CHAPTER 7

RESEARCH DESIGN AND METHODOLOGY

7.1 Introduction

This chapter addresses the research design and describes the research methodology employed in this study. The sample and sampling procedure is discussed, the measures used are briefly described, and the translation and administration procedures are outlined. Finally the methods used to analyze the data are described.

7.2 Primary Objective of the Research

Postpartum depression (PPD) is not uncommon – with up to 20 percent of all mothers, in all circumstances suffering from this type of depression. PPD is not always easy to identify without screening measures and may develop slowly any time during the first year of the baby's life. Every mother is different and may have a different combination of symptoms. Some may be more anxious or irritable than sad. It may be mild or severe. Some mothers have been depressed ever since the pregnancy, and sometimes "The Blues" just don't seem to go away. Some mothers manage well initially and then their mood becomes darker and darker. If untreated, it can adversely affect a mother's functioning as well as her infant's development. Screening all mothers after birth is therefore very important to ensure that they get the necessary help and support they need. With this in mind, the primary objectives of this research were to:

- Address the problem of the unavailability of suitable PPD screening measures for the majority of Afrikaans-speakers by providing an Afrikaans version of an existing PPD screening measure – the PDSS - that was developed for use with an American culture and that has not been standardized on a South African population;
- Determine the validity and the reliability of the PDSS and the Afrikaans PDSS for English and Afrikaans speaking South African mothers based on the Rasch measurement model procedures;
- Determine how well the PDSS, EPDS and QIDS correlate when used with a South African sample;
- Determine the magnitude of the relationship between a positive screen for PPD in a South African sample and known risk factors for PPD.

7.3 Research Methods and Designs Used in the Study

Both qualitative and quantitative methodologies were used. Qualitative analysis was performed using two translation techniques, namely Brislin's back-translation method advocated by Brislin (1970) and the committee approach. It is, however, considered unlikely that any one result can provide unequivocal evidence for such linguistic equivalence of a test (Kline, 1993). Rather, a whole series of results can build up a composite picture, which overall could demonstrate the equivalence of a test. With this in mind, various quantitative methods from the Rasch rating scale measurement

model were also used to examine the validity and the reliability of the Afrikaans PDSS (Linacre, 2009).

7.3.1 Multiple translation method: Brislin's back-translation method and the committee approach.

Brislin's back-translation method together with the committee approach was used in this study to qualitatively explore the linguistic equivalence of the PDSS and the Afrikaans PDSS. The multiple translation method was also used to translate the QIDS into Afrikaans. Brislin, Lonner, and Thorndike recommend that a multiple translation method be used to ensure semantic equivalence (as cited in Beck et al., 2003). The back-translation method involves the translation of items from the original into the target language by a translator. This material is then translated back into the original language by another translator. The original version and the back-translation are compared to determine the efficacy of the translation.

The back-translation method has been demonstrated to be especially useful in cross-cultural research for checking the equivalence of the translations of measures in different languages (Bracken & Barona, 1991; Prieto, 1992). The back-translation technique has been used successfully to translate from English to Afrikaans (e.g., Shillington, 1988).

A committee (or cross-translation) approach involves a group of experts (such as cultural, linguistic, and psychological) in preparing a translation (Nasser, 2005; Van de Vijver & Tanzer, 1997). The committee members discuss the instrument's questions with

each other during a collaborative effort to improve the quality of the translation, minimise bias, and reduce misconceptions (Ægisdóttir et al., 2008; Onkvisit & Shaw, 2004). Researchers often combine the committee approach with the back translation technique (Van de Vijver & Leung, 1997b).

7.3.2 Item response theory and the Rasch measurement model.

An item response theory (IRT) model, specifically the Rating scale model, a formulation of an extended Rasch model, was employed in this study as implemented by Winsteps (Linacre, 2009). IRT, also referred to as latent trait theory, is a paradigm for the design, the analysis, and the scoring of questionnaires and other instruments. A fundamental purpose of IRT is that it provides a theoretical framework which enables researchers to evaluate how well tests and measuring instruments work, and more specifically, how well the individual items on these measures work (Hambleton, Swaminathan, & Rogers, 1991). In a multidimensional instrument IRT allows researchers to determine how adequately the attitude continuum which underlies each dimension is assessed by the respective items in the instrument (Beck & Gable, 2001d). IRT is frequently used for developing and refining measuring instruments (Hambleton et al., 1991) and assessing performance across groups using conformable items where all the respondents did not need to respond to all the items (Andrich, 2004).

IRT models are based on the assumption that the items that are being analysed are unidimensional, in other words, a single construct, or single dominant affect or attitude is measured (Chou & Wang, 2010; Harvey & Hammer, 1999). Most IRT models assume

unidimensionality, in other words, all the items measure the same latent trait or underlying construct (Chou & Wang, 2010). The latent trait is the "unobserved characteristic that is presumed to be responsible for the observed responses that are made to the test's items" (Harvey & Hammer, 1999, p. 356). The latent trait is therefore that which is intended to be measured and "is defined by the items or agents of measurement used to elicit its manifestations or responses." (Linacre, 2009, p. 429).

Another assumption of IRT is local independence, meaning that a person's responses to one item are statistically independent to responses on any other items (Beck & Gable, 2001d; Linacre, 2009, p. 392). In local independence the latent trait measured, in this case PPD, is the only factor affecting the response to an item. This means that once the contribution of the latent trait to the data is removed, only random and normally distributed noise is left (Chou & Wang, 2010, p. 719; Linacre, 2009, p. 392). The local independence of items therefore implies that no residual associations are left in the data after the latent trait has been removed (Pallant et al., 2006). This means that all covariance among the items occurs as a result of the association between the items and the underlying construct being measured (Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006). Local independence is associated with unidimensionality because a data set can only be unidimensional when item responses are locally independent based on a single latent variable (Embretson and Reise as cited in Chou and Wang, 2010, p. 719).

Another fundamental premise of IRT is that an item characteristic curve (ICC) or function can describe the relationship between a respondent's item performance and the group of traits that underlie the item performance (Hambleton, Swaminathan, & Rogers, 1991). Accordingly, as the level of the latent trait increases, so too does the probability

that the respondent will endorse items and/or select categories that signify higher levels of agreement with the items (Beck & Gable, 2001d).

A number of different models have been developed within IRT. The different IRT models are named by the number of parameters that are used to describe the items of a questionnaire. Three popular IRT models are the one, two, and three parameter logistic models.

According to Yu (2010), IRT and Rasch measurement models are similar to each other in terms of computation, but their philosophical foundations differ immensely. Whereas IRT models may use up to three parameters, the Rasch model utilises only one parameter. The Rasch measurement model is often regarded to be a one-parameter IRT model. (Furr & Bacharach, 2007). The Rasch measurement model is, however, different to the one-parameter IRT model as it offers a completely different approach to conceptualizing the relationship between data and the theory (Andrich, 2004b; Royal, 2010). IRT attempts to fit a model to the observed data whereas the Rasch measurement model specifies that the data fit the model (Andrich, 2004b). The IRT approach would therefore adjust the model parameters to reflect the patterns that are observed in the data. A Rasch approach, on the other hand, specifies what the requirements are for fundamental measurement and emphasizes fitting the data to the model before any claims concerning the presence of a latent trait in the test or measuring instrument may be considered valid (Andrich, 2004b). Although IRT and the Rasch measurement model have diverse views regarding model-data fit, they are similar in that they take both person and item attributes into consideration in assessment methods, in contrast to classical test theory.

The Rasch measurement model is based on the assumption of a unidimensional measurement model (Bond & Fox, 2001). The Rasch measurement model assumes that if a person responds to a unidimensional construct then he or she ought to respond as expected according to his or her ability level (also referred to as trait levels) and according to the item difficulty level (Smith, Conrad, Chang, & Piazza, 2002).

Unidimensionality means the questions measure a single dominant affect or attitude (Smith et al., 2002) which, in this study, is the typical emotions or symptoms for the degree of depression experienced. Unidimensionality can always be determined on a particular level of reduction. Depression, for example, is unidimensional on a higher level but is multidimensional on a more basic level (Biondi, Picardi, Pasquini, Gaetano, & Pancheri, 2005). Smith et al (2002) state that "if an instrument is composed of multiple subscales, then unidimensionality refers to the set of items for each subscale" (p. 191). All items from the same subscale should therefore load on the construct measured by that subscale, and not on any other subscale.

In the Rasch measurement model, the item difficulty is the single item characteristic that is assumed to influence a respondent's performance (Rasch as cited in Beck & Gable, 2001d, p. 7; Smith, 2004). Or stated differently, the probability of a specific response by a specific person on a specific question is a function of the person's "ability" (level of depression) or theta (θ), and the "difficulty level" of the item (or d). In this sense "difficulty level" refers to the difficulty of endorsing an item (yes or no in a dichotomous case, or more or less as in the traditional 5-point Likert scale). The "ability" therefore indicates the degree of a latent variable, such as depression, that the item is meant to measure.

A distinguishing attribute of the Rasch measurement model is that the item difficulty parameter and the person ability parameter can be estimated separately from each other (Schumacker, 2004). As a result it yields a test free person ability calibration because the person's ability may be estimated independently of the sampling distribution of the test items. The Rasch measurement model also makes sample free item difficulty calibration possible where item difficulty is estimated independently of person abilities (Schumacker, 2004).

