

CHAPTER 3

VALIDITY AND RELIABILITY

3 VALIDITY AND RELIABILITY

3.1 INTRODUCTION

In Chapter 2, the study's aims of exploring how objects can influence the level of construct validity of a Picture Vocabulary Test were discussed, and a review conducted of the literature on the various factors that play a role as to how the validity level can be influenced. In this chapter validity and reliability are discussed. The reason why they are discussed in a chapter of their own is to provide a better understanding of what their role was in this study.

3.2 VALIDITY

Although there is much debating going on about validity and an entire thesis can be written about it only the relevant aspect to this study are mentioned and discussed in the sections to follow.

3.2.1 Validity

Bond (2003, p. 179) comments that:

...validity is foremost on the mind of those developing measures and that genuine scientific measurement is foremost in the minds of those who seek valid outcomes from assessment.

From this above quote, validity can be seen as the core of any form of assessment that is trustworthy and accurate (Bond, 2003, p. 179). Validity, according to Messick (1989, p. 6)

... always refers to the degree to which empirical evidences and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores.

Borsboom, Mellenbergh and van Heerden (2004, p. 1061) have a different take on validity stating that: "...a test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement". They do not agree with Messick's conception of validity. In this study Messick's viewpoint will be followed.

Forming the crux of this research project, not only is validity an essential issue for assessment but for measurement as a whole. In addition, the assessments can be used across countries and cultures, but if this is not the case, assessments can be seen as being biased. What is more, validity influences the way that instruction changes once the results of an assessment have been correctly interpreted (Gregory, 2000; Linn, 1998; Mahoney, 2008; Messick, 1989; Popham, 2003; Frederiksen & Collins, 1989; Gay & Airasian, 2003; Sullivan, Karlsson & Ware, 1995).

For instance, an Intelligent Quotient (IQ) test measures the intelligence of the learner (existing attribute), and not all learners will have the same intelligence (variations in the attribute). This implies that when a specific attribute needs to be investigated, the interpretations or inferences made from the test have to be valid. As Ukrainetz and Blomquist (2002, p. 60) put it: "...how well a test measures what it is purported to measure".

Validity is an evolving complex concept because it relates to the inferences regarding assessment results. Focusing on the consequences of the inferences made implies that they should be appropriate and adequate. Messick (1989, p. 6) points out that inferences are hypotheses, and when these inferences are validated it amounts to hypothesis-testing. As a result, validity is seen as evaluative judgements that are made on the inferences of assessment results or test scores, that is whether correct interpretations are made and actions are taken based on the inferences. These evaluative judgements need to be correct and reflective of the truth. An assessment

or test cannot be said to be valid, only the inferences about the test. In Messick's (1989, p. 5) own words: "...what is to be validated is not the test or observation device but the inferences derived from the test scores or other indicators..." An inference can be seen as the interpretation made by a person about a test or assessments results and for this reason it would be incorrect to say that a test is valid since only inferences about the test can be valid or not. In all research, the phenomena being researched must be accurately described through the findings, but if this does not happen then the level of validity is questioned (Gregory, 2000; Mahoney, 2008; Messick, 1989; Graziano & Raulin, 2000).

On the whole, validity is seen as a unitary concept. An example would be if various researchers had to examine one specific research study and also come up with the same conclusion, then the research study would be internally valid. Conversely, with external validity the results and conclusions can be generalised to other situations or with other subjects. Two different types of validity were portrayed in the above example but many other forms of validity exist, making validity a unitary concept (Howell 2002; Opie, 2005; Cohen, Manion & Morrison, 2000; McMillan & Schumacher; 2006). Validity cannot be adequately summarized by a numerical value but rather as a "matter of degree", as stated by Linn and Gronlund (2000, p. 75). The validity of assessment results can be seen as high, medium or low, or ranging from weak to strong (Gregory, 2000).

To summarise, validity refers to the appropriateness of the inferences made about the results of an assessment. Inferences being "...conclusions derived from empirical evidence bearing on score meaning..." (Messick, 1989, p. 6). Secondly, validity is a matter of degree and not a specific value. Thirdly, validity is applied to a specific purpose or use and therefore is not valid for all purposes. Fourthly, validity is seen as a unitary concept, meaning that there are a number of different types of validity. Lastly, validity is concerned with an evaluative judgment about an assessment (Gregory, 2000, p. 75). Of all the different types of validity that exist, construct validity is seen as the most important form. Construct validity forms the basis for any other type of validity and from a scientific point of view is seen as the whole of validity (Mislevy, 2007).

Content-related validity is also another type of validity. As its name implies it explores how the content of the assessment performs. In order to determine content-related validity the researcher is concerned with determining whether all areas or domains are appropriately covered within the assessment. Furthermore, it deals with how the assessment is designed, for example the size of the font, sufficiency of work space for learners, correct language usage and clarity of instructions (Fraenkel & Wallen, 2003). With the original PIPS instrument the content-related validity was carefully considered and of a high standard.

3.2.2 Construct Validity

In 1955, Cronbach and Meehl elaborated on the various methods to determine construct validity. For a construct to be scientifically acceptable it had to be located in a 'nomological network', made up of laws that are either statistical or deterministic. These laws tie observable properties to one another, in other words the same topics or constructs are grouped together (Garrison, 1994; Moss, 1992; Cronbach & Meehl, 1955).

The main purpose of a researcher by exploring construct validity is to determine whether the inferences made about the results of the assessment are meaningful and serve the purpose of the assessment. Construct validity is empirically explored by means of Rasch and, as mentioned above, is central to any quality assessment. Whenever a certain attribute has to be measured, construct validity is involved, as it is the most applicable form of validity to assess measurements (Andrews, 1984; Creswell, 2005; Mahoney, 2008; Messick, 1981, 1989; Popham, 2003; Embretson & Gorin, 2001; Gay & Airasian, 2003; McMillan & Schumacher, 2006).

In particular, construct validity is concerned with the efficacy of a test to gauge learner knowledge about the relevant topics of concern. The test must be relevant, appropriate and utilised correctly, with the focal point being the integration of evidence that produces inferences about assessment results. These inferences must be meaningful, trustworthy and serve the purpose of the assessment for construct validity to reach its goal (Messick, 1989).

If these three steps are integrated into an assessment then the degree of construct validity will be high. In the same way, this study explores whether these three steps are successfully implemented in the Picture Vocabulary Test, as discussed in Chapter 1. On the other hand, if the construct validity of an assessment is not the central focus, it means that the assessment does not assess what it is supposed to, causing the validity level to lower. If an assessment does not produce the same results across different groups then the level of construct validity comes into question.

