CHAPTER 3 MEASUREMENT THEORY

3.1 INTRODUCTION

Osborne (2008, pp. ix) stated that "researchers are romantic fools, research is magical". This strong and antithetical comment cited above is an unusual, yet influential way to remind all researchers and the receivers of the findings of research not to merely believe in its magic or charm. Osborne (2008) wished to think about the researchers of the 21st century as "intrepid explorers and adventurers striving to explore phenomena, understand processes, and, most of all, go where no other human being has gone before". All researchers strive to create knowledge and understanding that will change the world. Does this striving for the almost impossible make romantic fools of researchers?

3.2 CHALLENGES IN MEASUREMENT OF HUMAN BEHAVIOUR

Researchers might be perceived by others to be, "romantic fools" if they do not judiciously apply rigorous and effective principles of research to get the most accurate picture of the real world. According to Osborne (2008), it becomes a moral imperative of all researchers to produce valid and reliable results to others. This is easier said than done as many challenges face researchers when they attempt to produce valid and reliable results in the measurement of human behaviour.

Common challenges and problems have been highlighted by several authors (Furr & Bacharach 2008; Laver Fawcett 2007; Wright & Linacre 1989). Firstly, all measurements are subjected to some degree of error and even more so in abstract, unobservable constructs in human behaviour such as self-esteem, motivation, cognitive skills and the like. These unobservable constructs are often referred to as latent traits.

A second problem is that most measures that claim to 'measure' these latent traits are only an observation of certain behaviours and not measurement per se. Wright and Linacre were concerned that many researchers were confused between observation and measurement (Wright & Linacre 1989). They explained that although observation is the first step to measure, as these observations

can be counted and even the severity or amount of the construct can be rated on an ordinal scale, the down side is that ordinal scale data does not present equal distances between ratings. For example, the distance from none to mild, or mild to moderate cannot be calculated and cannot be treated as mathematical numbers. These ratings (1 = none, 2 = mild, 3 = moderate) cannot be treated as composite scores where the rating for each item is added to the remaining items in order to get a total score. The Mini Mental State Examination is an example where this principle is violated. For example, a total score of 21 out of 30 indicates cognitive impairment (Tombaugh & McIntyre 1992). This is a composite score derived from nominal and ordinal ratings.

A third challenge is that psychological constructs or attributes do not operate in isolation and some constructs have strong relationships with others. For example, memory and attention are interdependent, while self-esteem and motivation influence each other (Laver Fawcett 2007). How would a researcher know that the specific attribute under investigation is accurately measured?

A fourth challenge is closely linked to the above mentioned problem, that of score sensitivity (Laver Fawcett 2007). Will the scoring accurately indicate meaningful amounts of this attribute? This is important in outcome measurement as the main focus is to measure change. If the outcome measure cannot detect small amounts of progress (or deterioration), it lacks sensitivity.

When people are being measured, the challenge of behavioural inconsistencies is a reality. People are sometimes consciously or unconsciously influenced by the test items or the person implementing the test. Participants could change behaviour to give the preferred response, also called demand characteristic, or could avoid giving a true answer (Furr & Bacharach 2008).

The final challenge is the many measures and tests available to measure aspects of human behaviour. Before these tests can be administered to a specific population and for a specific reason, the user of the test must have inspected the psychometric properties of the test (Furr & Bacharach 2008; Laver Fawcett 2007).

Measurement principles have developed over the years and many sophisticated techniques and measurement tools are available to address the above mentioned challenges. Some of these tools and techniques are presented in the next section.

3.3 TOOLS AND TECHNIQUES IN MEASUREMENT OF HUMAN BEHAVIOUR.

All measurements are subjected to some error (Embretson & Hershberger 1999; Osborne 2008; Wainer & Tissen 2001). The Classical Test Theory (CTT), also called true score theory, has been used widely to account for some measure of error. Observed scores can be segregated into true scores and error, mathematically expressed as X = T + E. Here, X = observed score, T = true score, E = error. The true score is the average score expected when an examinee would take the same test several times. The error score is the difference between a score for a particular test and the true score. True score theory has many techniques (split-half, test-retest, internal consistency, intra-test etc) and are usually applied to test the reliability of the test scores (Wainer & Tissen 2001).

Osborne (2008, pp. 51) described several shortcomings of true-score theory:

- Measures of persons and items are test- and sample-dependent. Tests with high complexity yield lower results and vice versa. If an above-average sample performs a test (or items on a test), the average score on the items will differ from a below-average sample. Thus person ability and item difficulty cannot be generalised to other samples.
- Complete responses are needed to do comparisons. If there is missing data, true-score theory could give unfair results. Cases with missing data are not included in the calculations.
- Interaction between person ability and item difficulty cannot be predicted with true score theory.
- Mathematical calculations and statistical analysis performed on ordinal scale data lead to invalid inferences and spurious interaction effects.
- Few techniques in true score theory exist that can detect abnormal or invalid response patterns (for example, malingering) and will therefore not detect invalid or irrelevant items in a measuring instrument.

New developments in measurement techniques are constantly evolving to address the shortcomings of true score theory, such as Item Response Theory (Furr & Bacharach 2008; Wainer & Tissen 2001). There is an abundance of literature that suggests that the techniques used in Item Response Theory

are superior to the information produced by true score theory (Furr & Bacharach 2008; Embretson 1999; Osborne 2008; Tesio 2003; Wainer & Tissen 2001; Wright & Linacre 1989).

Item Response Theory is based on the principle that a person's response to a particular test item is influenced by qualities of the person and qualities of the item. Both a person's responses to an item and the properties of this item determine the trait-level estimates. A person with high ability is expected to score high on the difficult items. This is presented in an item-characteristic curve that provides the precise value of these probabilities. If these probabilities do not occur, one could suspect abnormal or invalid response patterns and inconsistent behaviours. Rasch analysis is a popular measurement tool that is used for this purpose (Clark & Watson 1995; Furr & Bacharach 2008; Laver Fawcett 2007; Osborne 2008; Tennant & Conaghan 2007; Tesio 2003).

The item-characteristic curve in Rasch will detect change in a person, that is, as a person's ability improves, he ought to score better on difficult items that he was unable to obtain previously in a testing session. The item-characteristic curve thus detects differences between individuals at different trait levels (Furr & Bacharach 2008).

Rasch analysis generated several other useful techniques that could address other challenges in measurement of human behaviour. Since the use of ordinal scale is viewed as observation and not as measurement, the ideal would be to convert ordinal data into interval data. The Rasch analysis has this capability, providing the data fits the Rasch model. If the data is based on Guttman scaling, it will most probably fit the Rasch model (Laver Fawcett 2007; Tennant & Conaghan 2007).

Guttman scaling is used when descriptions or statements for an item are formulated to indicate the level of a client's ability on a continuum from easy to difficult. In a Guttman scale, once a respondent fits a specific description on the continuum, he should also fit all the weaker or less severe descriptions of the item. It is a deterministic scale and does not allow for odd behaviors that could probably occur (Laver Fawcett 2007; Tennant & Conaghan 2007). Therefore if the data is on a Guttman scale, the Rasch model converts the ordinal data into a linear measure or interval data.

When data is converted into interval data, the one-dimensionality of items can be determined. All items are placed on a continuum according to their difficulty level. From this analysis one could determine if the items coalesce to form a single dominant construct (Furr & Bacharach 2008; Laver Fawcett 2007; Tesio 2003; Wright & Linacre 1989). This function will address the challenge that attributes in humans do not operate in isolation and that items that are one-dimensional contribute to one attribute. It is important to note that other analysis such as factor analysis could also be performed to address this issue, but only if the construct or attribute under investigation is firmly

grounded in a theoretical framework. This method is able to detect the hidden or covert structure in a series of measurements that purport to measure a particular trait. This implies that a researcher who develops a measuring instrument will include all possible items that contribute to a specific attribute according to theoretical frameworks and clinical experience. Statistical analysis will indicate redundant items but it can never suggest missing items (Clark & Watson 1995).

3.4 PSYCHOMETRICS

All measures or tests need to be investigated for their psychometric properties, namely validity and reliability. This is an important aspect in measurement which needs to be addressed in development of outcome measures.

Furr and Bacharach (2008) defined psychometrics as the science of evaluating the attributes of psychological tests. They mention three important attributes of tests, namely the type of data (the scores), issues concerning the validity of data and the reliability of the data.

