proof reading is language-bound. Although the majority of the correlations for the schools were above 0.3, the correlations were lower than the correlation between vocabulary and the English mark. A very weak relationship was found between skills and English for school 5. Of the four scales, non-verbal had the lowest correlations; this in itself is perhaps not surprising as non-verbal scales should not be as language bound as some of the other scales, vocabulary for example.

The emerging picture for the English marks and MidYIS in terms of the percentage of variance explained is similar to the one for mathematics marks and MidYIS. For vocabulary a large percentage of variance can be explained up to 57% (school 1 and 2) in some schools and as low as 10% in other schools (school 5). For Mathematics up to 54% (school 1) and as low as 14% (school 5) can be explained, a similar picture emerges for skills and non-verbal.

It is suggested that MidYIS could be used for prediction purposes; in answer to the question ***to what extent does the data predict future achievement?*** This was an initial first step in order to ascertain whether the MidYIS assessment could be used for prediction purposes. However, further analysis is needed with a larger sample (including rural schools and schools from other provinces) using a standardised school-based examination before definite inferences related to predictive validity can be made. Performance on its own can only account for so much variance. Other factors have to be considered as the quote in the beginning of the section suggests. It is proposed that a multilevel model be used in which other factors can be included in addition to ordinary least squares models which can be used to determine the value the school has added. It is important to consider that the results for mathematics and English are not standardised across the schools but rather a reflection of the assessment within the school. This could partly explain the fluctuations in correlations and in the percentage of variance. As a result, the exploration of predictive validity should be undertaken again, using standardised school scores.

7.6 Conclusion

> …validity of a test is not clear-cut, as was [is} the case with reliability. There is no single validity coefficient (Kline, 2000, p. 38).

The aim of Chapter 6 and of this chapter was to address the specific research question ***how valid and reliable the data generated by the MidYIS monitoring system are for South Africa?*** Different strategies for making inferences related to validity were presented, ranging from conceptual considerations as in the case of content-related validity (presented in Chapter 6) to empirical considerations as in the case of construct-related validity and predictive validity (presented in this chapter).

Three sub-questions were addressed in this chapter; two questions are associated with construct and predictive validity while the other is related to reliability. The sub-questions addressed in this chapter are (see Chapter 5 for a detailed discussion):

1.2.1 To what extent are the results obtained on MidYIS reliable?

1.2.4 How well do the items per sub-test function and do they form well-defined constructs?

1.2.5 To what extent does the data predict future achievement?

Sub-question 1.2.4 ***how well do the items per sub-test function and do they form well-defined constructs*** was addressed by means of item (Rasch analysis) and scale analysis (reliability analysis). What emerges from the Rasch analyses is that there are core items associated with sub-tests and that the sub-tests can be combined into scales, as was originally designed by CEM. However, there are items that seem to measure constructs other than the constructs they were designed to measure, and these were removed from further analyses. Thus the items which were identified, as misfitting should be revised or rewritten, based on an assessment framework for the assessment as a whole. The assessment framework should be developed both from a curriculum and psychometric perspective, thus satisfying conditions for conceptual forms of validity.

Sub-question 1.2.1 is related to the reliability of the MidYIS results (***to what extent are the results obtained on MidYIS reliable?***). The analyses were undertaken with the whole sample in addition to the different population groups. The results of the analyses indicate internal consistency of the set of items per scale and as a result, the items per scale seem to be measuring the same construct. It is suggested that in future larger samples for sub-

population groups should be included if inferences per population group are to be made with confidence.

The third sub-question addressed is related to the predictive validity of the assessment (***to what extent does the data predict future achievement?).*** The analysis was undertaken per school and not across schools. The results indicated that the scales do correlate with the results obtained from schools for mathematics and English. Therefore, MidYIS could possibly be used for prediction purposes, although more analytic work is needed in this area before definitive statements can be made (including a larger sample from other provinces and contexts). What does seem to emerge is that MidYIS on its own can only account for so much variation in performance. Other factors on the learner, classroom and school-level have to be taken into account. Thus a multilevel model should be used in addition to ordinary least squares models that can be used to determine the value the school has added. For trustworthy inferences to be made in terms of predictive validity, standardised academic results should be used, such as the Grade 9 exit-level examinations.

To conclude, Sicoly (2002, p. 174) encapsulates the aim of the first specific research question (***how valid and reliable the data generated by the MidYIS monitoring system are for South Africa***):

> Assessment results are expected to improve student performance by improving educational practices. The feedback provided by assessment results may be used to guide school wide planning, to adjust teaching practices, and to focus staff development efforts. If schools are to use assessment data as a basis for planning and decision making, we must satisfy the highest standards. Poor quality assessment results will only lead to misdirection and confusion instead of providing an opportunity for improving schools effectiveness.