Cronbach and Meehl (1955, p. 283) view construct validity as asking to what extent a test is culture-free. There are two major threats that exist for construct validity, the first major one being that the construct is underrepresented because it has limited facets of the construct, or too few relevant items to accurately assess the desired topic. The next major threat is 'construct-irrelevant variance', meaning that the test has too much reliable variance, for instance, making certain items easier or harder for certain learners, in such a way that is irrelevant to the construct being measured (Messick, 1989; Moss, 1992).

Linking the above to this study, the focus of construct for the overall SAMP assessment was to determine the relevant level of the academic knowledge and skills possessed by Grade 1 learners across different fields (see Chapter 1, Section 1.2.4). As for this study, the construct of the Picture Vocabulary Test was under investigation, in particular its implementation across three different language groups. When assessing learners using a Picture Vocabulary Test, the objects used in the test must be familiar to all participating learners, implying that most of the objects used in the three pictures of this study must have been observed by the learners at some time or another in their past. For the Picture Vocabulary Test to have a high level of construct validity, all the objects (items) that need to be identified by all learners have to perform the same across all three groups. Alternatively, if this is not the case this study will provide suggestions as to how the level of construct validity can be increased. Meaningful, accurate and justifiable inferences can only be achieved through a high level of construct validity.

To conclude, the level of construct validity plays a vital role regarding the inferences made about the test scores of the learners participating in a Picture Vocabulary Test.

3.3 RELIABILITY

A test is seen as being reliable when it can be used by a number of different researchers under stable conditions, with consistent results and the results not varying. Reliability reflects consistency and replicability over time. Furthermore, reliability is seen as the degree to which a test is free from measurement errors, since the more measurement errors occur the less reliable the test (Fraenkel & Wallen, 2003; McMillan & Schumacher, 2001, 2006; Moss, 1994; Neuman, 2003). In the same way, Maree and Fraser (2004) ask how far the same test would produce the same results if it was administered to the same children under the same conditions. This helps the researcher and educator to make comparisons that are reliable. The more errors found in an assessment the greater its unreliability, and *visa versa*. Reliability is a very important factor in assessment, and is presented as an aspect contributing to validity and not opposed to validity.

Messick (1989) transformed the traditional definition of validity - with reliability in opposition - to reliability becoming unified with validity. Thereby Messick (1989) has accepted a unified concept of validity which includes reliability as one of the types of validity; thus contributing to the overall construct validity. As Messick (1989, p. 8) states:

Hence, construct validity is a *sine qua non* in the validation not only of test interpretation but also of test use, in the sense that relevance and utility as well as appropriateness of test use depend, or should depend, on score meaning.

Here Messick (1989) explains that not only is construct validity essential for test interpretation but also for test use. The test must be relevant and be able to be utilised in a reliable manner.

With a Rasch analysis, the '*item reliability index*' was examined, reflecting whether the items could be replicated in the same order if they were given to another sample

group that had similar or equal abilities (Bond & Fox, 2001). If items, or in this case the objects order used in the pictures, are not replicable across the three groups, then the reliability of the test is in jeopardy. In this study the real person and real item separation reliabilities were explored. Similar to internal consistency, separation reliability values of between 0 and 1 had to be obtained, and those that are high are beneficial to an assessment (Scherman, 2007).

The items performance can be assessed through Rasch analyses which can alert a person to the ordering of the items. Each item's difficulty is situated along a logit scale, together with its degree of error. The more information available about the difficulty of the item, the more the estimation error decreases. If items are clumped then the difficulty of the items are not equally dispersed, which influences the level of validity as well as the reliability of a test. With the help of Rasch, items that are clumped or too easy or difficult can be identified and dealt with accordingly. If an assessment is focussed only on reliability, the validity level of the assessment will decrease. As was pointed out earlier Messick's (1989) unified concept of validity also includes reliability and overlaps each other. Consequently, a balance has to be reached between reliability and validity.

3.4 CONCLUSION

With this study empirical investigations took place by means of Rasch analyses, in order to determine the level of validity of the Picture Vocabulary Test. Following Messick's (1989) concept of validity and reliability interrelating within an assessment. These empirical investigations that took place lead the study to integrate a Positivist theoretical framework which aided in investigating the items and the level of validity of the assessment distinct from the learners or other factors. Statistical procedures were used as noted earlier and these are discussed in Chapter 4 to follow.

CHAPTER 4

METHODOLOGY

4 METHODOLOGY

4.1 INTRODUCTION

As mentioned before this research study was quantitative, with the research design illustrating the procedures followed and how the data was collected and analysed in order to answer the research questions. The results and findings of this study will be discussed in Chapter 5.

A detailed description of the Rasch analyses that were conducted to analyze the data of the Picture Vocabulary Test is provided in Section 4.1.6. In Section 4.1, the research methodology pertaining to this study is discussed, how the data was collected and the sample selected. The ethical considerations were taken into account in Section 4.2, before the conclusion in Section 4.3.

The main research question that was asked in this study was:

How do objects used in a Picture Vocabulary Test influence the level of validity?

The main research question has been broken down into more detailed questions that can lead the research study to explore objective answers.

What barriers to validity used in a Picture Vocabulary Test can be identified from literature?

Literature was explored to identify barriers most applicable to this study to the validity level of the Picture Vocabulary Test. These areas were language, culture and Visual Literacy as discussed in Chapter 2.

To what extent is a unidimensional trait measured by a Picture Vocabulary Test?

The objects presented in the pictures in the Picture Vocabulary Test are supposed to measure a single trait or ability of the learners. The trait or ability that the Picture Vocabulary Test is supposed to measure is vocabulary. This was investigated to determine whether this was the case.

To what extent do the items in a Picture Vocabulary Test perform the same for the different language groups?

On establishing that a unidimensional trait was indeed measured by the Picture Vocabulary Test the items were explored even further in order to determine whether the items were performing the same across the three language groups. Statistical procedures were used to get to an answer for this question.

How can the identified barriers that decrease the level of validity be minimized?

This question aims to provide suggestions as to how the objects that are barriers to the construct validity and the inferences made can be effectively addressed. The suggestions will try to provide insight into the means to increase the construct validity and decrease barriers that are detrimental to validity.

4.2 RESEARCH METHODOLOGY

The theoretical position within which the design of this research study followed was Positivist as mentioned earlier, which made use of a quantitative methodological approach to determine the actual reality regarding the above questions being researched, and to draw conclusions. A systematic form of measurement took place so that the conclusions or inferences made were objective (Eloff & Ebersohn, 2004) which is typical of a Positivist approach.