3.4.1 TYPE OF DATA

All data could be classified as either discrete or continuous data. Discrete data exists independently of each other, e.g. diagnoses, institution or type of programme. Continuous data forms a continuum, e.g. millimetres, age and temperature.

Data is also classified into scales of measurement, that is nominal data (assigning data to a specific category e.g. type of programme), ordinal data (assigning data in order of sequence), interval data (difference in standard units, also in rank order) and ratio data (similar to interval but originates at zero).

During the late 1980's Wright and Linacre (1989) were concerned that researchers could confuse observation with measurement and as a result, treat data improperly. Tesio (2003) also warned that measurement does not equal counting, but an abstract continuum on a continuous linear measure (e.g. a ruler). Quantitative observations are based on counting and these counts can be ordered in amounts of the underlying variable e.g. never, sometimes, always. This is a typical ordinal scale of measurement. The problem lies within the implied amounts as the distances between the scales are

not equal and the score is not a number in a mathematical sense. Nominal data (classifying variables into categories e.g. race, gender, age categories) are also observations and counts but not a number in a linear (interval) scale. Svensson (2001) explained these improper uses of ordinal data in more detail.

Wright & Linacre (1989) stated that measurements can only be true measurements if they are numbers with arithmetic properties (i.e. that can be added, subtracted, multiplied and divided). This is called interval or ratio scale of measurement. Svensson (2001) and Wright and Linacre (1989) labeled data on a nominal and ordinal scale as qualitative and data on interval and ratio scales as quantitative data.

To do valid statistical analysis of data from observed behaviours or latent traits, one must progress from counting observations to measurement. This progression is not new, and well known researchers like Thurstone and Thorndike invented techniques for this conversion in the 1920s. George Rasch developed even more sophisticated techniques in 1953 and these techniques are constantly refined in current research with ordinal scales of measurement (Osborne 2008).

It is not only the type of data that influences the choice of statistical analysis but also the types of validity and reliability that are being investigated.

3.4.2 VALIDITY

3.4.2.1 CONTEMPORARY VERSUS TRADITIONAL PERSPECTIVES OF VALIDITY

Validity raises questions such as: What does a test measure and does it measure what it is supposed to measure? Furr and Bacharach (2008) challenged this definition and argued that it oversimplifies the issue of validity. They proposed a more contemporary definition that reads: “the degree to which evidence and theory support the interpretation of test scores entailed by the proposed uses” (Furr & Bacharach 2008, pp. 168).

This definition implies that the test, or the set of items or the scores, are neither valid nor invalid. It is the interpretation of the scores that is valid or invalid. The next issue in this definition is that validity is related to the proposed uses of the scores and validity is only “valid” for that proposed use. It would be unjust to extend the use of a standardised test to measure associated or related constructs because a test’s validity is only valid for the intended purpose and intended population.

The third issue in this definition is that validity is a matter of degree and not an "all-or-nothing" issue. Furr and Bacharach (2008) further explained that there is no threshold for validity and therefore interpretation of scores should be expressed as strong or weak evidence for validity instead of simply valid or invalid. Clark and Watson (1995) shared this view by stating: "one does not validate a test, but only a principle for making inferences".

Furr and Bacharach (2008) made a distinction between the traditional three-faceted perspective (content, construct and criterion validity) described by Polgar and Thomas (2008) and the contemporary perspective on validity. According to Furr and Bacharach's (2008) contemporary perspective, construct validity becomes the focus point of psychometric investigation. This perspective makes sense, as one could argue that if there is no clear construct, there cannot be a meaningful measuring instrument. Construct validity therefore not only has to be the focus of a psychometric investigation but also needs to be the first property to be investigated. Figure 3.1 presents the types of information that can be used as scientific evidence to support the validity of test interpretations. Note that construct validity is at the centre while the five types of information on the outside represents a mix of reliability and validity aspects.

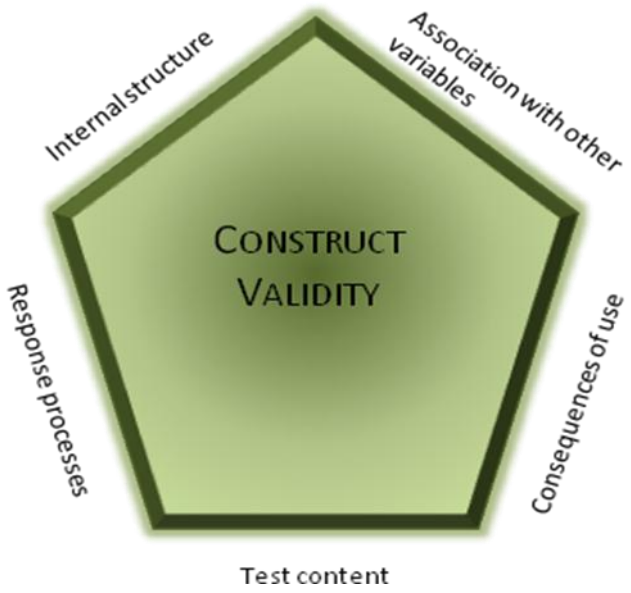

Figure 3.1 A contemporary perspective of types of information relevant to test validity. (Furr and Bacharach 2008, pp. 171)

Test content at the bottom of Figure 3.1 includes the aspect of content validity with specific reference to construct-irrelevant content and construct-underrepresentation. Face validity is also described as evidence of test content.

Response processes refer to the correct interpretation of instructions to use the test or, in case of self-report questionnaires, whether respondents gave true responses.

The internal structure of the test includes item correlation, how the structure matches the theoretical concept, if the items form clusters and whether dimensionality is single or multiple.

A structure is usually based on a theoretical concept or framework. When psychometric properties are being investigated, this theoretical concept must be included when the statistical analysis is interpreted. The crucial importance of the theoretical structure has been described as early as the 1950s by Cronbach and Meehl (1955). They described the nomological network of a construct and set out certain fundamental principles. One of these principles is that a construct occurs within a network of other constructs and the interrelationships should be presented in a nomological network. Some of these attributes in a construct needs to be observable but sometimes derivations could be made from remote observations. A nomological network visually represents the basic features of the concept, its observable manifestations and the interrelationships between them (Peterson & Zimmerman 2004).

When construct validity is being investigated the understanding of this network should be enhanced and thus the network could be expanded. New nomologicals emerge and enable the researcher to predict similar observations (Cronbach & Meehl 1955).

Although Furr and Bacharach (2008) did not refer to nomological networks, they do acknowledge the importance of the theoretical framework when interpreting findings from construct validity analyses. The internal structure of a test is dependent on a firm theoretical framework and associations with other variables should be clearly indicated.

Association with other variables can be investigated through convergent evidence. For example, the correlation with tests of related constructs can be investigated. This is similar to concurrent validity and discriminant evidence (non-correlation with unrelated constructs, also called divergent validity).

Consequences of use, also referred to as consequential validity (Lees-Haley in Furr & Bacharach 2008, p. 182) points to the fair (or unfair) and adverse effects of testing. Furr and Bacharach (2008) argued that if a test does not measure a construct equally well for the target population it is clearly an issue of validity. They gave the example of a test that predicts job performance where females score better than males; the researcher must have evidence that the test is measuring job

performance and that the items are not biased towards female characteristics or have an adverse effect on male subjects.

Face validity is sometimes considered as a weak type of validity. It is appropriate to use when many instruments for measuring a specific construct are available and one may choose from face validity which is the most appropriate (Brink, Van der Walt & Van Rensburg 2006).

Several authors classified criterion validity into concurrent and predictive validity (Laver Fawcett 2007; Payton 1994; Peat 2002; Seale & Barnard 1998). Concurrent validity is useful when two instruments measuring the same construct are available. It is administered to the same sample and the researcher could investigate a correlation between the two instruments.

Predictive validity is used to predict performance in future. A sample will be tested with a measure to establish how much of the criteria are met. Another assessment will be done at a later stage to measure the performance that was expected. If the measurement on the performance and the criteria correlate, the instrument has high predictive validity (Brink et al. 2006; Peat 2002; Seale & Barnard 1998). A typical example is intelligent tests. A subject must meet certain criteria for the different levels of intelligence. A certain level of performance (e.g. school performance) could be predicted using intelligence quotient.