A number of different Rasch measurement models have transpired to address the vast number of psychometric needs across various disciplines (Schumacker, 2004). The various measurement models provide the means for constructing interval measures from raw data. The family of Rasch measurement models include the dichotomous model – the simplest measurement model, the partial credit model, the rating scale model, binomial trials, Poisson counts, ranks, many-faceted, and multidimensional models (Wright & Mok, 2004). The different Rasch measurement models are defined by either the way in which a respondent is able to record a response to an item or the different response formats, by the number of dimensions in a multidimensional model, by the number of facets in the data collection design, or a combination of these factors (Schumacker, 2004). The Rasch rating scale model employed in this study describes the probability that a person will endorse a particular rating scale category on a specific item of a rating scale. In rating scale analysis the number of thresholds refers to the number of response categories. There is only one threshold in the dichotomous Rasch model.

The Rasch model makes it possible to construct linear additive measures from "counts of qualitatively-ordered observations, provided that the structure of quantity is

present in the data" (Linacre and Wright as cited in Salzberger, 2010, p. 1275). In an instrument where the test is unidimensional, or where the subscales are unidimensional, Rasch analysis is able to order the items of the scale or subscale from least to most difficult on a continuum. Rasch analysis is also able to calibrate person affect measures and place the respondents on the continuum – a linear scale – according to their item agreements. The person and item calibrations have equal interval units of measures on the continuum and are termed "logits". Logits are calibrated data with a mean of 0 and a SD of 1. A negative logit represents an item that is easy to endorse, whereas a positive logit represents an item that is hard to endorse (Smith et al., 2003; Wright & Stone, 1999). Logits less than -2 or greater than +2 are very easy or very hard to endorse respectively (Maree, 2004).

Since logits are linear metric units they may be used to compute item difficulty, trait ability, and item-fit indices to analyse the psychometric properties of an instrument for a certain population. To determine whether an instrument is appropriate for the sample, the overlap between item difficulty and trait ability distributions on the logit scale are examined (Hong & Wong, 2005).

Rasch analysis provides indicators of how well every item fits within the underlying construct using linear metric units, providing the researcher with insight regarding the relative difficulty of items and therefore allows for examining the construct validity of an instrument (Overston as cited in Kyriakides, Kaloyirou, & Lindsay, 2006, p. 785; Wu & Chang, 2008). Construct validity can only be achieved if the structure of the variable is supported by the item calibrations and if person characteristics can be substantiated by their placement on the variable (Wright & Stone, 1999). Construct

validity may therefore be determined by comparing both person ability levels and item difficulty levels on the logit scale. The difficulty indices allow for the examination of how well the items span the continuum giving an indication of how well the items measure what they are intended to measure. Better construct validity is achieved if the items are well differentiated or spread out on the logit continuum. This allows for a more complete score interpretation for both high and low scoring respondents because the content of the respective items provide a more adequate definition of the construct (Beck & Gable, 2000; 2001d; 2003; Bond & Fox, 2001; Bond, 2003; Smith, 2004).

Unidimensionality is analysed by principal components factorial analysis of the residuals as well as by analysis of fit statistics or indices (mean-square infit and mean-square outfit) – a necessary quality control technique to determine the validity of test items and person responses. Fit analysis is an important part of using latent trait models, like the Rasch model, if the interpretation of the calibration of results is to be meaningful. When the parameters of a Rasch model are estimated, they may be used to calculate the expected response pattern for every item and person. Comparison of observed and expected response patterns yields the fit statistics for the persons and items (Linacre, 2009, p. 428; Wright & Stone, 1999, p. 47). Fit statistics therefore enable researchers to determine how well the data cooperates with the construction of measurement and if and where misfit occurs – in other words, person and item response patterns that do not meet the requirements of the model and do not contribute to useful measurement. Confidence may be placed in person measurement and item calibration when the fit statistics fall within an acceptable range for the study (Smith, 2004; Wright & Stone, 1999). Person fit to the Rasch model is an indication of whether the participants responded consistently to

the items, or whether their responses were erratic. Item fit to the Rasch model is an indication the items performed logically. Item misfit may occur when the item is confusing, too complex, or it measures a construct other than what it was intended to measure.

Two aspects of fit are reported, namely infit and outfit. Non-standardized person fit and item fit are reported for infit and outfit as a mean-square statistic (MNSQ) and as a standardized value (*t*-statistic). MNSQ statistics are reported in this study. The infit statistic gives more weight to person performance closer to the item value. In other words, persons whose ability is close to the item's difficulty level should provide more sensitive insight regarding that item's performance. Outfit statistics are not weighted. They are more sensitive to the influence of outlying scores. Aberrant infit scores are normally a greater reason of concern than aberrant outfit scores. More attention is therefore paid to infit scores than to outfit scores by those who use the Rasch model (Bond & Fox, 2001).

MNSQ statistics of 1.00 are ideal values by Rasch specifications. Linacre (as cited in Chiang, Green, & Cox, 2009, p. 266) states that the values for differentiating fit and misfit should be sufficiently flexible to allow for researcher judgment. However, MNSQ statistics between 0.50 and 1.50 are considered to be productive for measurement (Linacre, 2002).

According to Smith, Schumacker, and Bush (1998, p. 78) the MNSQ statistic is dependent on sample size and relying on a single critical value for the MNSQ can result in an under detection of misfit. Wright (as cited in Smith et al., 1998, pp. 78-79) provides

a formula for calculating the critical value for mean squares which takes the sample size into account:

$$Critical\ value\ MNSQ(infit) = 1 + \frac{2}{\sqrt{x}}$$

$$Critical\ value\ MNSQ(outfit) = 1 + \frac{6}{\sqrt{x}}$$

Where x = the sample size. If this formula is applied to the samples in this study, the critical value for the MNSQ infit would be 1.15 for both the English and Afrikaans samples. The critical value for the MNSQ outfit for the Afrikaans sample (n = 178) would be 1.45, and 1.44 for the English sample (n = 187). The MNSQ infit value in this calculation is particularly stringent, and more in line with values for high stakes questionnaires, for which a range of 0.80 to 1.20 is recommended (Wright & Linacre, 1994). Values between 0.60 and 1.40 are more suitable for rating scales (Wright & Linacre, 1994). Bond and Fox (2007) also recommend that MNSQ infit and outfit values for persons and items be in the range of 0.60 to 1.40 for a Likert scale. Based on these recommendations a range of 0.60 to 1.40 was selected for differentiating between fit and misfit items and persons.

MNSQ statistics that are greater than 1.40 may suggest a lack of construct homogeneity with other items in the scale (Doble & Fisher, and Green, as cited in Hong & Wong, 2005, p.132). Items with MNSQ statistics which are smaller than 0.60 may suggest the presence of item redundancy.

7.4 Advantages of Item Response Theory and the Rasch Measurement Model over Classical Test Theory

Classical test theory (CTT) is a methodological approach which employs conventional techniques to analyse data. CTT is popular and has its purposes but it also has shortcomings which causes it to mask vast amounts of important information (Royal, 2010). "Modern test theory", or item response theory (IRT), is a measurement theory that was developed to address some shortcomings of CTT (Lord as cited in Beck & Gable, 2001d, p. 5).

IRT methods are gaining popularity in wide variety of psychological assessments and are not only limited to the traditional measures of aptitude and ability and measures with dichotomously scored items. This may be partly attributed to an increase in computer availability and advances in computer software in recent decades which has favoured the more computationally demanding techniques of IRT relative to those based on CTT (Harvey & Hammer, 1999). The use of IRT models capable of analysing items which are rated by means of ordered-category scales, such as the Likert scale, or unordered, nominal scales have gained increasing attention. Harvey and Hammer (1999) state that "the addition of these polytomous models renders the IRT approach applicable to virtually any type of standardized psychological assessment instrument" (p. 354). Furthermore, it has been predicted that IRT-based methods will, to a large degree, replace CTT-based methods in future years (Harvey & Hammer, 1999).

Two different approaches within IRT developed as a result of different viewpoints of the most prominent IRT pioneers, one articulated by Rasch (1960), and the other by

Lord and Novick (1968) and Birnbaum (1968). The primary difference between their approaches was concerned with the how the relationship between the data and the model was conceptualised (Andrich, 2004a). These two approaches are referred to as the traditional paradigm and the Rasch paradigm by Andrich (2004a). The traditional paradigm contends that measurement models must fit the data whereas the Rasch paradigm contends that the data must conform to the measurement model.

Although these traditional paradigm and the Rasch paradigm have have diverse views on model-data fit, they both offer several advantages over the classical test theory (CTT) in terms of test development and evaluation as well as the scoring process. The term "IRT" is used in the next section to refer broadly to these two paradigms, both which take person and item attributes into account. Where the Rasch measurement model, as an extension of IRT, has a unique advantage over the traditional paradigm in IRT, it will be indicated. Some of the advantages of IRT-based methods are highlighted below:

7.4.1 Focus on item-level.

IRT provides a more holistic, integrative view of item performance as it focuses more on the item-level than CTT, which places a greater focus on test-level indices of performance such as the overall reliability coefficient of an instrument (Harvey & Hammer, 1999). The sample dependency of item and test indices (like reliability indices, p-values, and item-total correlations) and the item dependency of person ability is a major limitation of CTT. Furthermore, although CTT is able to quantify the total sample

difficulty or item discrimination, it is unable to effectively combine and present this information simultaneously in a convenient format (Smith et al., 2002).