4.2.1 Positivism

'*Cogito, ergo sum*' – 'I think, therefore I am', a pronouncement by René Descartes (cited in Phillips & Burbules, 2000), greatly influenced modern philosophy. The

French philosopher, in his famous *Meditations*, wrote that he had accepted false opinions to be true from his youth, and promised to rid himself of the opinions he had adopted. Locking himself in a small room with a fireplace, during the winter, he examined all his beliefs. The false opinions created a desire in Descartes to establish a rigid superstructure of the sciences, but the only concrete, secure belief he could find was 'I think, therefore I am'. Descartes being a Foundationalist and a member of its rationalist division, identified the foundation of Positivism based on "...what could not possibly be rationally doubted and seemed indubitably true should be accepted as true" (Phillips & Burbules, 2000, p. 6). In other words, truth and reality do exist and can be measured and explained if found, leaving no room for doubt (Clark, 1998; Guba, 1990). These premises were the foundation upon which Positivism was built.

The Positivist approach has been a recurring theme since Plato, who believed that nature had certain unalterable ideas (Loving, 1997) that needed to be tested and proven true. These seeds of Positivism began sprouting in the 17th and 18th century during the period of Enlightenment (Ponterotto, 2005). Francis Bacon (1561-1626) had a sincere commitment towards the rules of evidence, feeling that modern science's purpose was to investigate a nature that was waiting to be discovered, and defined by man through induction. The two notions of induction proposed by Bacon was one of pure discovery and the other a method of observing and then testing hypotheses which lead to logic or justification (Abraham, 1996; Alexander, 2006; Loving, 1997; Phillips & Burbules, 2000). Later, in the 1920's the philosophy of Logical Positivism was developed by the Vienna Circle, a small group of philosophers, physical scientists, social scientists and mathematicians, who focused on the 'Received View' (Abraham, 1996; Phillips & Burbules, 2000), postulating that what is seen, is believed. It was seen as a pointless task to make statements about happenings that could not be verified by the senses. Positivism was developed further by B.F. Skinner, a behaviourist psychologist who also had a major influence on scholars in the Positivist movement (Abraham, 1996; Buchanan, 1998; Kidd, 2002; Lather, 2006; Phillips & Burbules, 2000; Schulze, 2003; Wardlow, 1989). In Table 3.1 the basic constructs of Positivism and how they link with this research study (highlighted in light green) are depicted.

Table 4.1: Positivism linked to the research study

(Adapted and combined from Scherman, 2007 & Guba & Lincoln, 1994)

Positivism linked to this research study	
Inquiry Aim	To explain, predict or control
In this study	To explain how the items function in the Picture Vocabulary Test
Method	Quantitative
In this study	A scientific Rasch analysis will make the study quantitative
Logic	Deductive
In this study	If the items do not function as expected then we can deduce that the Construct validity which has implications for interpretations about validity
Epistemology	Objectivist - objectively true
In this study	The researcher and the subjects are independent of each other
Ontology	Realism
In this study	When the items are analysed they show a true reflection of how they function in reality to the learners

The basic constructs of Positivism are linked to this study and further explained.

Inquiry Aim

The aim of the inquiry is to make use of a Rasch analysis to determine how each item functions in the Picture Vocabulary Test. The data will be analyzed and the order of difficulty of the objects as experienced by the learners will be revealed. This will all be shown in an item pathway, a figure in which the items are represented on a vertical axis from the easiest at the base to the most difficult at the top. If certain items do not follow the logical flow according to the Guttman scale they will be identified and an explanation given. On a Guttman scale items are arranged in a

pattern from the easiest to the most difficult (Bond & Fox, 2001; Cavanagh, Romanoski, Giddings, Harris, & Dellar, 2003).

Method

This research study followed a quantitative approach, with a Rasch analyses making use of various scientific formulae to explain certain phenomena occurring in an assessment and the order of difficulty investigated (Bond & Fox, 2001).

Logic

As shown in Table 3.1 (above), the logic is deductive. By using the knowledge received from the data analyses, the objects are arranged in order of difficulty.

Epistemology

By taking an objective viewpoint and observing the outcome of the Picture Vocabulary Test, knowledge is gained about the level of difficulty of the items and if the items are performing the same for the three language groups.

Ontology

Ontology is seen as the reality of a situation. In this study three different language groups of learners from diverse backgrounds are being assessed by an instrument originating from the UK.

4.2.2 Rationale for working with Positivism

With the help of Positivism, empirical investigations can be made to answer questions. Positivism is seen from the perspective that science does not need to have a prior sense of the whole to which different parts belong in order to study the different parts (Fischer, 1991). In this study the items alone will be explored, not the learner or any other related aspects.

A Positivist research approach to the educational sphere makes use of methods that directly investigate the questions asked. The different methods allow for a chain of reasoning that is lucid and rational, and that can be replicated across various fields of

study. The most beneficial factor of Positivism is that it is open to professional inspection and critique (Fischer, 1991).

Through Positivism, science determines to find the truth about how physical, social and personal worlds are configured. This is done by means of empirical testing and evidence in order for the truth to be claimed. Science is seen as universal, and though different methods are incorporated, the same methodology is used. Science may be considered rational if its truths are seen as similar across cultures (Mathews, 2004, p. 23), but such a claim is highly problematic since cultures vary widely in their understanding of even the basic tenets of reason and science. With the help of a Positivist approach the items performance across three languages were explored to create a better understanding of the roles they play in the assessment.

By making use of statistical procedures and empirical testing to determine how items perform in the Picture Vocabulary Test, a Positivist approach is taken. Simultaneously the level of validity of the Picture Vocabulary Test was also explored. With the help of Rasch analyses – a statistical procedure used in this study (discussed in the following chapter) the validity level was investigated. Rasch analyses can specifically be used to determine the level of construct validity of an assessment (Tennant & Conaghan, 2007).

Reliability and validity form the crux of any measurement since they are important in establishing the credibility and truthfulness of the findings. Both reliability and validity are represented in many types and forms and have multiple meanings (Neuman, 2003).

4.3 RESEARCH METHODOLOGY

Quantitative research was used in this study as described in McMillan and Schumacher (2006). The study was carefully planned and conducted in order to enhance the credibility of the results. An existing general assessment from the UK, that was used to assess Grade 1 learners, was explored so that statistical conclusion could be made regarding the data and level of validity. By employing quantitative

methods, measurements are made of each item's difficulty compared to the learner's ability, thus helping to establish whether the items follow the correct order of difficulty as well as to detect any possible bias that will influence the level of construct validity. Table 3.2 summarises the different instruments used and analyses conducted in order to address each research question.