3.4.2.2 ECOLOGICAL VALIDITY

One aspect of validity that Furr and Bacharach's (2008) contemporary perspective of test validity did not cover was that of ecological validity. Ecological validity has been discussed since the 1980's and refers to the relevance of tests' behaviour to the real world. In social sciences, ecology refers to the interrelationship between a person and his/her environment (Baker 1990). Brinberg and McGrath (1985, p. 138) defined ecological validity as the "extent to which a researcher can specify the scope and limits of a set of empirical findings with respect to the elements and relations selected from the substantive domain". Barrios (1988, p. 30) described ecological validity in terms of stability of obtained scores and states that a test will have high ecological validity if the same test has been administered in different settings and the scores show a high correlation. Polgar and Thomas (2008) discussed ecological validity as an aspect of external validity and refer to the difference in the situation in the test/intervention situation and real-life situations.

Tupper and Cicerone (1990) pointed out that internal and external factors influence everyday behaviour and performance in tasks. Therefore external and contextual factors need much more attention in test situations. Tupper and Cicerone (1990) defined ecological validity as the relationship between the results obtained in controlled experimental conditions and natural conditions, in other words, how the test performance corresponds to real-world performance. Sbordone and Long (1996) also pointed out that what a person is able to do in a simulated experimental environment is not necessarily what he/she can do in his/her natural environment.

All tests that relate to daily functioning or social aspects in a person's life should be ecologically sound. Ecological validity thus becomes a very important issue in occupational therapy since the profession's core business is about being able to perform the occupations that the environment demands of the individual; occupations that are both meaningful and provide satisfaction. Ecological validity has been taken seriously by neuropsychologists, but less so by occupational therapists. Numerous studies have been done in the field of activities of daily living and executive functioning by neuropsychologists. Van der Elst, Van Boxtel, Van Breukelen and Jolles (2008) found low to moderate correlations with three executive functioning tests, and argue that ecological validity in testing situations needs to be increased. Chaytor and Schmitter-Edgecombe (2003) highlighted the importance of adding test variables to simulate demands from the environment in order to enhance ecological validity. Their study demonstrated how the complex relationship between cognitive testing and real-world performance could be studied and better understood.

A study done on the cognitive skills of hospitalised patients with serious mental illness indicated that cognition in everyday tasks and common experiences implied ecologically sound results and that these types of testing should be relied on more often than traditional standardised cognitive testing (Heinrichs, Ammari, Miles & McDermid Vaz 2010; Thornton, Kristinsson, DeFreitas, & Loken Thornton 2010).

Studies done by occupational therapists to show ecological validity are scarce. The few studies that have been conducted showed good results. A study done by occupational therapists revealed high ecological validity in the Occupational Therapy Adult Perceptual Screening Test (Cooke, McKenna, Flemming & Darnell 2006).

Bottari, Dutil, Dassa, and Rainville (2006) emphasised the importance of context when assessing independence in everyday activities and urge occupational therapist to take context into consideration in standardised assessments.

An instrument to assess cognitive abilities and learning potential in children between 6 - 12 years was developed by Katz, Golstand, Bar-Ilan and Parush (2007) and showed good predictive validity. When an instrument has good predictive validity it could be an indication that it has assessed real-life variables (Chaytor & Schmitter-Edgecombe 2003; Moore et al 2007). This aspect of ecological validity has been referred to as veridicality by Franzen and Wilhelm (1996).

When developing instruments to measure outcomes, ecological validity has to be taken into consideration because the outcome of intervention in human behavior and social sciences is often only seen once the person has returned to his or her real-life situation. The implementation of ecological validity in the development of outcome measures has been reported in the literature. It is suggested that an outcome measure should reflect the satisfaction of the client with his or her real-life performance. This reflection could be done by the client or significant others that observe the performance in everyday life. While assessing performance, real-life situations should be simulated if assessment is impossible in the real-life situation (Franzen & Wilhelm 1996; Kielhofner 2006; Spooner & Pachana 2006).

Once instruments have been investigated for validity, reliability aspects should then be examined.

3.4.3 RELIABILITY

Reliability is the consistency with which a measuring instrument performs (Brink et al. 2006; Peat 2002; Payton 1994). Reliability is usually easier to establish than validity. However, if a measuring instrument is not valid there is little substance for investigating reliability. Although reliability is not a sufficient condition for validity (Leedy 1997), the one neither ensures nor precludes the other (Seale & Barnard 1998). Peat (2002) explained that poor repeatability (implying unreliability) cannot have good validity, but good repeatability does not guarantee good validity. However, validity will be higher if the instrument has good repeatability. Furr and Bacharach (2008, p.187) reiterated that: “Even though validity often requires reliability, the reverse is not true”.

Several reliability aspects can be investigated once validity has been established. These are stability, internal consistency, equivalence, intra- and interrater reliability.

Stability of an instrument refers to the extent to which the same results are obtained when the instrument is administered to the same sample twice. It is evaluated through the test-retest method for which a reliability coefficient will be calculated, statistically known as the correlation coefficient.

A coefficient above 0.7 is acceptable but coefficients of 0.85 and higher are preferable (Polgar & Thomas 2008).

Although stability is most appropriate for instruments that measure relatively enduring constructs such as personality and ability, there are instruments (specifically in outcomes measurement) that are developed to detect change in constructs such as motivation, self-esteem, participation and the like. Peat (2002:85) emphasised the importance of responsiveness of an instrument with outcomes measurement. Many instruments are inherently unresponsive to small changes in subjects. Peat suggested that in such cases the range of the scale needs to be extended by adding subcategories between main scores.

If therapists use their clinical reasoning and professional judgement about clients' progress or change, it should correlate with the correlation coefficient of the difference between the first and second assessment.

Laver Fawcett (2007, p. 202) refered to responsiveness as test sensitivity. She says that a sensitive test will minimise the chance of false negative results (when a person who has the deficit is not shown to have the deficit) and will be able to measure change due to intervention. The test must therefore target outcomes addressed in therapy. The test must be responsive to both the type and amount of change in function that is anticipated or desired as a result of intervention.

To calculate change, an effect size index is calculated. Effect size is a measure of the strength of the relationship between two variables and is ideal for outcome measurement; for example, the strength of the relationship between the base-line and final assessment. An effect size index of 0.2 or below is considered small, 0.5 is considered moderate and 0.8 or more a large effect size (Laver Fawcett 2007, p. 203).

An inconsistent classification of sensitivity and responsiveness was found in the literature. Laver Fawcett (2007) classifies test sensitivity as an aspect of reliability while Polgar and Thomas (2008) classified it as a validity issue. Clinically this should not have an influence on effectiveness of tests because the clinician should decide whether sensitivity is an issue. It should not really matter whether it is described as a reliability or validity aspect.

Internal consistency refers to whether persons in a trial respond consistently to the items in the test or measure. The Cronbach's alpha index is generally used to calculate the correlation coefficient for this consistency. A correlation of 0.7 and above indicates acceptable internal consistency while correlations above 0.9 could point to redundant items (Osborne 2008; Spiliotopoulou 2009).

Spiliotopoulou (2009) warned occupational therapists with regards to misconceptions about the use and interpretation of Cronbach's alpha. Sound interpretation of Cronbach's alpha will be enhanced if the researcher is cognisant of the following factors: the number of items in the test (large number of items could yield a high Cronbach's alpha), the width of the scale (wider scales could increase alpha), the nature of the data (nominal data is not suitable for Cronbach's alpha; rather use Kuder-Richardson), the sample size (small samples could yield large reliability coefficients), normal distribution and linearity (if not, Cronbach's alpha could underestimate the internal consistency of the test) (Spiliotopoulou 2009).

Equivalence is the term used when a test contains two similar sections or editions. A test user may then select one of the versions to administer. It is usually developed to counteract the threat of familiarity with a test. If a respondent scores the same on both editions of the test, equivalence is high (Furr & Bacharach 2008).

Interrater reliability refers to different observers or raters who use the same instrument to measure the same phenomena at the same time while intrarater reliability is a correlation between two or more ratings done by the same rater (Polgar & Thomas 2008).

Reliability thus deals with the property of reproducibility (Polgar & Thomas 2008). It is an important property in the measurement of outcomes since the minimum data collection points are at least two. This implies that the same measure will be applied at least twice on the same client and if the measure is not reliable, incorrect information will be used to make important decisions about the person's performance. Researchers and developers of outcome measures should thoroughly investigate the reliability aspects as discussed above.

Other concepts that may lead to confusion in research are internal and external validity. These two concepts warrant their own explanation as they may cause confusion in the realm of psychometric properties.