7.4.2 Better construct interpretation.

IRT models were established on the assumption that the items being analysed are essentially unidimensional (Bond & Fox, 2001; Smith et al., 2002). This does not, however, restrict the model only being applied to instruments that measure a single variable. Instruments which are composed of multiple subtests or dimension can also be analysed using a unidimensional IRT model as each subtest or dimension is analysed separately. The assessment of unidimensionality is important as it provides evidence of the construct validity of a measure (Van der Ven & Ellis, 2000). IRT analysis enables researchers to examine the construct validity of instruments more thoroughly and can result in finer construct interpretation. This allows for a more thorough description of high- and low-scoring respondents (Beck & Gable, 2001d).

7.4.3 Better measurement precision across the continuum of the variable.

In IRT-based methods, items with higher discrimination result in higher levels of information which indicates better measurement precision, and therefore lower undesirable errors of measurement. This is because information in IRT is inversely related to the standard error of measurement (SEM). In CTT-based methods it is the concept of reliability which indicates better measurement precision, and it is reliability

that is inversely related to the SEM (Harvey & Hammer, 1999). In CTT, however, only one reliability estimate is given and, because the SEM is calculated using the reliability estimate, CTT provides only one SEM which is applied to all the scores – despite the knowledge that extreme scores are less precise. CTT lacks procedures that would make it possible to determine how measurement error varies across the different ability levels (Smith et al., 2002). CTT assumes that the test is equally precise across the possible test scores and a single number, like the internal-consistency reliability coefficient, for example, is used to quantify the measurement precision of a test (Harvey & Hammer, 1999). IRT-based methods provide a SEM for each person and each item. This enables the researcher to determine the accuracy of item location or person ability estimates which cannot be accomplished with CTT-based methods.

7.4.4 Test development.

IRT-based methods have advantages of CTT-based methods when items need to be selected for test development. By using IRT-based methods the test developer is more easily able to determine the effect of deleting or adding a certain item or set of items by exploring the test's combined information function for the items being examined (TIF) and the test standard error (TSE) function for an item pool. Investigating the change in graphic curvature of the TIF or TSE functions after deleting or adding items and comparing this to the desired performance curve allows the test developer to tailor the test closely to desired specifications. Test developers using CTT-based methods need to

rely on far less sensitive measures like the test's global coefficient alpha or standard error of measurement (SEM; Harvey & Hammer, 1999).

7.4.5 Information on category functioning.

Another advantage of using Rasch analysis for the validation of latent trait measures is that it provides additional information on the category functioning of an instrument. Linacre states that this may serve to increase the measurement accuracy of the instrument (as cited in Wu and Chang, 2008). Winsteps (a Rasch analysis software program) provides rating scale category counts, rating scale category fit statistics, average measures, and step measures to assist researchers in determining whether there are potential problems with the functioning of the rating scale (Linacre, 2009). Determining whether the average measures or step calibrations advance monotonically across the rating scale categories, for example, enables researchers to optimize the effectiveness of the scale categories (Linacre, 2004).

7.4.6 Scoring methods.

IRT based methods offer substantial advantages over the scoring methods typically used in CTT-based tests. More sources of information can be considered simultaneously using IRT to estimate a respondent's score, including the specific items that were answered correctly or incorrectly, and the item's properties, like difficulty level and discrimination. This makes it possible to assess the degree to which the IRT model being

used provides a good fit to the individual's response pattern, to produce better estimates of the latent trait scores, and to produce quantitative estimates of the "quality" or likelihood of an observed response profile (Harvey & Hammer, 1999). A limitation of CTT-based methods is that it is not possible to determine how a person may respond to a certain item. As different metrics are used for persons and items it is not possible to predict the outcome of the interaction between item difficulty and person ability.

In Rasch measurement the scores are linear and are mapped onto a common metric with the same calibrations and steps. Co-calibration, a process in Rasch measurement, makes it possible for a number of instruments, which purport to measure the same construct, to measure in the same unit. This is possible even when the separate instruments have a different number of items, a different number of rating scale points, and rating scale points with different labels (Smith, 2004). Rasch measurement therefore assists in determining convergent validity between instruments that purport to measure the same construct (Smith, 2004).

7.4.7 Differential item functioning.

One of the major assumptions of the Rasch measurement model is the generalisability aspect of construct validity when the data fits the model. Parameter invariance is one major characteristic of the Rasch model (Smith, 2004, p. 109). Invariance of parameters (item invariance and person invariance) is an important distinction of the Rasch measurement model from other latent trait models and CTT (Bond, 2003; Smith, 2004, p. 109). This refers to the extent to which inferences regarding

item calibrations or person measures are invariant, within measurement error, across different groups, contexts, tasks, or time frames. Only the Rasch measurement model has sufficient statistics for estimating item and person parameters (Smith, 2004, p. 109). Examining item and person invariance "places the boundaries and context to which the frame of reference for interpretations can be extended or constrained." (Smith 2004, p. 110).

If parameter (or measurement) invariance is established then it means that there will be an equal probability that two individuals from different cultural or demographic groups will respond in the same way to an item, given that both individuals are at the same level of the latent trait being measured. Once parameter invariance is established, the differences on an instrument's scores accurately reflect the differences on the latent characteristics assessed by the latent trait or construct. Parameter invariance is determined through analysis of differential item functioning (DIF), which refers to distortions at the item level (Ægisdóttir et al., 2008; Küçükdeveci, Sahin, Ataman, Griffiths, & Tennant, 2004).

Psychological assessments should be free from DIF as items with DIF differ in psychological meaning across cultures or groups and have the potential to impact on the comparison of total test scores across cultures or groups. Therefore individuals from different cultures groups who have the same ability have a different probability of getting the item right (Hambleton et al., 1991).

Harvey and Hammer (1999) regard CTT-based methods of assessing bias as being limited as they do not allow for distinguishing between a scenario where the subgroups

have different means and the test is biased and a scenario where the means are different, yet the test is not biased. IRT techniques on the other hand offer a powerful alternative for examining DIF (Harvey & Hammer, 1999). Assessing bias, or analysis of invariance, can be conducted by using CTT-based methods by examining differences in item means by group or time, but such analyses are greatly simplified via use of the Rasch measurement model (Andrich, 2004; Chiang et al., 2009) due to its assumption of parameter invariance.

7.4.8 Administrative efficiency and item banking.

Administrative efficiency and item banking, where items may be selected from a calibrated item pool for every individual being assessed, differ significantly between IRT-based and CTT-based testing. Using a CTT-based approach, it is difficult to compare the performance of persons taking different forms of an assessment. It is also not possible to compare scores obtained from the same set of items unless the entire data set is available or a certain type of imputation method is used (Smith et al., 2002). The assumption in CTT-based testing is that the whole item pool will be administered to each individual whereas IRT-based testing allows for selecting different items for different individuals, or selecting a different number of items for different individuals. This is because the IRT model results in instruments that are sample free and test free (Schumacker, 2010). A sample-free measure means that the item characteristics do not vary with the sample being researched, in other words, the instrument transcends the group measured. When an instrument is test free, several items at a variety of difficulty

levels may be omitted from the scale without influencing the respondent's score. Furthermore, in a test-free instrument, it is not necessary for every person to complete the entire scale (Wright as cited in Beck and Gable, 2001d). The family of Rasch measurement models in IRT is rather robust when data is missing, and comparative estimates can still be made even for those individuals who did not respond to all items, as well as for items that only some individuals responded to (Wright & Mok, 2004).

In the context of Rasch measurement, the process of co-calibration of instruments that purport to measure the same construct, makes it possible to select a different mix of items depending on the desired precision at different locations on the variable (Smith, 2004). This has the advantage of tailoring the selection of test items and reducing testing time by limiting the number of items, or by administering more tests in the same amount of time while still producing a test with its highest degree of measurement precision for a specified latent trait (Harvey & Hammer, 1999). Typical scoring methods used in CTT-based approaches are highly dependent on each individual having the same list and number of items.

7.4.9 Additivity.

Additivity is an advantage that the Rasch measurement model has over other IRT models and over CTT. Additivity refers to the properties of the measurement units which are called logits (logarithm of odds). IRT techniques like Rasch analysis are capable of constructing linear measures from counts of qualitatively-ordered observations, provided the data fit the Rasch model (Linacre and Wright as cited in Salzberger, 2010, p.1275).

The ordering of items on a continuum (item difficulty calibration) and calibrating person affect measures by means of linear metric units (logits), which maintain the same size over the entire continuum, allows for computing multivariate parametric statistical techniques (Smith et al, 2002). Of all the IRT-based models, only Rasch analysis strives to provide invariance in scientific measurement with respect to estimates of item difficulty and person ability. The Rasch family of measurement models is the only model that produces linear measures, gives estimates of precision, and is able to separate the parameters of the measuring instrument and the object being measured (Wright & Mok, 2004).The use of an interval level of measurement for person and item estimates, as opposed to an ordinal level measurement like in CTT-based techniques, means that invariance of item and person estimate values always remain relative (Bond & Fox, 2007, p. 71). Although CTT-based statistical models typically assume an interval scale of measurement in order to allow parametric statistical techniques, they are based on raw scores that are mostly from ordinal scales of measurement that do not support the mathematical operations needed to compute means and standard deviations. When logits as opposed to raw scores are used, researchers are better able to calculate means and variances and this allows for a more accurate determination of reliability (Schumacker, 2004; Smith, 2004). This will be discussed in more detail in the section that follows.

