CHAPTER II

REVIEW OF LITERATURE
CHAPTER II

REVIEW OF LITERATURE

Review of literature is an important step in every scientific enquiry which influences conceptualization of a relevant question and design of an enquiry. Relevant literature reviewed for the task of development of a valid tool on aptitude towards nursing is presented in this chapter under the broad headings of assessment of nursing aptitude and process of tool development.

2.1 ASSESSMENT OF NURSING APTITUDE

Assessment of aptitude had been an area of interest and investigation of many researchers in the past. Several aptitude tests namely Scholastic Aptitude, Clerical aptitude, Teaching aptitude and the like are available and are in use in India and abroad. However interest shown by nurses in development of a valid tool to assess nursing aptitude was less and hence there is a paucity of research in this area especially in India. A summary of work done in this area is presented here under the following headings.

2.1.1 Nursing aptitude and its assessment

2.1.2 Nursing aptitude and academic performance

2.1.3 Nursing aptitude and clinical performance

2.1.1 Nursing Aptitude and its assessment

Aptitude (Pataliah, 2004) is the natural ability or the readiness to learn an activity. Traxler (Mangal, 2008) describes aptitude as a condition or a set of qualities indicative of the ability to acquire knowledge and skill under suitable training. However aptitude is not
intelligence, interest or achievement. Thus it is imperative that a test of aptitude should detect the potential ability for satisfactory performance in a job. In the words of Munn (Mangal, 2008), aptitude tests enable selection of candidates most likely to acquire skills in question and acquire them to a desirable level of proficiency. Aptitude tests (Navarro, Vitamog, Tierra, & Gonzalez, 2011) are capable of uncovering hidden talents, identifying strengths or weakness of an individual, predicting scholastic performance, identifying the match between professional demands and personal abilities and even direct how one should work with handicaps.

The review in this section will be presented in terms of available tools to assess nursing aptitude followed by the components to be considered in development of a valid test. Challenges in development of NAT are also briefed.

2.1.1.1 Tests available to assess nursing aptitude

Jiwan (2011) developed NAT with three subscales namely intelligence (cognitive), emotional quotient (affective) and psychomotor (conative) skills for the purpose of selection of candidates for nursing courses. The areas assessed in cognitive subscales were knowledge from applied sciences, judgement, language, reading skills and verbal reasoning. Self-awareness, motivation, self-regulation, empathy and social skills were assessed in the affective subscale. The conative scale had 10 tasks to be performed for which time was monitored. This 129 item composite scale was tested among a total of 115 diploma and B. Sc. nursing students. The internal consistency reliability of the cognitive and affective subscales was found to be .85 and the conative (KR 20) was .72. However though the tool was found reliable, owing to the methodological constraints in terms of its predictive validity and the
sample selection, the test cannot be recommended for assessment of nursing potentials of students seeking entry to nursing programmes in India.

Pataliah (2004) in India developed a tool to assess nursing aptitude of nursing professionals in the areas of memory, knowledge, vocabulary, judgement and ability to follow instructions. The study among 520 nursing professionals found an association of nursing aptitude with years of experience (higher mean aptitude scores among nurses with 11-20 years), religion (higher mean aptitude scores among Christians) and professional qualification (higher mean aptitude scores among graduate nurses). Variables such as age, sex, position and type of hospitals had no association with nursing aptitude scores. The study revealed that aptitude grew with experience and professional training. Thus, Pataliah recommended assessment of nursing aptitude of students on admission to course as well as assessment of teaching aptitude of faculty of nursing educational institution.

NAT developed in Philippines (Magno, 2010) is available for use in India and assesses abstract reasoning ($\alpha =.43$) verbal reasoning ($\alpha =.75$) and numerical reasoning ($\alpha =.75$). This tool assessed intellectual abilities than technical aspects of nursing job and had overall internal consistency reliability ($\alpha$) of .83. In line with this study, a tool was developed (Andrade & George, 2013) by the researchers with subscales namely abstract reasoning, verbal ability, reading comprehension, academic aptitude and numeric ability. The items of these subscales were Multiple Choice Questions (MCQ). In addition, a five point likert scale for assessment of nursing qualities namely caring, compassion, commitment, and connectedness was also designed. The aforementioned subscales of nursing qualities were termed as nursing potentials in Magno study. Though the subscales were named similar to
Magno study, items were developed by the researchers keeping in mind Indian socio-cultural background. The first trial revealed the internal consistency reliability ($\alpha$) of .932 of nursing qualities and split half reliability ($r$) of .42 of nursing aptitude. The scale was modified based on the limitations of the study, which replaced academic ability, verbal and abstract reasoning with decision making and mechanical aptitude. However, the modified scale had lower reliability coefficient and hence was regarded unsuitable for field test.