3.4.4 INTERNAL AND EXTERNAL VALIDITY

Internal and external validity are described in literature but should not be confused with the measurement properties of assessments or measuring instruments (Polgar & Thomas 2008). These two types of validity are not assessed when one investigates the reliability and validity of a specific measuring instrument. They are, however, applied in the interpretation of the results of the study.

In other words, psychometric properties will be included in research studies where instruments were developed and internal and external validity will be described in studies that investigate different variables (this includes almost all quantitative designs).

Internal validity is the freedom from bias in the interpretation of the results, that is, if the changes in the dependent variable may, with no doubt or bias, be attributed to the independent variable.

External validity refers to the extent to which the results may be generalised to the general population or subpopulations. When using mixed methodology a researcher could experience a problem in generalising the findings. During some phase of the research a researcher would have used qualitative methods to obtain data. Qualitative data does not lend itself to generalising and therefore external validity will be influenced. External validity is described as transferability in qualitative designs with qualitative data.

Polgar and Thomas (2008) classified external validity into population validity and ecological validity. Population validity refers to the generalisation of the results to the population from which the study sample was drawn. Population validity ensures that results are not generalised to populations with different characteristics to the study population. Ecological validity has been discussed earlier in this chapter and refers to the situation in which an investigation that has been carried out might not be accurately generalised to other situations.

Internal and external validity thus contribute to the validity of the findings and become important when the effect of service delivery is being investigated.

3.5 CONCLUDING REMARKS

This chapter discussed measurement principles and highlighted certain concerns with the measurement of human behaviour. Although some challenges exist, many tools and techniques are available to address the challenges. This information was helpful in the development of the outcome measure during Phase 2 of this study as well as with the interpretation of the data in Phase 3.

CHAPTER 4 RESEARCH METHODOLOGY

4.1 INTRODUCTION

The methodology comprised three phases. The aim for each phase was as follows:

Phase 1: Identifying the domains of the outcome measure

Phase 2: Designing and developing the outcome measure

Phase 3: Piloting the measure and examining selected psychometric properties.

The research approach and design for the entire study are set out in this chapter. The study objectives, participants who were involved in the study, data gathering methods and data analysis are described separately for each of the three phases. Chapter 4 concludes with a discussion of the ethical considerations of the study.

4.2 RESEARCH APPROACH FOR THE THREE PHASES

The researcher had several options with which to approach the challenge of developing an outcome measure. Use of a naturalistic approach as point of departure would delineate the context and meaning of the prospective measure in terms of professionals who intended to use it as well as potential clients who would be assessed by it. A participatory approach, in turn, would generate interpersonal responses from people participating in the real situation who, upon completion of the research, would continue to use the measure. A positivist approach, on the other hand, would provide the best external objective evidence of the measure's worth for the profession of occupational therapy and for clients. By settling for a pragmatic approach, a middle-of-the-road option, the researcher would create circumstance that would permit her to configure practical issues such as implementation, appropriateness and clinical utility during the conduct of the impending study (Babbie & Mouton 2001).

The researcher opted for a participatory approach to the study to address the research-practice gap (Taylor, Suarez-Balcazar, Forsyth & Kielhofner 2006). When researchers investigate problems in

clinical practice they have to be aware of the isolation that could be created between researchers and clinical practitioners. Researchers might not be aware of the important circumstances and constraints faced by clinicians. The development of an effective measure mandated scientific information from occupational therapists, based on their clinical experience, and from mental health care users about their personal experiences of occupational therapy services. Although the participatory approach is traditionally used in social science research with the intent to change social situations, it was often used to close the gap between research and practice. It was decided to apply this approach in a clinical setting where change in service delivery was clearly needed and where all the stakeholders had to become involved in order to maximise the chance of implementation in the clinical setting.

4.3 RESEARCH DESIGN FOR THE THREE PHASES

The researcher supplemented the participatory approach with a pragmatic approach that permitted her to merge the practical issues of an outcome measure (like clinical utility and psychometric properties) with the experiences of people from the research setting.

Whenever a compound approach, as was described above, is opted for, a mixed method design would yield the best strategic advantage: thorough investigation of each step or phase of the problem situation. A mixed method design is usually described as a diversified approach since it blends both qualitative and quantitative data. Creswell and Plano Clark (2007) suggested that where a researcher needed to explore a phenomenon and had not yet indentified appropriate variables, the application of an exploratory mixed method design could quickly point out relevant variables. The exploratory design, furthermore, specified the instrument development model for construction of measures or instruments such as an outcome measure. The research design for the entire study was thus a mixed method exploratory design that was aimed at instrument development.

Figure 4.1 presents the sequence of this research project according to the Instrument Development Model. Phase 1 constituted the qualitative steps of the research where the researcher relied on the participation of clinicians and mental health care users to achieve the research aims. Phase 2 entailed the development and designing of the outcome measure. This phase was a theoretical exercise where the researcher relied on previous research and literature reviews to generate item descriptions for the outcome measure. Phase 3 was the quantitative phase where the outcome measure was subjected to a trial and the psychometric properties were investigated.

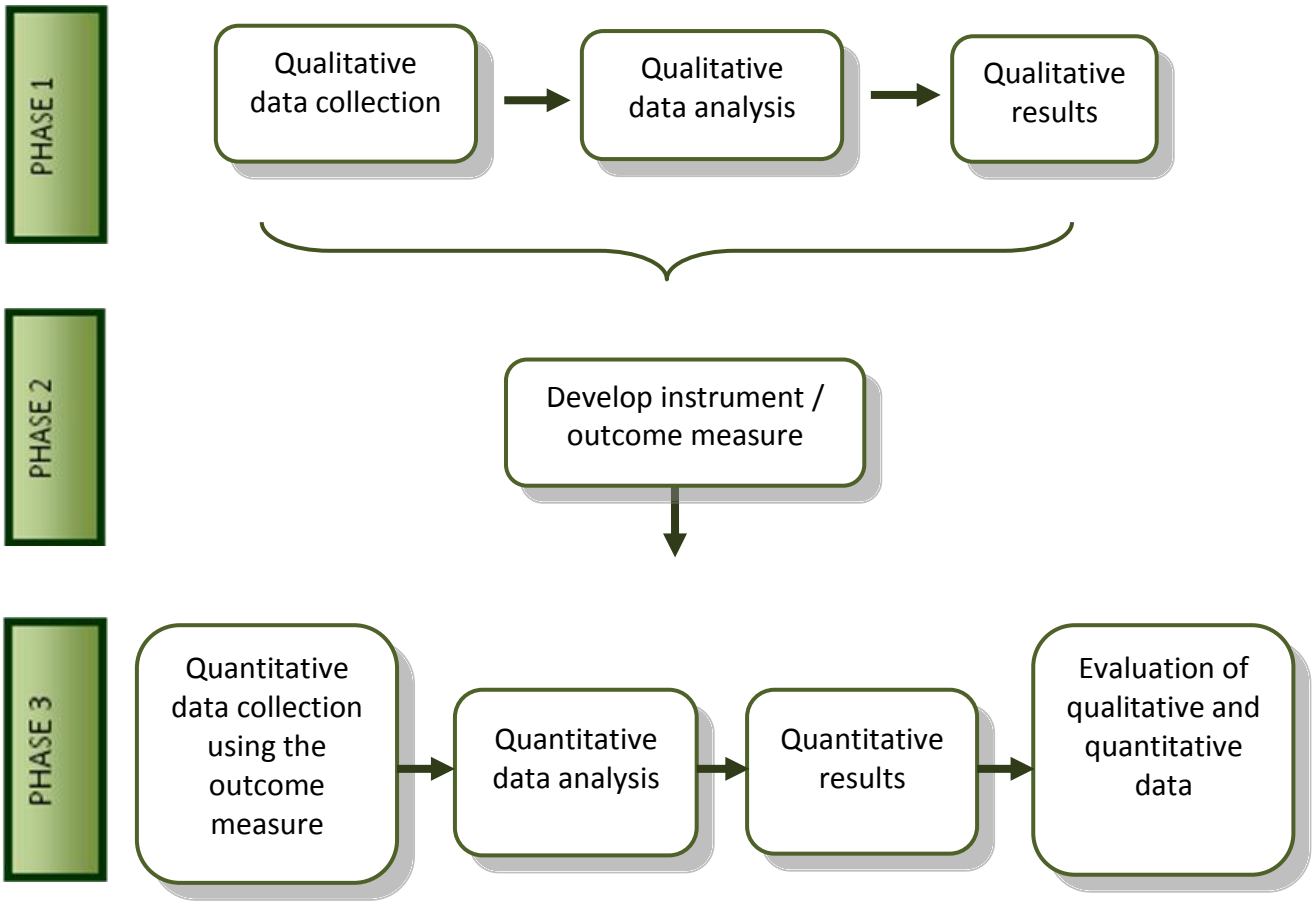

Figure 4.1 Mixed methods exploratory design: Instrument Development Model.