7.4.10 Superior reliability estimates.

Reliability, according to Schumacker (2004), is typically defined as 'the consistency of responses to a set of items or the consistency of scores from the same

instrument or parallel-forms instrument. Reliability is also defined as the degree to which scores are free from measurement error.' (p. 243)

In CTT, five different types of reliability coefficients are generally used, depending on the test situation. These are: 1) test-retest reliability, 2) rater consistency, 3) alternate forms reliability, 4) split-half, and 5) internal consistency. Rasch measurement models can also be used to compute reliability for various test administration designs. Rasch-based methods allow for identifying measurement error in the same type of testing situations, provides reliability estimates and individual SE's, and is able to yield more diagnostic information regarding individual person and rater performance. Rasch-based methods have been described as more advantageous to traditional methods in CTT (e.g. Harvey & Hammer, 1999; Schumacker, 2004; Smith, 2004) as they allow the researcher to pinpoint those individuals who exhibit consistent, declining, or improved performance on different forms of the same test, or on retesting. Those individuals who show declining or improved performance may then, for diagnostic purposes, be identified and the reasons explored. Furthermore, the rater reliability design provides more extensive information regarding individual raters, like rater consistency, severity levels, and potential bias (Schumacker & Smith, 2007).

Rasch analysis for calculating internal consistency also has distinctive advantages over traditional measures of internal consistency. In CTT, Cronbach alpha is the traditional measure of internal consistency, or reliability coefficient, indicating the extent to which items measure a single construct. It examines the average inter-item correlation of the items in a questionnaire (Cortina, 1993). If all items in a questionnaire are measuring the same construct (without any error), then Cronbach alpha will be equal to

one. If there is no shared variance in the items, then only measurement error is reflected which results in Cronbach alpha being equal to zero (Hinton as cited in Spiliotopoulou, 2009). A Cronbach alpha value of one does, however not necessarily imply unidimensionality of the questionnaire (Helms, Henze, Sass, & Mifsud, 2006). The presence of more than one construct may be determined by factor analysis. When there is a one factor solution, the Cronbach alpha is likely to be high, which indicates that the items are measuring the same latent construct. A Cronbach alpha value equal or greater than 0.70 is conventionally regarded as an acceptable level of internal consistency (Bland and Altman as cited in Spiliotopoulou, 2009). Caution should be taken, however, when judging estimates of internal consistency as a low coefficient alpha value might not always indicate problems with the construction of the tool and a high value does not always suggest adequate reliability. Spiliotopoulou (2009) indicates that these reports might rather be a reflection of the data characteristics of the construct and suggests that researchers, reviewers, and practitioners should consider several guidelines for interpreting internal consistency estimates. These guidelines may include consideration of the variability of the data, whether the statistical tool is appropriate for the level of measurement of data, whether the data are normally distributed and linear, the scale's length and width, and the sample size.

Cronbach alpha utilises nonlinear raw scores in calculating the sample variance, and like other traditional estimates of reliability, normally include extreme scores (i.e. zero scores and perfect scores) that do not have any error variance (Schumacker & Smith, 2007). Including these scores therefore decreases the average error variance which results in an increase in the reported reliability (Schumacker & Smith, 2007; Smith, 2004). The

sample variance is therefore potentially misleading (Smith, 2004). Rasch analysis, on the other hand, typically excludes extreme scores due to their SEMs being infinitely large and that they provide little information regarding the person's location on the underlying variable (Linacre as cited Smith, 2004, p. 99).

Concern has also been raised about the use of raw scores in the SEM (Schumacker & Smith, 2007). The classical method of estimating the SEM uses the reliability coefficient and the score standard deviation (SD):

$$SEM = SD_x(1 - R)^{1/2}$$

Where SD_x represents the observed spread of the sample raw scores and R represents the reliability estimate. The average error variance for the test, and hence the confidence intervals around the scores are represented by the SEM. When determining the precision of every score on the scale, this method may be misleading due to extreme scores being less precise than central scores (Smith, 2004).

Smith (2004), Schumacker (2004), and Schumacker and Smith (2007) address these concerns within the context of Rasch measurement, where each person's ability and each item's difficulty are indicated on a linear scale as logits, as opposed to raw scores on an ordinal scale – provided the data fit the model. These estimates, due to them being on a linear scale, are more appropriate for calculating means and variances (Smith, 2004). Schumacker (2004, p244.) concurs that 'reliability determination in the Rasch model is more directly interpretable because logits (linear measures) rather than raw scores (ordinal measures) are used. Logits rather than raw scores are used in Rasch analysis

because logits satisfy the following criteria for measurement: logical ordering, linear scales, and objective comparisons. The calibration of items and persons on a common linear scale provides information on criterion-referenced and norm-referenced information for person measures and item calibrations.

A further advantage of Rasch-based methods is that it yields a direct estimate of the modelled error variance for each estimate of a person's ability and item's difficulty rather than sample dependent averages used in CTT (Schumacker, 2004; Wright in Smith 2004, p. 96). CTT lacks procedures for determining how measurement error varies across person ability levels (Smith et al., 2002). The SEs in Rasch models provide a quantification of the precision of every person measure and item difficulty. They can also be used to describe the confidence intervals in which each item's 'true' difficulty, or person's 'true' ability lies. The individual SE may be more useful than a sample or test average which overestimates the error variance of persons with extreme scores. Should a group estimate of reliability be required, the individual SE may be squared and summed to yield a correct average error variance for the sample (as opposed to the error variance for an 'average' person sampled) which is then used to calculate formulas for internal consistency. A superior estimate of internal consistency is produced due to numerical values being linear (provided the data fit the Rasch model), and due to the actual average error variance of the sample being used as opposed to the error variance of an 'average' person. The result is a person variance that is adjusted for measurement error, which represents the 'true' variance in the person measures. Furthermore, in Rasch measurement, person separation reliability (person reliability estimate) is calculated as

the ratio of the adjusted true variance to the observed variance. This represents the proportion of variance that is not due to error.

Correlation-based reliability estimates (including KR-20 and Rasch person reliability) are unfortunately nonlinear and suffer from ceiling effects as their estimates are restricted in range from zero to one. The Rasch measurement model addresses these shortcomings by yielding a person separation index and an item separation index which have a range from zero to infinity.

“Separation” in Rasch analysis is the measure of the spread of both items and persons in standard error units. The separation index should exceed 1.00 for the instrument to be minimally useful. Higher separation values represent a better spread of persons and items along a continuum. Lower separation values indicate redundancy in the items and less variability of persons on the trait. The separation value provides evidence of reliability, with higher values yielding higher reliability.

The item separation index allows the researcher to determine whether the items discriminate different levels of person performance and therefore provides evidence of "test" reliability (Linacre, 2009). Conventionally, only a person reliability estimate is reported, which also provides an indication of test reliability (Linacre, 2010). Larger item separation indices demonstrate better confidence in the spread of items across the targeted continuum (Beck & Gable, 2001d; Bond & Fox, 2001).

Rasch analysis also produces a person separation index. The person separation index enables the researcher to determine whether persons are able to discriminate differences in item calibration (Linacre, 2009). The person separation index is on a ratio

scale and is able to compare the true distribution of person measures (in logits) with their measurement error, which results in an indication of the spread of person measures in SE units (Fisher as cited in Smith, 2004). The higher the person separation index, the more spread out the persons are on the variable being measured and the better the reliability. According to Linacre (2009) a separation index of 2 signifies that high measures are statistically different from low measures.

It is useful to examine the person separation index across several analyses of the same data, as an increase in person separation index signifies an increase in reliability even when Rasch person reliability remains unchanged due to it's maximum value of one (Smith, 2004; Schumacker & Smith, 2007).

In Rasch measurement, the person reliability estimate (person separation reliability) provides evidence for internal consistency reliability. It is calculated as the ratio of the adjusted true variance to the observed variance:

Person Reliability Estimate = True Variance / Observed Variance.

This represents the proportion of variance that is not due to error. The Rasch person reliability estimate is conceptually equivalent to Cronbach's alpha, but is computed without extreme scores making its value lower than that for Cronbach's alpha. Winsteps provides a person reliability estimate as well as an item reliability. CTT does not typically compute an estimate of item reliability (Linacre, 2009).

7.5 Participants and Sampling Procedures

For the purpose of this study three different categories of participants are needed: 1) participants for the translation process; 2) participants for the administration of the PDSS, the EPDS, and the QIDS; and 3) participants for the administration of the Afrikaans version of the PDSS, the Afrikaans version of EPDS, and the Afrikaans version of the QIDS. Two different sampling procedures were employed in this study. The first pertains to the translation process, and the second to the administration of the screening questionnaires. These are outlined below.

7.5.1 Participants for the translating process.

The main purpose of this study was to provide an Afrikaans version of an existing PPD screening measure – the PDSS – and to determine the reliability and validity of the Afrikaans PDSS and the English PDSS on respective South African mothers. The QIDS-SR and the EPDS are two additional screening questionnaires that were selected as convergent instruments to provide additional data on the construct validity and the equivalence of constructs across the translations. At the time this study was undertaken, an Afrikaans version of the EPDS was available from Postnatal Depression Support Association South Africa (PNDSA), but not an Afrikaans version of the QIDS-SR. Therefore accredited translators and persons with a thorough knowledge of the subject matter were required for the translation of the PDSS as well as the QIDS.