Table 4.2 Research questions, instrument and analysis

Research Question	Instrument	Analysis
What barriers to validity used in a Picture Vocabulary Test can be identified from literature		Literature review
To what extent is a unidimensional trait measured by the Picture Vocabulary Test?	Picture Vocabulary Test	Rasch analysis Developmental pathway
To what extent do the items in the Picture Vocabulary Test function the same for the different language groups	Picture Vocabulary Test	Differential item functioning
How can the identified barriers that decrease the level of validity be minimized?	Picture Vocabulary Test	Reflections on the analysis

4.3.1 The SAMP Sample

SAMP chose the target population of Grade 1 learners speaking Afrikaans, English and Sepedi within Pretoria, Gauteng, South Africa. These languages were selected because they are the most dominant in the Pretoria area and were also the most accessible population for the SAMP project. Multi-phase sampling was used whereby schools were stratified according to medium of instruction. Eight schools were

selected randomly from each medium of instruction from the DoE databases. A sample of 22 schools was selected, including 2 dual medium schools. The sample was inspected to ensure geographic representation of the Pretoria area and found to be satisfactory.

4.3.2 Instrument

The instrument in this study is a Picture Vocabulary Test that was used to assess the sample group of Grade One learners' ability to identify certain objects. A paper-and-pencil test (Gay & Airasian, 2003) was used, and the learners had to identify various objects presented in the Picture Vocabulary Test as pointed out by the fieldworker, who noted the answers on a sheet with a pencil. A paper-and-pencil test makes use of a standard set of questions presented to the learner, requiring cognitive tasks to be completed (McMillan & Schumacher, 2001, p. 189).

The PIPSSA Picture Vocabulary Test, as described in Chapter 1, was the original instrument developed in Durham specifically for UK learners. The pictures were slightly contextualised to accommodate South African learners but the difficulty order of the objects remained the same as presented in the PIPSSA Picture Vocabulary Test. South African learners were thus asked to identify objects in the modified SAMP Picture Vocabulary Test which followed the order of difficulty used in the original PIPSSA Picture Vocabulary Test.

The objects presented in the pictures acted as stimuli for the learners to answer the questions asked by the fieldworkers. There were three different pictures in the Picture Vocabulary Test, each with progressively difficult objects for the learner to identify. The first picture was a kitchen, in which the learners had to identify 7 objects. The second picture was of a bedroom window overlooking a field where 10 objects had to be identified and the third a child's bedroom where 5 objects that had to be identified (see Chapter 4 as well as appendices A, B and C).

4.3.3 Data collection

The SAMP assessment takes place at the beginning and end of the year. The schools participating in the SAMP assessment were visited on various days. The

Picture Vocabulary Test which is a subtest of the SAMP assessment was administered to the Grade One learners on a one-to-one basis by the fieldworkers, who were trained to ensure that the assessment was administered in a standard way to all the learners. The assessment took place in the area designated to the fieldworkers by the participating school over a two day period. The fieldworker fetched each Grade One learner from his or her classroom then followed the correct protocol by setting the learner at ease before the assessment started. The fieldworker then asked the learner to identify various objects from different pictures used in the Picture Vocabulary Test. The learner received a mark of 1 for each correct answer and 0 for an incorrect answer. Each object was worth one mark, and these make up the total number per picture a learner could achieve. The correct answers given by the learner were compared to the total number of objects the learner was asked to identify per picture. The results for the Picture Vocabulary Test were then worked out per learner, and captured electronically.

The responses were marked on an optical reader in pencil by the fieldworkers administering the test. An optical reader is a form that allows the fieldworker to colour in a circle next to the correct response. If the response is incorrect the circle is not coloured in. Once the participating schools had completed the SAMP assessment, the optical readers were sent to independent data capturers. The optical readers were processed through specialised machines. The circles coloured in pencil allowed for the data to be magnetically screened. The lead in the pencil allows for easy recognition by the machines capturing the data. This data received from the optical markers was then sent via e-mail to the CEA to be further processed.

4.3.4 Data analysis

“Data analysis is the vehicle used to generate and validate interpretations, formulate inferences, and draw conclusions”, as stated by Scherman (2007, p. 147). The data analysis for this research study followed a quantitative approach; using Rasch analyses (see Section 3.2.6). By following statistical measures of enquiry, exact measurements can be made to determine whether there are any significant differences in the performance of the items used in the test. The inferences made

about the results of the learners with regard to the Picture Vocabulary Tests can help to determine the level of construct validity.

By making use of Rasch analyses, the functions of the items of the test can be scientifically investigated. Rasch analyses are quantitative in nature because of the attributes of 'additivity' and 'ordinality' (Acton, 2003, p. 902). It has been used over the past 40 years but is being used more as a research tool by many researchers as the adequacy of the instrument and its level of construct validity can be verified (Callingham & Bond, 2006; Rasch, n.d.; Tennant & Conaghan, 2007). Analyses of the data made use of Descriptive Statistics (Section 3.2.5), Rasch Analyses (Section 3.2.6), and Differential Item functioning (Section 3.2.7). These are discussed as follows.

4.3.5 Descriptive statistics

In order to report on the data analyzed, the mean, mode, median, range of scores and minimum and maximum standard deviation were measured. This was done for each object as well as each individual picture. The descriptive statistics facilitated the process of writing about the results of the data that was analyzed (Scherman, 2007). For all the necessary analyses, pathways and graphical representations described in the above sections, a statistical programme, *WINSTEPS* (Section 3.2.8), was used.

4.3.6 Rasch Analysis

In the 1960's, Georg Rasch, a Danish mathematician, introduced a simple logistic model to construct objective measures (Boone & Rogan, 2005). Designed to overcome the problem of defining the difficulty of an item independently of the subject, it also determined the ability of the individual independently of the items. Relevant to this study, the data of the Picture Vocabulary Test can be analyzed independent of the subject (Bond & Fox, 2001; Bush & Schumacker, 1993; Linacre, 1993; McCamey, 2002; Tennant & Conaghan, 2007; Waugh, 1999).

The Rasch model can be applied to analyse dichotomous data and polytomous data (Pallant & Tennant, 2007), and various types of questions or items. Dichotomous data can only be right or wrong, assigned the value of 1 or 0 respectively. There are

also multidimensional Rasch models that deal with more complex forms of assessment, as noted by Rost and Carstensen (2002) and Briggs and Wilson (2003). These forms of assessment, such as the partial credit model and the rating scale model, are not relevant to this study and will therefore not be discussed (de Beer, 2004; Henson, 1999). This study utilized the dichotomous model, also known as the one parameter model or b-parameter model; it deals with the difficulty value of items in an assessment, and focuses on whether the items follow the correct order of difficulty (de Beer, 2004; Dinero & Haertel, 1977).