The Instrument Development Model guided the research through the three phases. Figure 4.2 gives an overview of the three phases of the research, methods used in the phases, the products that each phase yielded and the developmental stages in the instrument or outcome measure.

The methodology for each phase is next described in detail. Each section introduced the objectives of the particular phase, in an attempt to improve readability and understanding.

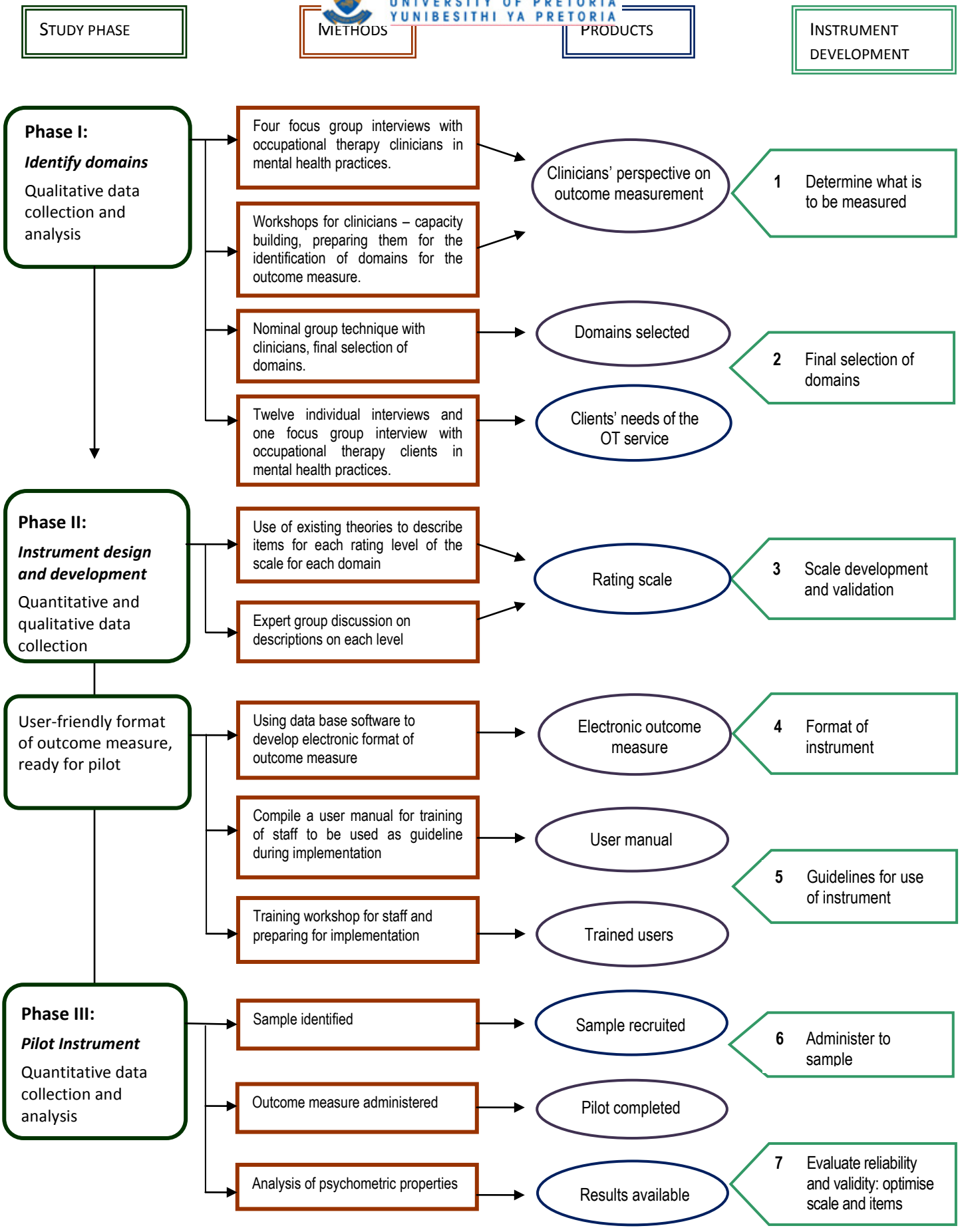

Figure 4.2 Phases, methods, products and instrument development steps of the study.

4.4 PHASE 1

4.4.1 AIMS AND OBJECTIVES FOR PHASE 1

The aim of the first phase of the research was to identify the domains of the outcome measure. Two objectives had to be met to achieve this aim: (i) establishing the views and perceptions of occupational therapy clinicians with regard to measuring of outcomes in order to identify relevant domains for an outcome measure; and (ii) evaluating mental health care service users' expectations about outcome domains selected by the clinicians.

4.4.2 PARTICIPANTS

4.4.2.1 RESEARCH POPULATION

The population to whom this study findings will apply are occupational therapists that deliver services to mental health care users in acute care, subacute care, long-term care and forensic services. These occupational therapists should use the Model of Creative Ability as their theoretical framework and be conversant and literate in English. The types of diagnoses of the mental health care users may include any of the disorders in the DSM IV. The severity of the illness may range from acutely psychotic or disorientated to stable and fully coherent. The age group of the mental health care users should be older than 15 years and could be from any cultural or socio-economic background.

4.4.2.2 OCCUPATIONAL THERAPY CLINICIANS

This phase of the study was initiated by selecting occupational therapy clinicians who were able to give rich and informative data on measuring outcomes. The clinical placement settings that were being used for student training at the University of Pretoria were judged as a fair representation of the scope of established occupational therapy mental health services. Occupational therapists from five mental health care settings in and around Pretoria were asked to participate. During personal discussions with the researcher, clinicians at these settings indicated that their departments would benefit from an outcome measure and that they were eager to participate in the study.

A situational analysis intended to assess the current state of knowledge and experience of the participating clinicians with regard to measurement of outcomes in occupational therapy mental

health practices, preceded determination of the domains. All participants came from mental health practices that delivered different services e.g. acute care (two to three weeks), subacute care (three weeks to three months), long-term care (more than three months), and forensic services. Qualified occupational therapists, each with unique personal experiences, coming from various training centres throughout South Africa, partook in this phase. Acknowledging the different backgrounds and service settings of the participants, it was imperative to go through a process where all could understand each other's views and perceptions of measuring outcomes.

Occupational therapy clinicians from the following settings were invited:

- Weskoppies Hospital (acute, subacute, forensic and long term care)
- 1 Military Hospital's psychiatric unit (acute and subacute care)
- Witbank Hospital's psychiatric unit (acute and subacute care)
- Denmar Psychiatric Clinic (acute and subacute care)
- Vista Psychiatric Clinic (acute and subacute care)

Clinicians, who had worked at least one year in a mental health setting, was eligible for inclusion. The research mandated a convenience sample since all participants that fitted the inclusion criterion above, were included. They had to respond to specific research question, namely what their views and perceptions about outcomes were and what outcomes they considered suitable for inclusion in an outcome measure. The researcher aimed to generate a wealth of detail from all the available participants from the above mentioned settings.

The researcher personally contacted every setting and invited clinicians who met the sampling criterion to attend the first focus group on a date negotiated between the clinicians and the researcher. Sixteen clinicians accepted the invitation. Two first-round focus groups were arranged to accommodate the number of participants.

4.4.2.3 MENTAL HEALTH CARE USERS

A purposive sample using the maximum variation sampling technique was applied to select mental health care users. The inclusion criteria stipulated a variation in age, gender, length of stay and attending different types of programmes at the occupational therapy department. These criteria implied that the mental health care users had to be familiar with occupational therapy and regularly attend the programme in order to give valuable information. They also had to be able to provide reality-based answers to questions put to them, thereby ensuring access to rich information. Mental

health care users who were interviewed individually were recruited from one institution. Clinicians at this institution provided the researcher with names of appropriate participants.

The individual interviews were followed up by two focus groups that verified the presence of data saturation. Mental health care users for the focus groups were selected from two other mental health care settings using the same inclusion criteria explained above. Clinicians from these settings identified and assigned suitable participants.