A non-probability sampling technique – purposive sampling – was used to select the translators, back-translators, as well as experts to review the translated versions. The

reason for this is that the researcher wanted to ensure that individuals who translated the PDSS met certain requirements: a) they had to be bilingual in English and Afrikaans; b) they must have had experience in translating in these languages; and c) they must have some knowledge about the subject matter they will be translating. Knowledge about the subject matter being translated is important to avoid literal translations being made, which could cause misunderstanding in the target population (Hambleton, 1994). In order to meet these requirements, translators registered with the South African Translators Institute, who were accredited in English and Afrikaans translating, and who had translated subject matter in the field of psychology were selected for the back-translation process. The purpose of the PDSS or QIDS was briefly explained to the translators responsible for translating the respective questionnaires. The experts who were selected to evaluate and review the translations were all psychologists with a clinical background and experience in the assessment of depression and were bilingual in English and Afrikaans.

Six people were involved in the back-translation process and refining of the Afrikaans version of the PDSS. Two bilingual professionals and translators registered with the South African Translators Institute (SATI) translated and back-translated the PDSS. One psychologist from PNDSA and two psychologists, both senior lecturers in the field with experience in the validation of psychological measures, evaluated and reviewed the translations. The author of the PDSS, Professor Cheryl Beck, was also involved in the final revision.

The translation of the QIDS into Afrikaans was performed by four individuals. Two translators from SATI (different translators to those mentioned above) translated and

back translated the QIDS into Afrikaans. The two psychologists/senior lecturers who evaluated the translations of the PDSS also evaluated and reviewed the Afrikaans translation and back-translation of the QIDS. The author of the QIDS was also consulted for advice and permission on the metric conversion of items 8 and 9.

7.5.2 Participants for the English PPD screening process.

A total of 187 English-speaking postpartum mothers of mixed parity were selected through convenience sampling for screening. Participants were therefore selected on the basis of availability. An advantage of this technique lies in the relative ease with which the sample can be made available. However, a disadvantage of this technique is that the sampling method may be seen as arbitrary and not a true representation of the population which limits the generalisability of the results.

Participants were eligible if they met the following criteria:

- Mothers between 4 and 16 weeks postpartum;
- A South African citizen, residing in South Africa;
- Able to speak and read English or Afrikaans fluently; and
- Gave birth to a healthy baby without a disability.

7.5.3 Participants for the Afrikaans PPD screening process.

The same inclusion criteria as listed above applied to this sample. This sample comprised 178 Afrikaans-speaking postpartum mothers of mixed parity who were also selected through convenience sampling. A total of 365 mothers (187 English and 178 Afrikaans) were therefore screened for PPD in this study.

7.6 Measures

Data were collected with a demographic questionnaire, comprising socio-demographic and obstetric data, the Postpartum Depression Screening Scale (PDSS), the Afrikaans version of the PDSS, the Edinburgh Postpartum Depression Scale (EPDS), and the Quick Inventory for Depressive Symptomatology – Self Report (QIDS-SR16).

7.6.1 Demographic questionnaire.

The demographic questionnaire collected data about the mother's home language, language proficiency in either English or Afrikaans, ethnic group, marital status, education level, employment status, mother's age, obstetric history, perception of level of care during labour and delivery, perception of level of support after childbirth, psychiatric history, baby's health, baby's current age and gestational age at birth, and the baby's sex. Questions relating to known risk factors for PPD were also included.

7.6.2 The Postpartum Depression Screening Scale (PDSS).

The PDSS (Beck & Gable, 2000) is a self-report, 35-item Likert response scale consisting of seven dimensions, each containing five items. The dimensions include Sleeping/Eating Disturbances, Anxiety/Insecurity, Emotional Lability, Cognitive Impairment, Loss of Self, Guilt/Shame, and Contemplating Harming Oneself. Each item describes how a woman may feel after the birth of her child. The mother is asked to indicate her degree of agreement or disagreement on a five-point scale from (1) strongly disagree to (5) strongly agree. The woman is asked to circle her answer which best describes how she has felt over the past 2 weeks. After completing the PDSS, or its Afrikaans translation, the mothers in this study were asked to indicate if there were any items that they found difficult to understand.

The PDSS can be completed by the mother in 5 to 10 minutes. The scale may be administered by any health practitioner the postpartum woman comes into contact with. The conceptual basis of the PDSS is based on Beck's series of qualitative studies on PPD (Beck, 1992, 1993, 1996c).

The PDSS is intended to provide an overall score for PPD, but also considers the multidimensionality of PPD and gives seven subscale scores. The summative scoring results in a total score range from 35 to 175. The total score may be sorted into one of three categories: 1) normal adjustment (total score of <59), 2) significant symptoms of PPD (total score of 60 to 79), and 3) positive screening for PPD (total score of ≥80). The psychometric properties of the PDSS are presented in chapter 3.

7.6.3 The Edinburgh Postnatal Depression Scale (EPDS).

The Edinburgh Postnatal Depression Scale (EPDS; Cox et al., 1987) designed to screen for the risk of PPD in women by measuring emotional and cognitive symptoms of PPD and sleep difficulty. The EPDS excludes somatic symptoms of depression as this may be affected by normal postpartum recovery rather than signify a mood disorder. The EPDS does contain items which pertain to anxiety specifically, but opinions are divided on whether the EPDS screens for the presence of anxiety as well as depression (Brouwers et al., 2001; Pallant et al., 2006).

The EPDS is a 10-item self report measure with a 4-point Likert scale. Each of the 10 questions has 4 answer choices that are scored between 0 and 3. The EPDS total score ranges from 0 to 30. The total score is obtained by adding the scores for each item. The cut-off point of the EPDS was calculated to be 12 or 13 for probable depression, and at 9 or 10 for possible depression (Cox et al., 1987). Boyd et al. (2005) have suggested, however, that different cut-off scores may be warranted for different cultural groups. Cultural groups may vary in the manner that depressive symptoms and postpartum experiences are expressed. This has an impact on the assessment of PPD as the symptom presentation may differ across cultural groups, and therefore also the optimal scores for PPD screening (Affonso et al., 2000; Barnett et al., 1999; Bashiri and Spielvogel, 1999).

The EPDS is a widely used screening scale for PPD and demonstrates moderate to good reliability properties across samples from a wide variety of countries and languages (e.g., Barnett et al., 1999; Benvenuti et al., 1999; Berle et al., 2003; Garcia-Esteve et al., 2003). The EPDS has moderate to good correlations with other depression measures (e.g.

Flynn, Sexton, Ratliff, Porter, & Zivin, 2011) and has been found to be a valid screening instrument, when administered verbally, in an urban South African community in a study by Lawrie et al. (1998). The above mentioned factors, and that the EPDS is available freely as a screening tool for PPD, and is brief and easily administered made it a desirable instrument to include in this study.

7.6.4 The Quick Inventory for Depressive Symptomatology – Self Report (QIDS-SR16).

The Quick Inventory of Depressive Symptomatology (QIDS; Rush et al., 2003) is derived from the 30-item Inventory of Depressive Symptomatology (IDS). The 16-item QIDS is a shorter, more time-efficient version of the IDS and is used in daily practice and in clinical research. The 16 items were identified as needed to rate the nine criterion domains of major depression: sleep disturbance, psychomotor disturbance (agitation and retardation); appetite or weight disturbance or both (appetite increase or decrease and weight increase or decrease), depressed mood, decreased interest, decreased energy, worthlessness or guilt, concentration or decision making, and suicidal ideation.

Just as there are two versions of the IDS with identical items: a clinician rating (IDS-C30) and a self-report (IDS-SR30), there are also two versions of the QIDS: Quick Inventory of Depressive Symptomatology – Clinician Rating (QIDS-C16) and the Quick Inventory of Depressive Symptomatology – Self Report (QIDS-SR16; Rush et al., 2003).

The (QIDS-SR16) was selected for use in this study due to it being a self report measure of depressive symptom severity; it provides a specific assessment of al the core

criterion DSM-IV symptoms of MDD; its brevity was considered ideal for the population being screened; it has demonstrated highly acceptable psychometric properties and it has proven useful as a brief rating scale of depressive symptom severity in both research and clinical settings (Rush et al., 2003). The IDS – and the QIDS – were designed to assess depression for a patient population, but the IDS has thus far proven to have excellent sensitivity, good specificity and moderate PPV when administered to women during the postpartum period (Yonkers et al., 2001). Research has shown that the QIDS-SR (16) correlates well with the IDS-SR (30) (0.96) and the Ham-D (24) (0.86), and that the QIDS-SR (16) is as sensitive to symptom change as the IDS-SR (30) and HAM-D (24), signifying high concurrent validity for all three scales (Rush et al., 2003).

The QIDS-SR16 total scores range from 0 to 27. The total scores were obtained by adding the scores for each of the nine symptom domains of the DSM-IV MDD criteria. To score domains which consist of more than one item, the highest score of the item relevant for each domain is taken. The QIDS-SR16 takes approximately 5-7 minutes to complete. Table 8 presents the thresholds that are recommended for major depression screening with the two versions of the QIDS.