The learners had to identify various objects. For this study, a correct response was awarded a mark of 1 and an incorrect response 0, known as the '*observed score*' (Fox & Bond, 2001, p. 173). Each question or item became progressively difficult, following the Guttman scale.

4.2.6.1 Unidimensionality

The Rasch model that this study used is a unidimensional measurement model, focusing on one attribute, trait or ability at a time. Thus, items represent only one trait or dominant factor (Henson, 1999). This allows the researcher to develop useful, meaningful and descriptive insight from the analyzed data. The main principle of unidimensionality requires that analytical procedures are to be incorporated to test the degree to which learners participating in the assessment and items fit this idea of a unidimensional line or whether a single trait is being measured (Bond & Fox, 2001). In this study, the measuring of a single trait had already been investigated and implemented when the original PIPS instrument was designed.

By means of a Rasch analysis, evidence can be provided as to whether a particular item over- or under-discriminates, and if any anomalies exist in the ordering of the items. If objects in the Picture Vocabulary Test are too easy or too difficult Rasch analyses will demonstrate exactly how these items performed. Also, if an object is more difficult for one group than another then this item can also be identified. The Rasch model can also provide diagnostic opportunities in which the items can be explored further. This attribute of a Rasch analysis was most beneficial to this study since the order of difficulty of the items were in question (Andrich, de Jong & Sheridan, n.d.). To determine whether anomalies existed with the ordering, 'Fit

Statistics' were used in the study to shed more light on the different objects' difficulties. Bohlig, Fisher, Masters and Bond (1998) argue that misfitting items are not to be thrown out but rather contemplated and explored for possible reasons.

4.3.6.1 Item Characteristic Curve (ICC)

The probability of the learner responding correctly to the item is expressed through an Item Characteristic Curve (ICC). A correct response to an item is dependent on both the learner's ability and the item's characteristics. The ICC is a mathematical function or a visual representation of the learner's ability and the item's characteristics. An ICC has two asymptotes, the upper asymptote is on the vertical axis at 1.0 and the lower asymptote never reaches 0, as seen in Figure 3.1:

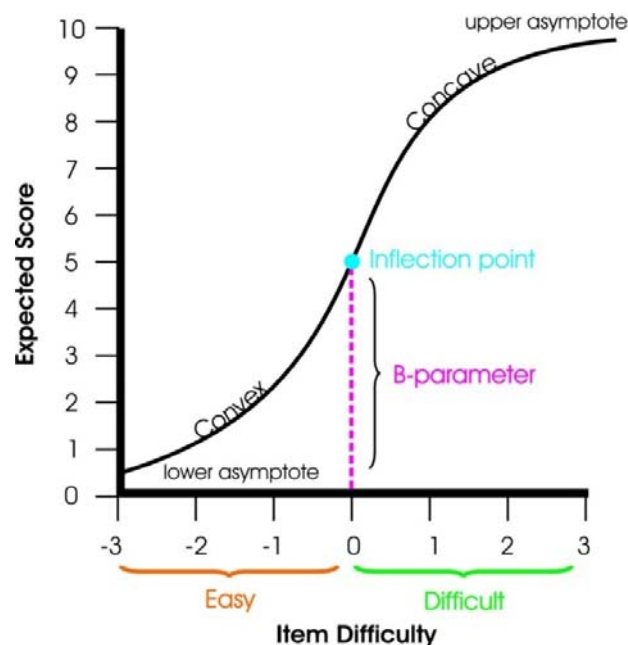


Figure 4.1: Item Characteristic Curve
(de Beer, 2004)

The probability of a correct response to an item by the learner is a continually increasing curve (de Beer, 2004). ICC's differ from one another, with the horizontal location of the inflection of the ability axis shifting more to the right or the left. An inflection occurs when the ICC goes from concave to convex, showing the difficulty level of the item. The ability axis is also known as theta (θ) axis on an ICC. The

horizontal point where the inflection occurs is known as the difficulty level or b-parameter, value or item difficulty (Bond & Fox, 2001). If there are any objects detected that function differently for one group after a DIF analysis has been done, ICC's will be created by Rasch. The ICC's will then show where the inflection occurs and how the objects function.

The b-parameter reflects the point where the ability of the learner to get a correct response is 0.5 or 50%. The greater the value of the b-parameter, the more difficult the item. Theoretically the b-value is from minus infinity to plus infinity ($-\infty$ to $+\infty$), but a value of -2.5 to +2.5 is the typical range. -2.5 indicates a very easy item and +2.5 a very difficult item.

4.3.6.2 The One-Parameter Dichotomous Rasch model

Using the Rasch model that follows a Guttman scale will result in some learners being seen as having more ability than others, and there is a greater probability that the learners with high ability will get the easier items correct. If this is not the case, then the assessment is faulty or has a low level of construct validity (Sick, 2008). Rasch analyses generate separate estimates of each item's difficulty and the learner's ability. These estimates give the researcher a value relative to every individual's ability and every item's difficulty. In other words, a Rasch analysis tells the researcher how the item is functioning relevant to the ability being assessed. It also provides indices to determine if there are items that are spread out or in 'clumps'. The items should move up in difficulty at equal levels and not be grouped on one difficulty level. If this happens in an assessment, the level of construct validity would be in jeopardy since the items do not follow the true Guttman style, each question becoming progressively more difficult (Bond & Fox, 2001).

The Rasch model further provides an opportunity to examine the responses received from the learners to see if they form a pattern that suits the expected outcome. These response patterns from the assessment are tested against what is expected from the specific assessment. The Rasch model is a powerful tool for determining item ordering (Tennant & Conaghan, 2007, p. 1361). In this study, the response patterns

of the learners were compared to the various items' levels of difficulty. This helped determine if the item ordering for the Picture Vocabulary Test was correct.

Rasch also allows the unification of various measurement issues that are required to verify the validity of an assessment. With unification, Rasch measures a single latent trait or ability of a learner and endeavours to specify what occurs when a learner attempts a specific item (Engelhard & Osberg, 1983; Henson, 1999; Wright, 1977). Measuring a single latent trait in this research study involved measuring the learner's ability to identify objects in the pictures used in the Picture Vocabulary Test. If a number or letter had to be added to the Picture Vocabulary Test, a single latent trait could not be measured because now additional abilities or traits were involved (identifying letters and numbers not only objects).

Rasch is useful for reviewing the measurement properties as well as the unidimensionality (see Section 3.2.6) of an assessment. An interval level scale is created by the Rasch model to show the interaction between the learners participating in the assessment and the items used in the assessment (Callingham & Watson, n.d.). Rasch analysis "...provides a complete solution to almost every measurement problem encountered in science" (Wright & Mok, 2004, p. 24).