4.4.3 DATA GATHERING METHODS

4.4.3.1 FOCUS GROUP DISCUSSIONS

Focus groups are acknowledged means of eliciting information (De Vos, Strydom, Fouche & Delport 2005). The method originated in the early 1960's and have been extensively used by social scientists since the 1990's. Schurink and Poggenpoel (1998, p. 315) described it as a group of between eight to ten people purposively discussing a topic of concern or common interest. In practice it is conducted as an open but focused conversation, usually in a series to generate confirmatory data and ensure its saturation.

The advantages of focus groups are numerous. They are rapid and cost-effective means to gather data and useful when little is initially known about a specific situation or topic. The researcher directly interacts with participants and can thus probe for deeper meaning of specific views and perceptions. Participants are allowed to expand the responses of others in order to increase the generation of rich and comprehensive information about a topic.

The disadvantages of focus groups are that data-based findings cannot be generalised because the participants usually are part of a purposive sample or one of convenience and as a result are not necessarily representative of the population being studied. The qualitative and diverse nature of the data could also make it difficult to summarise.

A moderator facilitates the discussion of a topic or concern. The moderator does not participate in the discussion but encourages interaction between participants and allows the discussion to flow naturally. It is important to listen openly and intensely in order to keep the discussion within the boundaries of the topic without limiting response. The moderator, interacting with participants to

probe deeper meaning, has to remain non-authoritarian and non-judgmental throughout the focus group.

The moderator usually prepares an interview guide to subtly direct the discussion (refer to Appendix A for the interview guide for the focus groups). If the first question of the focus group is open, non-threatening and inviting, discussion should flow naturally.

Sixteen occupational therapy clinicians volunteered to participate in the intended study and they were divided into two groups of eight each. The groups were kept small in order to encourage participants to contribute maximally. Each focus group was planned to last two hours. It was important to pose the same opening question to both groups, and the interview guide was therefore essential for consistency.

The interview guide covered questions on the knowledge and attitude of clinicians about outcome measurement, existing measurement systems in use (including assessment methods and techniques), and various needs for an outcome measure. These questions were piloted with a colleague and the feedback was implemented. The original direct opening question was converted into an open statement with an invitational tone: “Let’s talk about outcome measurement”.

Data saturation occurred after each focus group participated in two rounds of discussion. Participants expressed a need for more information on available outcomes to assist them in identifying and selecting domains for an outcome measure during the second round focus group discussion. All the participants agreed to attend a workshop that addressed these needs. The researcher conducted the two workshops, as requested. Some participants found the dates inconvenient and the workshop, therefore, was repeated.

The content of the workshop covered the most recent trends in mental health practices, an update of philosophical, theoretical and practice frameworks as well as existing outcome measures. Towards the end of the workshop participants felt equipped to identify domains for an outcome measure. Step 1, 2 and 3 of the nominal group technique as described below formed part of the workshop.

4.4.3.2 NOMINAL GROUP TECHNIQUE

The nominal group technique (Lloyd-Jones, Fowell & Bligh 2002; Willcox & Zuber-Skerritt 2003) was utilised to determine the domains for the outcome measure. This technique starts off with a focal question using brainstorming. A public list of all responses that the participants had contributed is next compiled. The third step involves discussion and clarification of similarities, duplications or

unclear statements. During step four each participant prioritises and selects the top three statements from the public list. The final step is to rank the chosen statements in order of priority.

The focal question posed to the participants in this study was: "What are the domains that you wish to include in an outcome measure for your practice?"

Step 1 (10min)

Individual brainstorming in writing followed. Each participant received a small booklet with 20 pieces of paper (8cm x 8cm). They had to write each outcome on a separate piece of paper. The number of outcomes per participant was unlimited. (They could request more paper if necessary).

Step 2 (20 – 30 min)

A public list (on a flip chart) was compiled after a round-robin collection of ideas. Each participant received a turn to nominate one domain; if someone else had already mentioned it, the participant chose another one from her list. No criticism or judgement was allowed during this step.

Step 3 (30 – 45 min)

Hereafter discussion and clarification of public statements/domains started. Participants could ask questions about the domains. Duplicated domains were eliminated while others were renamed to enhance clarity of understanding. After the workshop, a final list of the domains was compiled and distributed to all participants via e-mail. The final list consisted of two columns: Column A contained a domain, e.g. process skill, and column B contained all the aspects for the domain, e.g. attention, pace, adaptation. (See Appendix B for the list that was distributed).

Step 4 (via e-mail)

Each participant was asked to select the three most important items/domains. Participants were asked to select from column A and then mark the aspects in column B which they viewed as relevant.

Each participant had to rank their three chosen domains (A = priority 1, B = priority 2 and C= priority 3).

Step 5 (ranking)

After receiving the selections of the participants, the domains were counted and weighted. Weighting of domains was done by assigning 3 to all As, 2 to all Bs and 1 to all Cs. The list was then reordered to reflect order of priority.

4.4.3.3 INDIVIDUAL INTERVIEWS AND FOCUS GROUPS WITH MENTAL HEALTH CARE USERS.

Interviews with mental health care users were conducted. An interview guide was prepared and used with volunteering clients. See Appendix C for this interview guide. The first interview with each client served as a trial interview. It was audiotaped to permit the researcher and a colleague to co-jointly review questions and answers obtained from the interview, at a later stage. Some questions were rephrased or eliminated after review. Twelve health care users were individually interviewed. The individual interviews continued until no new themes or categories emerged from the data. Two focus group interviews were held with nine mental health care users at a second and third institution, to confirm the data from the individual interviews and to check if new data emerged from participants. Three users who had participated came from one institution and six from another.

Domains were inferred from this information and were then compared with the clinicians' selection of domains.

4.4.4 ANALYSIS OF THE DATA

4.4.4.1 FOCUS GROUP INTERVIEWS WITH CLINICIANS

The content of the focus groups was transcribed verbatim, categories were identified and themes were created. Thematic content analysis (Green & Thorogood 2004) was used to categorise common themes. The key elements in each participant's versions were compared with those of all other participants and then classified into an existing category. A new category was created if the key element did not fit in an existing category.

4.4.4.2 NOMINAL GROUP TECHNIQUE

In Step 5 data generated by the Nominal Group Technique was analysed. The frequencies of responses were calculated where after and the weighting of each domain was determined.

4.4.4.3 INTERVIEWS WITH MENTAL HEALTH CARE USERS

The interviews with mental health care users were transcribed verbatim. Categories and themes that they came up with were compared with the domains that the clinicians had generated. The specific data set was examined to track additional or new categories or themes that might have emerged. The focus was on the content or range of the data and not on the frequency of responses: in other words, the number of times a theme or category was mentioned was not important.

4.4.5 TRUSTWORTHINESS OF DATA

Subjective meanings, experiences and perceptions from participants in this first phase of the research formed the starting point of the scientific inquiry. Subjective interpretation, however, could compromise data and make it unreliable and invalid, or in qualitative terms, data that is neither plausible nor trustworthy. Krefting (1991) suggested four strategies to establish trustworthiness in qualitative enquiries: credibility (internal validity in quantitative terms), transferability (external validity), dependability (reliability) and confirmability (objectivity). These strategies were applied during different stages of the research e.g. in the course of the research design, data collection and data interpretation. Application of the four strategies to ensure true reflection and presentation of the data that emerged from the first phase of the study is explained as follow.

4.4.5.1 CREDIBILITY

Identification of recurring patterns required spending adequate time with participants. The researcher had to submerse herself adequately in the research setting in her attempt to identify and verify response patterns that reflected the true circumstances and established confidence in the data. Krefting (1991) reminded researchers that the truth value is subject-oriented but, nevertheless, it is the researcher's responsibility to present the truth value as it is. The following methods were used to ascertain the presentation of the truth during the data collection.

Lincoln and Guba (1985) introduced the phrase "prolonged engagement". It refers to the researcher's need to ensure intimate familiarity with the research setting so as to permit the discovery of hidden facts (Anthony, Onwuegbuzie & Leech 2007a). It was often noted in qualitative

studies that as time went by, participants offered more sensitive information. In the current investigation the researcher engaged the research setting for an extensive period of two years since she first accessed it with the intention to study outcomes. She immersed herself in the research process by paying regular visits to settings, by having informal discussions with clinicians and students who did their training at the settings and often returned to supervise students in training. By the time the focus groups and interviews had started, the researcher valued and sometimes even identified with the comments from the clinicians and mental health care users.