Table 8 Severity Thresholds for the QIDS-C16/QIDS-SR16

	QIDS-C16	QIDS-SR16
No depression	≤ 5	≤ 5
Mild	6-10	6-10
Moderate	11-15	11-15
Severe	16-20	16-20
Very Severe	≥ 21	≥ 21

7.7 Procedure

7.7.1 Procedure for the translation of the PDSS.

A multiple method translation incorporating the back-translation method together with a committee (or cross-translation) approach was used in this study. Permission was sought from Western Psychological Services (WPS) for the translation of the PDSS into Afrikaans. Once approval was given, an accredited translator registered with SATI translated the PDSS into Afrikaans. This material was then back-translated into the original language (English) by another accredited translator registered with SATI. The promoter of the study examined, and commented on, the quality of the Afrikaans translation. The original version was compared to the Afrikaans translation, and finally the original version to the back-translation to determine whether there were significant discrepancies. Suggestions were made to improve 8 of the 35 items. A bilingual psychologist, who is also a senior lecturer in the field, with experience in adapting and translating psychological questionnaires, compared the original version to the back translation. The original version was then compared to the Afrikaans version and

discrepancies in linguistic equivalence were pointed out and better alternatives to 14 items were suggested to improve the Afrikaans translation.

The translation was evaluated further by a board member from PNDSA, a bilingual psychologist with extensive experience in the assessment and treatment of PPD. It was suggested that she review all three versions – the original, the back-translation, and the Afrikaans version – along with the comments and suggestions for further improvement made by the study promoter and the bilingual psychologist. The PNDSA psychologist then made recommendations for further improvement to the Afrikaans translation and suggested alternatives to 19 items that would keep the translated version as close as possible to the original version while keeping the language simple and easy to understand.

All the evaluators' comments and recommendations for improvement were then incorporated. The promoter of the study and the researcher evaluated these and, together with the recommendations made for improved quality of the Afrikaans translations, selected the most suitable Afrikaans translations. However, discrepancies with items 16 (I felt like I was jumping out of my skin.) and item 33 (I did not feel real.) remained. The author of the PDSS, Cheryl Beck, was contacted to provide insight into the real meaning of these two items. With her clarification the most suitable Afrikaans translation was selected. The standard translation, back-translation, adjustment sequence has been utilised in many studies requiring the translation of instruments into African languages (Parry, 1996).

7.7.2 Procedure for the translation of the QIDS-SR.

Permission was sought from the author of the QIDS-SR for the translation of the screening scale into Afrikaans so that it could be used as an additional screening scale for the purposes of this study. Once approval was obtained, an accredited translator registered with SATI translated the QIDS-SR into Afrikaans. The Afrikaans version was then back-translated into English by another accredited translator registered with SATI. The researcher and the promoter of the study examined, and commented on, the quality of the Afrikaans translation. The promoter of the study compared the original version to the Afrikaans translation, and the original version to the back-translation to determine whether there were significant discrepancies. The bilingual psychologist/senior lecturer who assisted with the translation of the PDSS also compared the original QIDS-SR to the back translation and the original version to the Afrikaans version. Discrepancies in linguistic equivalence were pointed out and better alternatives were suggested to improve the Afrikaans translation. The researcher and promoter of the study evaluated these comments and recommendations and selected the most suitable Afrikaans translation.

7.7.3 Procedure for the screening process.

A variety of professionals who have contact with postpartum women were approached and informed about the research. These included nursing staff at antenatal, postnatal, and immunisation clinics, obstetricians, general practitioners, staff at maternity hospitals, psychologists, and individuals offering antenatal and postnatal exercise classes. They were asked to assist by identifying participants for inclusion in the study.

Pamphlets were distributed to those professionals who agreed to assist with the study by referring mothers for participation, regardless of whether or not they presented with symptoms of depression or anxiety. Referring professionals were also given information about the research and recruiting lists on which mothers could complete their contact details if they wished to participate or be contacted with more information about the research. Mothers who were interested in participating could opt to contact the researcher in person or could leave her contact details on the recruiting list provided. Regular contact was maintained with referring professionals in order to obtain this information and to encourage the referral of additional mothers.

Mothers who expressed an interest in participating in the research were contacted by the researcher to determine whether they met the following criteria:

- A South African citizen, resident in South Africa;
- Able to speak and read English or Afrikaans fluently;
- Were between 4 and 16 weeks postpartum;
- Gave birth to a healthy baby without a disability.

Referring professionals were also encouraged to refer antenatal mothers for participation in the study. The researcher made contact with these mothers to discuss the research and to make arrangements for participation once their babies were delivered. The researcher then followed up with these mothers between 4 and 16 weeks after their expected due date.

Where possible, the researcher assessed mothers in person by paper/pencil administration. Mothers with internet access could opt to complete the questionnaires confidentially online via a secure, password-protected website. Participants who wished to participate online obtained this information from the researcher. Online assessments allowed the researcher to assess mothers from across South Africa. This method of assessment also meant that mothers could be assessed with minimal disruption in the postpartum period as they were able to complete the questionnaires at home, online, and in their own time. Individuals who suffer from disorders that affect their ability to complete self-report measures reliably and validly were not asked to volunteer for this study. During the paper/pencil administration the participation criteria were discussed with the mothers. The researcher was able to determine from interaction with the mothers whether they were coherent and understood the administration procedure. Those mothers who opted to participate online were provided with information regarding the research and the participant criteria prior to completing the research questionnaires. The researcher was of the opinion that those mothers who could communicate their intention to participate via email or telephonically with the researcher and who could subsequently navigate successfully through the website pages during participation did not suffer from disorders that would negatively impact their ability to participate.

The researcher found that about one third of the mothers who had indicated their desire to participate online refrained from doing so. The researcher assumed that this was due to the time-consuming and demanding role of the early postpartum period. The researcher followed up with these mothers by sending a written reminder about the

research and added that participation is especially valued considering the demands of early motherhood. Mothers who still refrained from participating were not pursued.

Anastasi (1988) states individuals who are assessed for research purposes should be assessed by a suitably qualified person and the participants should receive feedback from the assessment. In accordance with this recommendation, individuals were informed prior to completing the screening questionnaires that they would receive feedback in the form of brief reports of the results of their screening. It was assumed that this would serve as an incentive for participants to complete the questionnaires accurately according to how their mood has been, thereby allowing for more reliable results.

Mothers who screened positively for symptoms of PPD were referred for counselling, to a PPD support group – if one was available in their vicinity – and for further assessment by their doctor if required. Mothers were also given the contact details for PNDSA for additional support and information.

7.8 Ethical Considerations

Good psychological research can only be made possible with mutual respect between the participant and the researcher. The participant should also have confidence in the researcher. Therefore a number of ethical guidelines must be considered when conducting research with human participants (British Psychological Society, 2009). The following ethical principles were followed to ensure that the guidelines as stipulated by the British Psychological Society (BPS, 2006) were adhered to:

- Required ethical approval was obtained from the Western Psychological Services (WPS), who holds copyright for the PDSS, to translate the PDSS to Afrikaans for use in local populations, and to adapt the PDSS for on-line administration, in English and Afrikaans, via a secure on-line environment for administration and scoring.
- Permission was obtained from the author of the QIDS-SR for the translation of the screening scale into Afrikaans so that it could be used as an additional screening scale for the purposes of this study. The author of the QIDS was also consulted for advice and permission on the metric conversion of items 8 and 9.
- Approval was obtained by The Royal College of Psychiatrists for using the EPDS as an additional screening scale for the purposes of this study and for online administration on a password-protected website.
- Ethical clearance was obtained from the University of Pretoria.
- The researcher provided mothers with information regarding the objectives of the study and obtained their informed consent prior to participation. For online participation a procedure was followed, recommended by Kraut et al., (2004), whereby mothers clicked a button on an online form to indicate that they have read and understood the consent form before they could complete the research questionnaires.
- Being involved in a research experience may be a safe and anonymous means for participants to explore thoughts and feelings that they may not want to confide to family and friends (Cooper, Turpin, Bucks, & Kent, 2005). The researcher may also find evidence of psychological problems of which a participant may be

unaware (BPS, 2007). With these factors in mind, and the knowledge that many women with PPD are reluctant to reveal their postpartum distress, the researcher recognised that mothers may use this study as an opportunity to explore symptoms, thoughts, and feelings that they may be experiencing. It was therefore regarded as important to provide the mothers with the results of the screening. In particular, mothers who screened positively for PPD, mothers who indicated the presence of suicidal thoughts or thoughts of harming their babies, and mothers who were unaware that their symptoms would result in a positive screen for PPD, needed to be followed up with information about PPD, information about where to seek support or treatment, and prompt referral to their doctor if required.

- A secure password-protected website was established for participation online. Participants had to contact the researcher in order to obtain a username and password for online participation.
- The participants' biographical questionnaires as well as their screening questionnaires were anonymised and scored by the researcher prior to being sent to the University of Pretoria for statistical analysis to ensure anonymity and confidentiality of the results.
- Individuals who suffer from disorders that affect their ability to complete self-report measures reliably and validly were asked not to volunteer for this study.
- Mothers were informed that participation in the study was voluntary.
- Screening for symptoms of PPD was done at no financial cost to the participants, nor were the participants financially reimbursed for their participation.

- The participants' information was treated with utmost confidentiality and a mother's data was destroyed if she decided to withdraw. No participants wanted to withdraw from the study after completing the research questionnaires.
- Referring health practitioners did not have access to the results of the screening.

7.9 Data Analysis

7.9.1 Descriptive statistics for the PDSS.

The participants' demographic questionnaire data were collated and charted. The data collected from mothers included demographic information as well as known obstetric and psychosocial risk factors for PPD. The descriptive statistics for the English and Afrikaans samples in this study were examined to determine if there were significant differences between them. Furthermore, the demographic and obstetric characteristics of the participants and their PDSS screening results across three screening outcome categories were investigated.