4.3.6.3 Exploring the data using Rasch

Another reason the Rasch model was chosen was that only a single attribute or latent trait, namely vocabulary, was measured. Each item is expected to contribute meaningfully to the construct being measured, in a hierarchical order from easy to difficult, in the Picture Vocabulary Test. For Picture 1 there were 7 items, Picture 2, 10 items and Picture 3, 5 items, making a total of 22 items. It was important to explore whether the items followed the specifications of hierarchy of item difficulty.

With the help of Rasch analyses, the extent to which the learner performs and the difficulty of the items can be determined along a continuum. Ordinal data is converted to interval data, allowing inferences to be made about the difficulty of the object and to investigate the construct validity (Kyriakides, Kaloyirou & Lindsay, 2006). By fitting the data from the Picture Vocabulary Test to the Rasch model, detailed examination

took place of the level of construct validity, how the items were ordered, the unidimensionality, and whether the items worked the same across all three language groups. An assumption on which the Rasch model is based is that the difference between item difficulty and person ability indicates the probability of a learner being successful with a particular item (Kyriakides et al., 2006). The data of the Picture Vocabulary Test will be explored to see if it follows the Rasch model accurately. A misfit in unidimensionality is indicated by the real standard error. A value of 0.9 or over indicates unidimensionality, while 0.5 and below indicates multidimensionality (Tenant & Pallant, 2006). This was examined to detect any possible misfitting items in the Picture Vocabulary Test.

In the Picture Vocabulary Test a value is given to a response as either 0 (incorrect) or 1 (correct). A 50% chance exists that the learners will get the item correct or incorrect. The probability of a correct response is a logistic function that is determined by the difference in a learner's ability to correctly identify the object and the difficulty in doing so. Items that fit the Rasch model have an item INFIT range of 0.77 to 1.30 (Kyriakides et al., 2006) and a related Z statistic of -2 to +2 (Beaton & Wright, 2005). Any items that do not fall into this range indicate a tendency that they do not follow the expected response pattern of easy or difficult. The response patterns of all three language groups were explored to determine whether any unexpected responses occurred. Furthermore, any items that were experienced as being too difficult or too easy for the learners were identified.

In Rasch analyses the items and the odds ratios are evaluated. Here the odds refer to the probability of successfully answering an item correctly divided by the probability of answering the item incorrectly. The odds ratio is the natural logarithm called natural log-odds, which in turn are referred to as logits (Schumacker, 2004).

Items that do not yield the same results across two or more groups show bias, known as DIF (see Section 3.2.7), which allows comparison of results to be made between various groups (Huang, Church & Katigbak, 1997). The performance of items across the three language groups will be compared in order to detect any items that were not performing in the same way across the groups.

Misfitting items show an unexpected response and an obscured relationship of the probabilities compared to the other items (Lundgren-Nilsson, Grimby, Ring, Tesio, Lawton, Slade, Penta, Tripolski, Biering-Sørensen, Carter, Marincek, Phillips, Simone, & Tennant, 2005). Individual item fit statistics are acceptable within the range ± 3 . Any items that are misfits were identified by the Rasch model used in this research study.

4.3.6.4 Fit Statistics

Fit statistics help a researcher detect any discrepancies found between the Rasch Model's expectancy and the actual results of the test that is whether a learner or item's performance is consistent with others (Kyriakides et al., 2006). In order to determine how well the tests data fits the Rasch model's data, chi-square fit statistics were used. With a Rasch analysis, two chi-square ratios are reported: *INFIT (weighted)* and *OUTFIT (unweighted)* Mean Square statistics (Fox & Bond, 2001). Most researchers are more concerned with the INFIT statistic, since it gives more insight into the learner's performance. The learner, whose ability is closer to the item's difficulty, allows for greater understanding about the specific item's performance. *OUTFIT statistics* are concerned with the difference between the expected and observed scores, while with *INFIT statistics* extreme items or persons are detected for targeted items (Tenant & Pallant, 2006, p. 3).

In order for fit statistics to be interpreted, there is a need for experience that is related to that specific measurement context. It is essential to know whether a mean is too large or too small, since each test has its own unique situation. Wright and Linacre (in Fox & Bond, 2001, p. 179) produced a set of general guidelines for researchers, according to whose table a reasonable Item Mean Square range for a multiple choice test is used, from 0.7-1.3 (Bond & Fox, 2001; Tenant & Pallant, 2006).

INFIT is a sum that carries much information, as mentioned above. The statistical information is its "...variance [and] the Standard Deviation (SD) of the estimate squared..." in a Rasch observation (Fox & Bond, 2001, p. 176). To calculate INFIT, each squared standardized residual is weighted by its variance and then added. The total is then divided by the sum of the variances. This produces the same distribution

as the OUTFIT but the differential effects of the weighting can also be seen (Beaton & Wright, 2005).

OUTFIT is the sum of squared standardised residuals, a residual being the difference between the observed score and the expected response. The residual contributes toward misfit in that the greater the residual value, the greater the possibility of misfit. OUTFIT is calculated by squaring each residual, then adding the residuals together and dividing by the number of items to get the mean square (Fox & Bond, 2001).

There are two aspects of fit on which fit statistics focuses. One aspect is the standardised form, known as the t statistic, with acceptable values of -2 to +2. The t -statistic is also known as the INFIT t and OUTFIT t . When the observed data conforms to the Rasch model, the t value is near 0 and the SD near 1. If the data is less compatible, the t values are greater than +2 or less than -2. When a t -test value has infinite degrees of freedom or the t -statistic has been modified to a unit normal value then ZSTD (standardized as a z-score) is used (Linacre, 2009).

The other aspect is the unstandardised form, known as the 'mean square' or 'average value' of the squared residuals of a specific item. The residual values are the differences between the Rasch model's theoretical expectation of how the item will perform and the actual performance of the item used in the assessment (Tenant & Pallant, 2006). The greater the residuals, the greater the difference between how the item was expected to perform and how it actually performed. All residuals are squared in order to make any minuses into plusses so they can be added in order to give a sum of differences. Therefore INFIT and OUTFIT are always positive, allowing for the mean square fit statistic to be used to monitor the compatibility of the item with the Rasch data (Bond & Fox, 2001). An INFIT mean square of greater than 1 indicates more variation than the Rasch model predicted (underfit). An OUTFIT mean square of less than 1 indicates less variation than was modelled by the Rasch model (overfit).