Krefting (1991) cautioned researchers against the social desirable responses that participants often give, instead of relating their personal experience. This can also be countered by prolonged engagement in the setting. Questions sometimes had to be reframed to elicit the exact perceptions or experiences from participants.

The researcher's influence on the results of a qualitative inquiry must be acknowledged and explained. Reflexivity is an effective way of revealing the researcher's own background, perceptions and interests in the process. Even the relationship between the researcher and the participants needs to be explained. For this reason a vignette about the researcher has been included (see Appendix D).

As the researcher became more enmeshed in the research setting, she had to be careful not to lose the ability to interpret the findings and constantly had to separate her own ideas from those of participants. She did this by keeping a self-reflexive journal where she wrote up her perceptions and interpretations after each data gathering session.

Member checking prevents misinterpretation of the data (Anthony et al. 2007a). After completion of the focus groups, the researcher transcribed the data and partitioned it into themes and categories. The researcher presented the preliminary interpretations to the participants at the workshop to check if it represented their experiences and perceptions about outcomes. Participants used the opportunity to clear misunderstanding and verify some of the wording of certain themes and categories.

4.4.5.2 TRANSFERABILITY

Krefting (1991) mentioned that transferability might not necessarily be an issue depending on the purpose of the research and its situational uniqueness. If a case study, for example, had been investigated to understand the dynamics of that case but for no other purpose, transferability becomes irrelevant. In the current study, transferability was an issue since outcomes that were to

be measured in engaged mental health care settings preferably had to be appropriate for and relevant in other mental health care settings. Transferability was ensured through assembly of a representative sample of participants, and by drawing up a thorough profile of the participants (Anthony et al. 2007a; Anthony, Onwuegbuzie & Leech 2007b).

Participants were able to give information-rich data on what needed to be included in an outcome measure. Participants, after all, worked in mental health care settings and were experienced in whatever was at issue. As was said earlier, almost all mental health care settings, at the time of data collection, had already been invited to participate in the study.

The onus to provide an index of transferability did not rest on the researcher but it, nevertheless, was his or her responsibility to provide adequate information for outsiders to decide if findings of a study were transferable to their own situations (Lincoln & Guba 1985). By providing a comprehensive profile of the sample, namely the occupational therapy clinicians, the settings they represented as well as the mental health care users, external stakeholders could assess the transferability of the findings to their own situations.

4.4.5.3 DEPENDABILITY

Dependability refers to the consistency of the findings and repeatability of the study (Anthony et al. 2007a). The use of the code-recode procedure added to dependability of this study. During the analysis phase, the themes and categories were coded. The researcher went back to the analysis after a few weeks to see if she would still code the themes in the original categories. The participating clinicians were consulted again and some categories were recoded.

4.4.5.4 CONFIRMABILITY

Confirmability of a study indicates the degree of bias present in a study. Due to the nature of qualitative studies, the researcher has to be subjectively involved in the entire process, and is expected not to take an objective or neutral stance. Lincoln and Guba (1985) pointed out that neutrality in a qualitative study shifted from the researcher to the data set. Instead of the researcher being objective, the rigor lies in the data being neutral. The findings are solely those of the participants and any involvement of the researcher is clearly explained.

Triangulation was used in this study to support the confirmability or neutrality of data (Guba 1981). Obtaining data about what to include in the outcome measure from occupational therapy clinicians

as well as mental health care users enhanced the neutrality of the data. Informal discussions with experts during colloquial seminars and congresses were used to confirm some aspects of the data.

Reflexivity (Krefting 1991) was used to further support confirmability. The researcher kept a research journal in which she noted her personal experiences, perceptions and feelings after data gathering sessions. Neutral interpretation of the data was achieved by acknowledging subjective involvement and discussing this with research experts.

The matter of trustworthiness primarily related to the qualitative phase of the study. Since the study was a mixed method design, trustworthiness or internal and external validity of the other phases will be explained in following sections.

4.5 PHASE 2

4.5.1 AIMS AND OBJECTIVES FOR PHASE 2

Once the domains of the outcome measure were identified in Phase 1, Phase 2 commenced with the intent to design and develop the outcome measure.

The strategy for constructing and developing outcome measures was described in the literature review and applied in the prospective research. It is summarised in Figure 4.2, the last column labeled instrument development, commencing with *Determine what is to be measured* and ending with *Evaluate reliability and validity*. Each step in this instrument development strategy will now be described.

4.5.2 DEVELOPMENT OF THE OUTCOME MEASURE

4.5.2.1 DETERMINE WHAT IS TO BE MEASURED

As set out above, the first two objectives of Phase 1 of the study focused on what was to be measured. This was done in collaboration with occupational therapists and mental health users in

focus groups and with occupational therapy clinicians at workshop. Mental health care users were also individually interviewed. The final product of these data gathering sessions was a list of potential outcome domains.

4.5.2.2 FINAL SELECTION OF DOMAINS

The original list of possible outcome domains was overwhelming: it was impossible to construct an outcome measure with so many domains. The nominal group technique was used to guide the clinicians in prioritising the initial contributions.

4.5.2.3 SCALE DEVELOPMENT AND VALIDATION

Domain outputs derived from individual interviews and focus groups with mental health care users were compared with those chosen by the clinicians with the intention to confirm domains or items for the outcome measure.

The next step in the development of the outcome measure was to construct and validate a dedicated scale that measured each domain. The ultimate objective of scale development was to come up with a valid measure of a specific construct. Clark and Watson (1995) emphasized that it was essential to begin with a clear conceptualisation of the target construct. The initial content preferably had to be overinclusive, not limited to the needs of an institution, based on a particular individual's experience and ideas, or consisting of domains that are easy to measure. The entire range of outputs from the target population had to be represented in the domains. Overinclusion was necessary because techniques of data analysis could pinpoint weak or irrelevant items while they, simultaneously, were unable to detect items that should have been included.

One needed to take cognisance of the theoretical framework(s) of assessment and treatment procedures during outcome measurement (Law et al. 2001a). Theoretical frameworks are important indicators of key outcome measures. For example, if the framework was based on Vona du Toit's Model of Creative Ability (Du Toit 2004), volition and action as expressed in activity participation, would be two of the key measures. A theoretical framework, such as that of Creative Ability, could further contribute to a sound outcome measurement system since this model describes a person's activity participation and occupational performance in consecutive levels. Each level gives a detailed description of characteristics expected from a person at that level. These levels were used as the consistent rating scale across different domains and items for the outcome measure.

4.5.2.4 OPERATIONALISING THE DOMAINS

Each domain of an outcome measure was studied theoretically and operationalised by breaking it up into observable and measurable components. Items that represented the domain were phrased. This output made the domain observable and measurable. Each domain and its subset of items were meticulously delineated to permit domain differentiation and promote clarification of the meaning attributed to a specific outcome measure. Discussions in focus groups and theoretical definitions were used to identify domains and describe the items.

Many factors were considered in formulating item descriptors. Meticulous grammar, phrasing and semantics ensured precise formulation of a specific level of creative ability and representativeness of a particular item. The outcome measures were specifically developed for therapists who had knowledge and experience of the Model of Creative Ability, an issue that mandated item descriptors that were not cluttered by assumptions, constructs or principles of the theory. Item descriptors served as cues at a particular level and thus to be brief and to the point.

4.5.2.5 FORMAT OF INSTRUMENT

The instrument or measure was specifically developed for use by occupational therapy clinicians to measure outcomes of their intervention programmes. The format had to be commensurate with their level of knowledge and field of clinical practice. The format had to be user-friendly and explanatory, yet sufficiently practical for use with all types of clients in a mental health setting.

4.5.2.6 GUIDELINES FOR USE OF THE INSTRUMENT

A user manual was compiled that guided clinicians in the use of the outcome measure. The manual also had to facilitate the training of the clinicians. Its contents had to be comprehensive: the user manual had to illustrate the principles of outcome measurement, explain the difference between assessment and outcome measurement, detail procedures required for using the outcome measure, instruct users on how to determine a specific rating for a client, and define the various domains and items.

A four-hour training session was planned to explain the user manual. Adequate time for questioning and further explanation were allocated at the end of the training session. Clinicians were asked to use the outcome measure on five of their mental health care users.

4.6 PHASE 3

4.6.1 AIMS AND OBJECTIVES OF PHASE 3

Phase 3 required application of the instrument with the intention to identify clinical utility problems, investigate aspects of its validity and reliability, assess the sensitivity of the outcome measure to detect change and optimise the items where necessary.