7.9.2 Qualitative data analysis.

Qualitative analysis of the screening questionnaire items was done to arrive at a satisfactory Afrikaans translation of the PDSS. This was achieved by familiarising the translators with the subject matter of the inventory, and then comparing the items on the PDSS to the items on the back-translated version. Face validity was used to determine whether the items in both instruments appeared to measure similar concepts.

7.9.3 Quantitative data analysis.

7.9.3.1 Rasch analysis.

Rasch analysis was conducted as implemented by Winsteps software (Linacre, 2009). The specific measurement model employed was the Rating Scale Model, which is a formulation of an extended Rasch model based on IRT.

The main objective of this study was to analyse the PDSS and the Afrikaans PDSS in South African mothers within the Rasch framework. This would allow for determining the validity and reliability of these screening scales in a South African sample.

Given that the Rasch model distributes items along a level of difficulty, it was possible to determine whether some individual items on the PDSS and Afrikaans PDSS, or in turn, on each of the PDSS and Afrikaans PDSS dimensions, were harder to endorse than others. The psychometric properties of the PDSS and an Afrikaans translation of the PDSS were examined within the Rasch framework to determine how well the items defined the underlying construct of PPD in a South African sample. The PDSS was, however, developed as a multidimensional construct of PPD, incorporating seven individual dimensions. Rasch analysis was also performed to determine how adequately the attitude continuum which underlies each PDSS dimension (or construct) was assessed by the five items which constitute the dimension. These additional analyses of the dimensions were considered essential due to the fact that PPD is a multi-faceted phenomenon.

Fit statistics were computed to show how well the raw data fit the Rasch model. The Rasch model assumes that the items assess a unidimensional or single construct. The PDSS was, however, developed to assess the multidimensional construct of PPD and therefore incorporates seven individual dimensions. Rasch analysis was therefore performed on the PDSS and Afrikaans PDSS as a whole, as well as on each separate dimension.

The hypothesis of unidimensionality is that the items of the same factor should ideally load only on that factor. The assessment of unidimensionality is important as it provides evidence of the construct validity of a measure (Bond & Fox, 2001; Van der Ven & Ellis, 2000). Unidimensionality is equally important to the subtests or, in this case dimensions, that comprise a measure. Assessing the unidimensionality of each dimension of the PDSS and Afrikaans PDSS is therefore an important requirement for unidimensionality of the overall measure. Dimensionality was assessed by examining Rasch principal components analysis of residuals (PCA) and by examining item fit statistics.

PCA residuals were analysed to determine if secondary dimensions were present (Linacre, 2009). The residuals are the difference between the observed and the predicted scores. Using raw data in an analysis leads to non-linearity present in the data being accumulated in the PCA. This analysis was therefore performed using calibrated data (logits).

The item fit statistics – the global infit and outfit mean-square statistics – were also examined to assess the overall fit of the data to a unidimensional structure.

Unidimensionality of the PDSS and Afrikaans PDSS dimensions were determined by individual item fit. This was performed by examining the individual item infit and outfit mean-square statistics. These statistics also provide an indication of how well the data fit a unidimensional Rasch model and if any items misfit was present. Items with mean-square fit values above 1.5 contribute little value to the measure (Linacre, 2009). In this study a range of 0.60 to 1.40, as recommended by Bond and Fox (2007) and Wright and Linacre (1994) for rating scales, was selected to differentiate between fit and misfit persons and items.

When the dimensional structure of the PDSS and Afrikaans PDSS were confirmed, an item analysis using item-total correlations was performed. The Pearson item-total correlation (r_{it}) allows for identifying item misfit thereby providing an indication of the construct validity and whether there are coding problems present. This analysis is similar to the discrimination or item-total correlation in CTT. It does, however, differ in that extreme values are omitted (Maree, 2004). The Pearson item-total correlation (r_{it}) was compared to the expected score (EXP) to determine if discrepancies were evident which could indicate that an item did not fit the dimension well.

Indices of reliability of the PDSS and Afrikaans as well as individual dimensions were determined by Rasch analysis through item and person separation coefficients. Internal consistency reliability was determined by the person reliability estimate. Classic reliability coefficients were also calculated. The item-person map and item and person separation reliabilities were investigated to determine the appropriateness of item difficulty.

The data was also examined to evaluate the effectiveness of the Likert response categories as this impacts on how well the response data defines the dimension. The following six criteria were applied, as suggested by Linacre (2004), to evaluate the appropriateness of the Likert response categories for the PDSS and Afrikaans PDSS:

1. There should be at least ten observations in each category as low frequencies in the category can lead to unstable or imprecise estimates in the step calibrations.
2. There should be reasonably regular observation distribution for each category.
3. The average measures should increase monotonically with each category. People with higher abilities are thus expected to endorse higher categories and people with lower abilities are expected to endorse lower categories.
4. The outfit mean-square statistic for each category should be less than 2. Values greater than 2 suggest the presence of more unexplained variance than explained randomness as anticipated from the Rasch model, therefore indicating that some data did not support the definition of the underlying variable.
5. The step calibrations should advance orderly from easy to hard. An essential conceptual feature of a rating scale design is that a greater amount of the underlying variable in a respondent corresponds to a greater probability that the respondent will be observed in a higher category of the rating scale. When items have disordered categories it causes concern about the appropriateness of the item for measuring the underlying latent variable.
6. Step difficulties should advance by at least 1.40 logits, but by less than 5 logits. If the threshold distance were too wide, a “dead zone” is created in the middle of the category which means that the scale will not be precise in targeting

respondents between two successive categories. A five category rating scale should ideally advance by at least 1 logit (Linacre, 2004). If the advance is less than 1, the categories may need redefining to have wider substantive meaning, or categories should be combined.

When investigating the quality of a new measure, it is important to establish invariance before instruments may be deemed to be equivalent in a measurement sense (Küçükdeveci et al., 2004). Only then do the differences on the screening scales' scores accurately reflect the differences on the latent characteristics assessed by the construct. Invariance is determined through analysis of DIF which is a powerful means of checking for item bias in Rasch analysis (Bond, 2003). The foundation of DIF is to determine whether items have shifted in meaning across different groups or across different time points. Inconsistency in an item's difficulty estimate location across samples, with variation greater than the modelled error is a clear indication that DIF exists and indicates that the item has significantly different meanings for the different groups (Bond, 2003; Bond & Fox, 2007, p.92). Invariance analyses can be conducted via CTT by examining differences in item means by group or time, but such analyses are greatly simplified via use of Rasch model software (Chiang et al., 2009).

In a study by Allalouf and Sireci (1998), a panel of translators and researchers reviewed each DIF item to determine the possible sources of DIF and to formulate general conclusions about the sources of DIF in translated verbal items. The following were found to be the four main causes for DIF:

1. Changes in the difficulty of sentences or words despite an accurate translation;
2. Changes in the meaning of the item or the item content during the translation, thereby creating a different item. This may happen as a result of an incorrect translation causing a change in the meaning of the item, or a word that has a single meaning in the source language is translated into in a word that has more than one meaning in the target language;
3. Changes in the format of the item, for example, longer or shorter sentences in the target language;
4. Items remain the same, but differ in terms of their cultural relevance in the source and target language. The content of an item may, for example, be more familiar to one culture than to another.

Gierl and Khaliq (as cited in Allalouf, 2003, p. 56) identified the following sources of DIF in achievement tests: the addition or omission of phrases or words that affect the meaning of the item; differences in the words or expression either inherent or not inherent to the target language or culture; and format changes of the items.

The translation method used is important to consider in an attempt to reduce the likelihood of measurement bias. Translating an instrument for use in another culture can have a significant impact on the instrument's psychometric properties (Ramirez, Teresi, Holmes, Gurland, & Lantigua, 2006). Back-translation is a method widely used in cross cultural research used for addressing semantic equivalence. Ramirez et al recommend, though, that it be used in addition to other qualitative methods like cognitive interviews

and random probes in order to address aspects of item equivalence and conceptual adequacy both within and across populations from diverse ethnic or cultural backgrounds and from different language groups.

Allalouf and Sireci (1998) formulated a basic flow chart which depicts the process involved in identifying the sources of DIF in an item. They recommend that translators use this flow chart, presented in Figure 2, to identify the causes for DIF. This chart was used as a guidline to examine the sources of DIF in this study.

DIF is generally not anticipated prior to administering tests or screening scales. Researchers must therefore rely on post-hoc explanations to determine the presumed causes of DIF. One recommendation in this regard, is to focus on items in a test that did not display DIF in order to help determine how it is different from items that did display DIF (Allalouf & Sireci, 1998).

Analysis of variance was conducted for each item of the PDSS and the Afrikaans PDSS. This analysis allowed the researcher to determine if DIF was present and if items have significantly different meanings across the two samples.