Item difficulty is estimated from the proportion of learners who succeeded on each item, while person ability is calculated by the proportion of items of which each learner succeeded in. These processes lead to the items being calibrated into logits

and a set of all the learners' measures. For every item used in the assessment, an estimate is given for its difficulty shown in logits.

Any data that was missing in this study was not coded as incorrect because it was interpreted that the learner did not achieve that level of difficulty in the Picture Vocabulary Test. The missing data was not discarded but kept as the missing data can be handled by the WINSTEPS program.

While item difficulty is important item spread along the continuum is also of importance. With the Rasch model reliability indices are generated for both persons and items. The person reliability index indicates the replicability of the order of the persons that could be expected if the sample of learners were given a parallel test measuring the same construct. For person reliability to be generated ability estimates and well targeted items are needed but also a large enough spread of ability across the sample so that a hierarchy of abilities can be measured. Low person reliability indicates that more data has to be collected to reduce error of the estimates. The item reliability index indicates the replicability of items and the placement of items if given to another sample. High item reliability indicates there are items that are more difficult and some that are easier (Bond & Fox, 2001).

4.3.6.5 Item-Learner Map

An Item-Learner Map is a pathway used to represent the development between the items and the learners. It has basic information fundamental to Rasch measurement (Bond & Fox, 2001). In an Item-Learner Map a vertical line can be seen that has X's on the left side, representing the number of learners and their abilities along the variable (y-axis). On the right hand side of the vertical line the objects and their difficulty measures are displayed. Ideally there should be an even spread of items along the variable with no large gaps or clumps. The items should also be lined up with the learners' abilities. The variable (y-axis) measure starts from a minus (easy) value at the bottom to a plus (difficult) value at the top. At the bottom of the Item-Learner Map a '# is 10' is shown with a number. For example '# is 10' if there are 3 next to each other - '# # #' - that means there are 30 learners at that specific measurement of the variable (Linacre, 2009).

4.3.6.6 Item Pathway

An item pathway is a diagrammatical representation of where the items lie along the unidimensional line according to the responses from the learners and the difficulty of the items in an assessment. The learner's abilities are also usually represented on the pathway so that an easy diagnosis can be made of the learner's ability and the item's difficulty (Bond & Fox, 2001).

The location of the items on an item pathway allows for a better understanding of how they function (Bond & Fox, 2001). In this study, the focus is more on the items and therefore the Developmental Pathway in Bond and Fox (2001, p. 22) has been adapted to include only items, not learners, as can be seen in Figure 4.2:

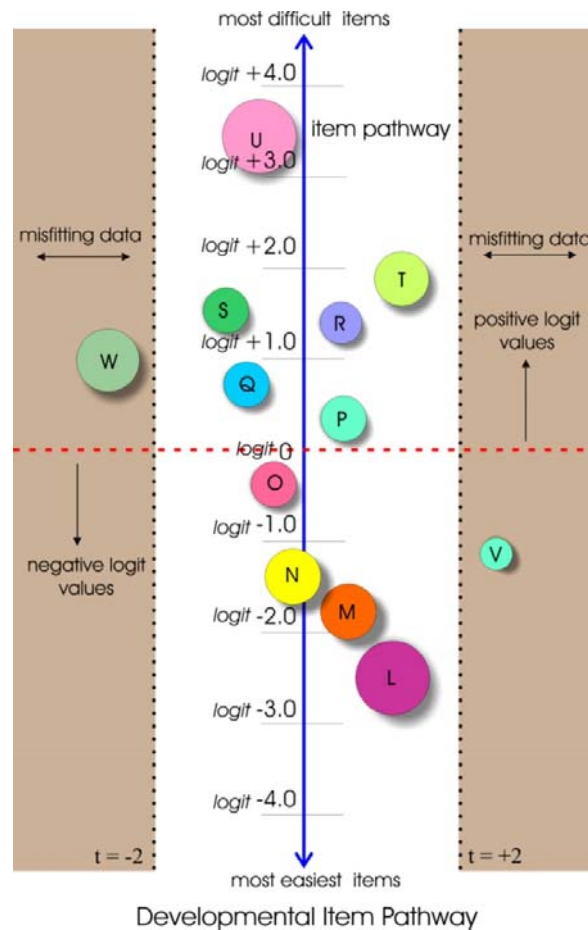
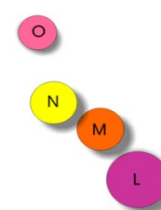


Figure 4.2: Item Development Pathway
(Adapted from Fox and Bond (2001))

The blue line (right) in the Figure 4.2 is the centre and represents the ideal that carries on to infinity and is the unidimensional line along which the items are situated. Interval scales or the dividing segments ‘represent the levels of difficulty of the items which are subjected to a log transformation represented along a ‘logit scale’. Each logit unit has a consistent value. The mean of item difficulty is 0, with any item above 0 seen as being difficult and any item with minus (-) as easy (Bond & Fox, 2001, p. 33).



The multi-coloured buttons, or ‘stepping-stones’, (Bond & Fox, 2001, p. 21) represent all the various items and persons. The pathway at the lower end is typically easier items, such as items L, M and N. Those items at the top are typically more difficult, such as items S, T and U. The distance between the locations of the buttons represents the level of difficulty between the items. The ‘greater’ the distance to the next button the greater level of difficulty from one item to the next. Looking at the IDP, item N is quite a step away from item O. In an ideal test, the buttons should be the same distance apart indicating that each item moves up in difficulty at an equal level. The buttons are also situated at different distances from the blue line, the closer to the line the closer to the ideal of what is meant to be measured in the assessment, while the ones further away are slightly less ideal for what is being measured. As long as they are close enough to the blue line and fall in the pathway, then the assessment is well targeted for what it is meant to be measuring. In addition, some of the buttons are larger than others as the difficulty of the item in the test is located at a specific point and has a ‘zone of error’ or degree of error associated with it. The smaller the buttons, the smaller the error, which helps the researcher to be more accurate in determining the level of difficulty of that specific item. With the larger buttons, the ‘zone of error’ is larger, therefore the item’s difficulty cannot be determined so accurately and they are not located as precisely (Bond & Fox, 2001).



The vertical dotted black lines on the left and right of the IDP and shaded in brown represents the edges of the pathway. Buttons that fall within these two boundary lines, or the white area, are seen as items that are useful, but items that fall in the light brown shaded areas are seen as problematic, such as items V and W. These items cannot



be interpreted meaningfully in relation to the other items as the possibility exists that they may be measuring another trait (Bond & Fox, 2001).

The dotted horizontal line reflects the point where the items cross - - - - - over from easy to difficult.