4.6.2 ADMINISTER TO SAMPLE

Eleven participating clinicians were asked to recruit five mental health care users each and subject them to the outcome measure. The researcher was present whenever they assessed users so that she could respond to questions with regard to descriptions in the measure. The researcher also noted problems related to clinical utility, e.g. instructions for use, time required to complete the outcome measure and the effectiveness of the electronic format.

4.6.2.1 THE SAMPLES

Different samples and sample sizes were chosen for different types of validity and reliability estimation. Six occupational therapists were invited to judge content validity while the rest of the investigations were done on the performance of the mental health care users. The sample for intra- and interrater reliability was five clinicians from the setting who participated in the piloting of the outcome measure. Table 4.1 below indicates the sample size for each property that was investigated.

Table 4.1 Sample sizes for the validity and reliability investigations.

PSYCHOMETRIC PROPERTY	THE SAMPLE (UNIT OF ANALYSIS)	SAMPLE SIZE
Content validity	Occupational therapists acting as subject matter experts	6
Construct validity	Mental health care users' performance rating on the outcome measure	41
Interrater reliability	Mental health care users' performance rating on the outcome measure as rated by five clinicians	5 clinicians 1 mental health care user, 2 measurements
Internal consistency	Mental health care users' performance rating on the outcome measure	41
Sensitivity	Mental health care users' performance rating on base-line and final assessment	31

4.6.3 EVALUATING VALIDITY AND RELIABILITY: DATA COLLECTION PROCEDURE

4.6.3.1 CONTENT VALIDITY

Six experts on the subject matter of the Model of Creative Ability were asked to judge the relevance of each item and its descriptions in terms of overarching domains of the outcome measure. The experts had to assign a value of between 1 and 5 to each item to indicate their judgment of the relevance of the item to its domain. Each description was then judged for accuracy at the specific level of creative ability. The experts were asked to rewrite or add to descriptions where they did not agree.

Five of the experts had more than 30 years of experience and one had 10 years of experience in the Model of Creative Ability and also had been involved in the training of students.

4.6.3.2 CONSTRUCT VALIDITY

Clinicians attended two training sessions prior to the implementation of the outcome measure. They also received a training manual with relevant information for reference purposes during the data collection phase. The clinicians hereafter applied the outcome measure to collect data from the mental health care users. Base-line assessments were used to investigate construct validity.

4.6.3.3 INTRA- AND INTERRATER RELIABILITY

Data for interrater reliability was acquired from five clinicians who rated a single mental health care user that they knew well. A base-line measurement was collected, followed by second measurement five months later, with the same clinicians and health care user.

The same data set for the interrater reliability was used to calculate intrarater reliability.

4.6.3.4 INTERNAL CONSISTENCY

The data set used in the construct validity investigation was also used to investigate internal consistency.

4.6.3.5 SENSITIVITY

It was important to assess sensitivity to detect change in activity participation between the base-line and follow-up measurements. The effect size was calculated for 31 subjects who were rated by clinicians at two data collection points, namely the base-line and final assessment. The data set contained both a base-line and final score for each subject. This data set was similar to the construct validity data set, except that the data set for construct validity consisted only of the base-line assessment.

4.6.4 DATA ANALYSIS

The Table below sets out the statistical analysis performed for each property that was investigated.

Table 4.2 The statistical analysis per psychometric property.

PSYCHOMETRIC PROPERTY	THE SAMPLE (UNIT OF ANALYSIS)	ANALYSIS
Content validity	Occupational therapists acting as subject matter experts	Content validity index Qualitative amendments to the descriptors
Construct validity	Mental health care users' performance rating on the outcome measure	Principal Component Analysis
Intra- and Interrater reliability	Mental health care users' performance rating on the outcome measure as rated by five therapists	Correlation coefficients
Internal consistency	Mental health care users' performance rating on the outcome measure	Cronbach Alpha coefficient
Sensitivity	Mental health care users' performance rating on base-line and final assessment	Cohen's difference and t-test for paired observations

The contents of Tables 4.1 and 4.2 warrant explanation. The research design selected for this study was an exploratory mixed method design and required both qualitative and quantitative analysis. Combined (mixed) applications of the two opposing methods to data analysis in studies serve different scientific purposes. The different approaches to data analysis often might supplement one another, such as one generating research data that the second approach requires. In other applications the approaches perform a confirmatory function by using data generated by one approach to validate research data generated by an alternative approach. In a third research design simultaneous application of qualitative and quantitative approaches to data analysis initially might appear to clash and even result in radical differences of opinion among scientists. The latter assumptions hold for the current investigation (Giorgi 1985; Tabachnick & Fidell 1989).

An experienced quantitative researcher would hesitate to accept statistical calculation generated by an advanced multivariate statistical method (factor analysis mentioned in Table 4.2) on a limited sample (N = 41, as mentioned in Table 4.1). The quantitative examiner instantly will be conscious of the fact that the preconceived statistical criterion of having at "… least five cases for each observed variable" (Tabachnik & Fidell, 1989, p 603) has not been met. The qualitative researcher has a different opinion on this issue. Giorgi (1985, p 23), in explaining phenomenological thinking, set out

some of the principles on which quantitative and qualitative analysis differ, by saying that it: "is intrinsically difficult, since it goes against the natural tendency of consciousness to go toward things rather than its own processes ….". The quantitative researcher's natural tendency would be to become aware of the "thing" (statistical criterion) at the expense of the "noteworthy phenomenon" (the research data generated by the factor analysis) that is being investigated. In contrast, the qualitative researcher initially would consider the results of the factor analysis as a naïve data set that necessitates further reduction to convert it into meaningful data. Any qualitatively generated data set, in its initial format is embryonic. The embryo, through further data gathering that increases the sample size, will in due course evolve into a fully-fledged data set capable of yielding excellent research outcomes that can expand the discipline's body of scientific research knowledge (Giorgi 1985).

Presentation of factor analytic statistics based on a naïve data set, is solely intended to encourage the current researcher, other South African researchers and renowned scientists abroad, to accept the challenge of applying the innovative outcome measure in different contexts, with the intention to generate further data that, when pooled with the naïve data set, will convert the latter data set into mature research data. This "devious" approach has one other notable benefit. The qualitative analyst, as professional, will also develop and mature scientifically and better understand the psychosocial dynamics that underlie the outcome measure's factor analytic structure as it evolves from its embryonic structure into statistically structured scientific information of excellence.

4.6.5 OPTIMISE THE SCALE AND ITEMS OF THE OUTCOME MEASURE

Once the results of the psychometric investigation were available, the final step in the research process was activated. The researcher identified what changes were required to optimise the scale. A decision had to be made on possible removal of items or domains from the scale and adjustment of instructions in the training manual.

4.7 ETHICAL ISSUES CONSIDERED

Three fundamental ethical principles guided this research: respect for persons, beneficence and justice (Brink et al 2006).

Respect for persons was evident by respecting their right to self-determination, privacy, anonymity and confidentiality. Consent to become involved in the study was obtained from hospital management (Appendix E1), occupational therapy clinicians (Appendix E2) and mental health care users (Appendix E3). Potential participants were given the choice to participate or decline the request. The researcher undertook to manage data responsibly and ensure confidentiality and anonymity at all times.

Participants were protected against discomfort and harm by the researcher's adherence to the principle of beneficence. Although the nature of the study and type of data required did not expose participants to physical harm, caution was taken in the focus groups to ensure that questions asked and interviews did not cause emotional discomfort.

The principle of justice was exercised by acknowledging the participants' right to fair selection and treatment. All participants were selected for reasons directly related to the research and were not recruited and empirically manipulated because they could offer favourable information. Participants were treated fairly, no promises were made and their privacy was respected.

The Ethical Committee of the University of Pretoria and the Pretoria Academic Hospital gave ethical approval prior to the commencement of the study (Number: 118/2005, refer to Appendix E4). Additional ethical approval was obtained for the Human Research Ethics Committee of the University of the Witwatersrand (Number: M091025, refer to Appendix E5).

4.8 CONCLUDING REMARKS

A mixed method exploratory design, specifically the instrument development model, was chosen to guide this study. Development of the instrument progressed through three phases. Domains for the outcome measure emerged from Phase 1. The development of the outcome measure and descriptions for observable behaviours took place during Phase 2. Phase 3 subjected the instrument to a trial to investigate selected psychometric properties of validity and reliability. The identification

of issues to be optimised for the final implementation of the outcome measure was also addressed during Phase 3. Final routine implementation was not planned to be part of this study but would continue as a new study.