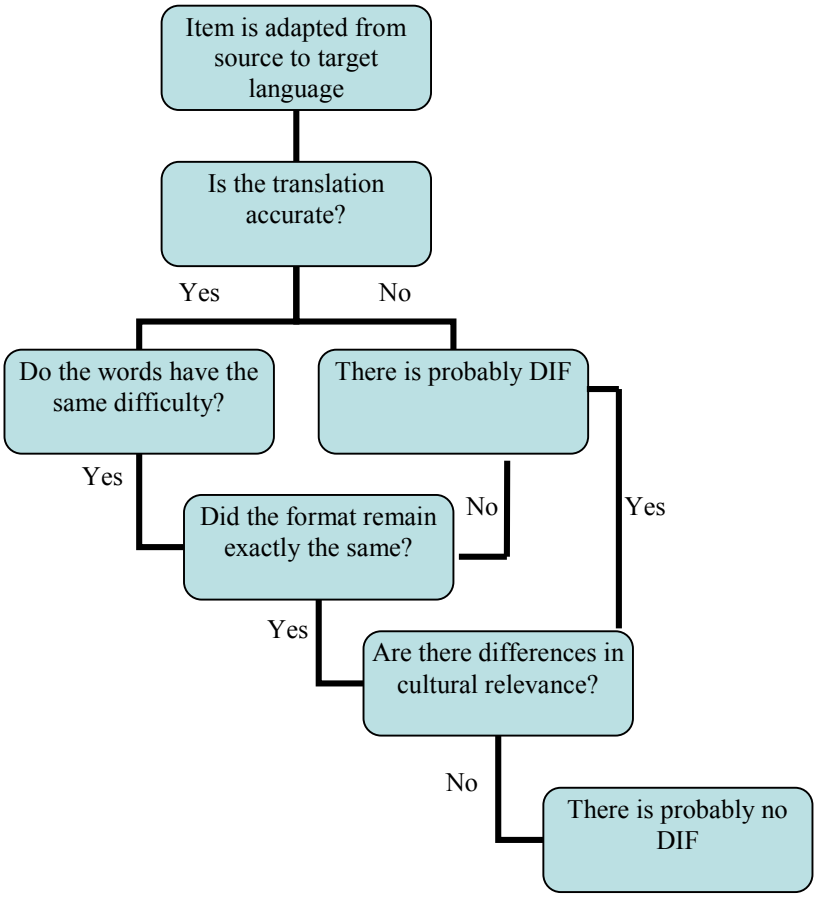

Figure 2 Flow chart for examining the sources for differential item functioning. (Adapted from Allalouf & Sireci, 1998, p. 19).

Convergent validity is a subtype of, and also an important aspect of construct validity that is determined by examining an expected overlap between measures that theoretically measure the same construct. Convergent validity therefore refers to the degree to which a measure is similar to (or correlated with) other measures that it is theoretically predicted to be similar to (or correlate with). High correlations provide evidence of convergent validity (Trochim, 2006). Convergent validity was examined to ascertain whether the PDSS and the Afrikaans PDSS correlate positively with other self-report screening scales for depression, namely the EPDS and the QIDS-SR16 and their respective Afrikaans translations.

7.9.3.2 Multiple regression analysis.

Various factors have been reported to be associated with the development of postpartum mood disorders. The relationship between known risk factors for PPD and scores on the PDSS amongst women in South Africa was investigated through multiple regression analysis. This statistical method was used to analyse the dataset because it is able to depict the relationship between several independent variables (predictor variables) and the dependent variable (the outcome or response) on a continuous scale, such as the severity of PPD.

Multiple regression is able to determine the relative influence of several independent variables (hereafter referred to as “predictor variables” for those variables that may be useful in predicting scores) when they are used to predict or explain a dependent variable (also referred to as the outcome; Field, 2005). The outcome is

therefore explained by the predictor variables and how much influence they have. The fact that multiple regression makes it possible to determine how important the predictor variables are and takes into account how important the associations between the predictor variables are, has made it an extremely popular method of data analysis in the past couple of decades (Cramer, 2003; Foster, Barkus, & Yavorsky, 2006).

Based on the literature of risk factors for PPD, predictor variables were selected that were likely to correlate with the dependent variable. The dependent variable in this study (the PDSS score) is described by the following predictor variables, namely a history of psychiatric illness, antenatal depression in recent pregnancy, postpartum blues, feeling negative or ambivalent about expecting this baby, fearful of childbirth, lack of support from the baby's father, lack of support from friends, infant temperament, concern about health related issues regarding the infant, like colic, sleeping and feeding problems, and allergies, difficulty conceiving, and life stress.

The stepwise selection method was used in the multiple regression analysis. This method relies on computer software to select the order in which predictor variables and is based purely on mathematical criteria (Field, 2005). The predictor variables are entered in sequence and the software selects the predictor that has the highest simple correlation with the outcome. The predictor variable is retained if its addition contributes to the model (i.e. the theory that the predictor variable is likely to indicate a high PDSS score). The remaining predictor variables are, however, subjected to re-testing to determine if they are still making a contribution to the success of the model. These remaining predictor variables are removed if the re-testing indicates that they are no longer contributing significantly (Field, 2005). Employing the stepwise method therefore

ensures that the minimum possible number of predictor variables is included to predict the outcome variable, in this case, the total PDSS score.

According to Foster et al (2003, p. 30), multiple regression is used to answer three types of question, namely:

1. What is the relative importance of the predictor variables included in the analysis?
2. Does a particular variable add to the accuracy of the prediction?
3. Given two alternative sets of predictors, can it be determined which is more effective? For example, can PPD be predicted better by the mother's demographic characteristics or by obstetric factors?

In simple regression the degree of the relationship between two continuous variables is expressed as a correlation coefficient which may vary from -1.00 to +1.00. When two variables are correlated, then predicting the score on one variable is possible if you know what the score on the other variable is. The stronger the correlation, the closer the scores will be to forming a straight line, and the more accurate the prediction will be (Foster et al., 2003). The scattergram for simple regression, which depicts the relationship between only two variables, is a visual representation of the following regression equation:

$$y = c + m(x)$$

In this equation, *y* (plotted on the vertical axis) is the predicted score on the dependent variable, *x* (plotted on the horizontal axis) is the score on the independent variable, *c* is the constant, or the intercept or point at which the line crosses the *y* axis, and *m* is the regression coefficient or weight, which indicates by how much *x* must be multiplied to obtain the predicted value of *y*. Put in another way, the regression coefficient (*m*) is the amount of change in the dependent variable (*y*) resulting from a one-unit change in an independent (*x*) variable when all the other predictor variables are held constant. The difference between the predicted *y* score and the actual score is known as the residual (Foster et al, 2006).

In this study *y* is the PDSS score. The predictor variables (*x*) are categorical, for example, the presence of a history of depression is coded 1, and no history of depression is coded 0. When the predictor variable (*x*) has a value of 0, the variable disappears and the leaves only the constant value (*c*). All participants therefore start off with the constant value. The presence of any predictor variables (*x*) are therefore expected to add to the overall PDSS score.

Multiple regression is simply an extension of this correlation principle. In multiple regression a prediction to one variable is based on several other variables as opposed to just one, as in simple regression. For every predictor variable that is added, a coefficient is added, so that each predictor variable has its own coefficient (Field, 2005). This enables not only the prediction of the dependent variable, but also determining the relative influence of each of the predictor variables on the outcome – the PDSS total score – and gives an indication of the combined ability of the predictor variables in

predicting or explaining the variation in the outcome variable (y). The multiple regression equation is therefore slightly more complex:

$$y = c + m_1(x_1) + m_2(x_2) + m_3(x_3)...m_\kappa(x_\kappa)$$

The aim of multiple regression is to find the regression coefficient (the weight, i.e. m_1, m_2, etc.) for each of the predictor variables (x_1, x_2, etc.) which will produce the values of y which are closest to the actual values (Foster et al., 2003). The regression coefficients therefore maximise the correlation between the predicted y values and the combination of the predictor (x) variables. In this study the relative influence of a number of predictor variables (the known risk factors for PPD) as they relate to the outcome variable (the PDSS total score) was investigated.

When two or more predictor variables correlate strongly with each other, known as collinearity, then making assumptions about the relative contribution of each predictor variable is difficult. SPSS is able to determine if collinearity in the data was present.

SPSS 19 was used in this study to examine the following requirements for multivariate analyses, namely sample size, independence of residuals (Durbin-Watson test), presence of multicollinearity (the variance inflation factor or VIF, the tolerance statistic, and collinearity diagnostics), the influence of outliers (casewise diagnostics using Cook's distance, and Mahalanobis Distance), homoscedasticity and non-linearity (plots of the standardised predicted values against the standardised residuals), and normality of residuals (Field, 2005).

7.9.3.3 Correlation of PDSS, EPDS, and QIDS-SR16 total scores.

Statistical analyses were performed on the total sample (N = 365) to determine the comparison of the participants scores across the three screening scales. Descriptive statistics for the three screening scales were calculated and the frequencies were determined according to the participants screening results at the published cut-off thresholds recommended. Chi-square tests were performed on the categorical depression screening status to compare participants who scored positive for symptoms of PPD on the three measures.

The Pearson correlation was used to investigate the relationship between the PDSS and the EPDS, and the PDSS and the QIDS-SR16. The Pearson correlation measures the strength of the correlation (linear dependence) between two variables. It is sometimes referred to as "Pearson's *r*" and is denoted by *r*. It yields a value that may range from +1 to -1. The stronger the association between two variables are, the closer the Pearson correlation coefficient will be to either +1 or -1, depending on whether the association is positive or negative. A value of 0 signifies that there is no association or linear correlation between the two variables. A general guideline is that a positive coefficient *r* of 0.50 to 1.00 indicates a strong positive association between the variables, and a negative coefficient of -0.50 to -1.00 indicates as a strong negative association between the variables (Cohen, 1988).