4.3.7 Differential Item Functioning (DIF)

A test has a high level of construct validity when the items perform the same way across different groups. In order to determine this, various methods can be incorporated to test for Differential Item Functioning (DIF) (Mahoney, 2008), the most appropriate for this study being a Rasch analysis.

Within assessments across gender, language and culture it is expected that the items function invariantly, provided that all the learners have the same amount of knowledge for a given subject (Badia, Prieto & Linacre, 2002; O'Neill & McPeck, 1989). This is not always the case and for this reason DIF is a vital source to help identify bias in assessments across dissimilar groups, thereby helping to improve upon the items found in an assessment that displays bias. Items that give different success rates across two or more groups display DIF (Huang, Church & Katigbak, 1997; Tennant & Pallant, 2007).

Numerous studies have been done on DIF, especially with regard to the test performance of different groups, however only in the last decade has the focus moved more towards the differences in test performance among ethnic groups as opposed to sexes. These differences among groups have been extensively reviewed by Green, Crone and Folk (1989), Kim, Cohen and Park (1995) and Wang and Wilson (2005).

When items do not perform in the same way across different groups that have the same abilities or traits DIF occurs, which means that there is a difference in the statistical properties of items. These items are said to 'operate invariantly' (Andrich, 2004, p. 3). DIF is also known as 'bias' (Andrich & Hagquist, 2004; Maller, 2001). A definition found in Smith and Smith (2004, p. 391) regarding bias is said to be the

“differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers.” Also in Smith and Smith (2004, p. 392) a more comprehensive definition is provided of bias as being the “significant and persistent interaction between some (but not all) persons and some (but not all) items.” Mahoney (2008, p. 15) elaborates: “Bias...creates a distortion in test results for members of a particular group”. With the help of DIF analysis, a statistical procedure, items that may have different meanings for different groups can be investigated, which is often overlooked by conventional processes for reviewing items (Freedle & Kostin, 1990; Scheuneman & Gerritz, 1990).

Through DIF analysis, the researcher is able to monitor whether the level of validity and fairness of the assessment is jeopardised by biased items. The ideal for valid quantitative judgments to take place is for all items to perform the same way across different groups with the same knowledge (Badia, Prieto, & Linacre, 2002). Likewise the construct validity level of an assessment is threatened when items exhibit DIF.

DIF analysis further helps to create a better understanding of the difficulty of an item and the characteristics of the group participating in the assessment, indicating the group’s relevant strengths and weaknesses (Hagquist & Andrich, 2004). The reasons some items appear to be biased can be attributed to factors such as ethnicity, exposure to various resources, differing opportunities, background, education, culture, language and life experiences (Green, Crone & Folk, 1989; Maller, 2001; Scheuneman & Gerritz, 1990; Zwick, Donoghue & Grima, 1993; Zwick & Ercikan, 1989).

Many educators discard items that appear to be exceedingly biased in favour of investigating the possible reasons. Three possible factors within a broad environment that can contribute towards bias have been identified by O’Neill and McPeck (1989, p. 256):

1. Surface features or content characteristics of the question
2. Real differences in the groups’ knowledge and skills (such as those resulting from different educational experiences)
3. The nature of the criterion used for matching.

These abovementioned factors evidently show that items with a high DIF value are not unfair items but rather items that perform differently across diverse groups with matched knowledge.

DIF can be graphically represented with the help of statistics programmes, such as *WINSTEPS*.

4.3.8 WINSTEPS

WINSTEPS is a programme used to analyse data, developed by people who, on a daily basis, were involved with analyses in the work environment. In the area of educational research, it is helpful with the many applications of the Rasch model, and was designed to “...construct measurement from the responses of a set of persons to a set of items” (Linacre, 2009, p. 29).

There are a number of advantages of using *WINSTEPS* (Bond & Fox, 2001; Linacre, 2009; Scherman, 2007):

- Letters as well as integers can be used
- Easily used with other programs such as Statistical Package for Social Sciences (SPSS) and *EXCEL*
- Data can be analysed from dichotomous, multiple-choice, rating scale or partial credit items
- Missing data can be included in the analysis
- Items and learners are analysed in depth.

With *WINSTEPS*, diagnostic procedures are used to provide information on outliers, unexpected data points and whether the test is not unidimensional. Items and the response structure are calibrated and a central estimate for each learner calculated. These are represented in the form of graphs, plots and tables. For the learner and the items measured, standard error, fit statistics and reports on item or person responses that cause the misfit are also included in the output.

One of the limitations of *WINSTEPS* is that it cannot calculate two or more parameter models, as it was designed specifically for the one-parameter model (Scherman, 2007). However, for the purposes of this study, *WINSTEPS* is ideal.

4.4 ETHICAL CONSIDERATIONS

For the purpose of this study permission by the DOE was attained when the larger SAMP research project went into the field. Clearance was obtained for the project as well as developments from it. Additional letters were submitted to the Ethics Committee for changes that took place. A letter requesting permission to conduct the study was sent to the schools as well as a consent letter to the parents wherein the project was explained, the CEA's contact numbers were provided and the benefits and/or risks of participation stated. The learner was allowed to withdraw or refuse to take part in the study at any time. The letter also ensured confidentiality and anonymity.

4.5 CONCLUSION

In conclusion, this chapter gave a methodological overview of how this research study explored the research questions. By means of this study, the construct validity of the Picture Vocabulary Test was explored, focussing on how the objects represented in the various pictures perform across the three different groups partaking in the study. Since the methodology was based on statistical procedures to explore the answers to the research questions, a Positivist viewpoint was taken, resulting in this research study incorporating Positivism as a research paradigm. The items were investigated apart from the learners or other related aspects, which is a Positivist belief i.e. parts can be studied apart from the whole (Fischer, 1991). The chain of reasoning was lucid and rational and can be inspected by other researchers and they will come to the same conclusion as this study. The Positivist viewpoint of the study lead to the research questions being empirically investigated to reach suitable answers.

The main research question followed an exploratory angle, making the study quantitative. The data was collected at the beginning of the year. The sample

consisted of learners from Afrikaans, English and Sepedi speaking schools in Pretoria. The learners participated in a Picture Vocabulary Test that consisted of 22 objects that they had to identify. The Picture Vocabulary Test is a sub-test that forms part of a larger instrument that was originally from the UK. These objects were arranged from easy to difficult for the learners in the UK. Since learners from different language groups are being assessed with the same instrument the most important and relevant to this study, is the issue of validity. How validity is influenced together with reliability. This chapter concluded with a brief discussion of the ethical considerations for this study. The focus turns to the results of the study, with the data that was analysed and discussed in the chapter to follow.