Tests are used to assess nursing potentialities or aptitude in the developed countries. Some of them are: 1) Test of Essential Academic Skills (TEAS): Used for student selection and was found to be predictive of first semester nursing program success rather than attrition (Pitt et al., 2012). 2) Nurse Entrance Test (NET): A scholastic Aptitude test, manufactured by Educational Resources Inc., identifies students who are at high risk for attrition from graduate nursing programmes through reading comprehension scores of NET 3) HESI (Health Education Systems Inc.) Admission Assessment exam: Used to predict success in the first semester nursing courses and found to be highly correlated with final course grades (Knauss & Wilson, 2013). 4) National League for Nursing (NLN) Admission Test: Predictive of NCLEX RN success 5) American College Test: Predictive of NCLEX RN success (Pitt et al., 2012) and 6) Mosby Assess Test: Predictive of NCEX RN success.

### 2.1.1.2 Components of nursing aptitude test

As the available tools developed in US or Philippines were not readily usable in Indian setting and tool developed in India had limitations for use, review of literature with a purpose to prepare a blueprint for development of a tool to assess nursing aptitude was continued. An integrated review of available instruments on non-cognitive construct instrumentation and
measurement revealed the available instruments in nursing required further psychometric testing on a larger random sample (Megginson, 2009). This review recommended focus on assessment of discipline specific attributes identified using a grounded theory approach (Megginson, 2009) apart from review of existing reports and opinions about qualities expected of nurses.

Nursing literature reported several characteristics and qualities expected of nurses (Andrade & Mayya, 2013a). Varghese (2010) reported deterioration of virtues over the years at different levels namely prudence, justice, temperance and fortitude among nurses. Her article reminded nurses to possess emotional and spiritual quotient inclusive of empathy, self-control, self-motivation and social competence in interpersonal relationships. Jeyadeepa (2010) was of the opinion that nursing is a common sense work and hence nurses should possess the skills of emotional stability, keen observation, sound judgement and meticulous analysis. However in a study carried out with an assumption that nurses because of the nature of their work would have more emotional intelligence in-fact revealed nurses and masters of nursing students had the lower Emotional Quotient scores (Adams, McCabe, Zundel, Price, & Carey, 2011).

A study among undergraduate nursing students (Ward, Cody, Schaal, & Hojat, 2012) reported decline of empathy. Francis (Entwistle, 2013) emphasized on assessment of the prospective students caring and compassionate qualities to be a good nurse than the technical aspects of nursing or the knowledge. Nursing literature too propagates caring and compassion as the core values of nursing however these qualities are elusive to measure (Watson, 2009; Flynn & Mercer, 2013; Beck, 1999). In a review (Romanelli, Cain, & Smith, 2006) the
authors pointed out that the role of emotional intelligence in academics is less studied moreover expressed concern about the abstract nature of the concept of emotional intelligence. A study conducted in Philippines (Magno, 2010) among nursing students revealed that nursing traits were independent of nursing aptitude and achievement.

Lalitha (2010) reported ability to utilize many varieties of communication, demonstration of sensitivity and gentleness, abilities of work organization, problem solving and teaching, the qualities of accuracy, consistency, thoroughness and continuity and all of these performed in a caring manner and to appropriate standards are the essential competencies in nursing. She was of the opinion that nursing being a feminine skill, a natural part of a woman’s role require little development or reward.

While assessment of non-cognitive constructs were given emphasis, assessment of math, reading, English language abilities were also carried out as essential component of nursing aptitude. While basic math skills, reading comprehension, vocabulary/general knowledge and grammar were some of the components of HESI Admission Assessment test (Knauss & Wilson, 2013), English, science, math and reading were the components of TEAS (Newton et al., 2010). The overall TEAS score were predictive of pass in medication calculation assessment in nursing than the math aptitude in one of the studies in USA (Newton, Harris, Pittilgio, & Moore, 2009). In addition to similar finding, another study revealed reading comprehension had a moderately strong positive relationship with medication calculation assessment. The authors (Newton et al., 2009) recommended assessment of reading comprehension as a contextual factor stating that if a student is unable
to read a question correctly or do not understand it, then she/he may fail a medication calculation assessment test.

Arkell and Rutter’s (2012) study to assess numeracy skills and attitude towards numeracy skills among nursing, pharmacy and midwifery students revealed nursing and midwifery students had poor numeracy skills. Only 15-18% of nurses and 26% of midwives expressed that they liked working with equations and formulae. A study by the researcher (Andrade et al., 2013c) using an Indian version of standardized timed numeracy test (Vohra, 1994) also revealed similar finding wherein 52.1% of the entry level graduate nursing students had poor numeracy skills and only 3.3% had good skills.

Assessment of critical thinking is essential in nursing according to the National League for Nursing Accrediting Commission (Clifton & Schriner, 2010). Further a study (Wangensteen, Johansson, Bjorkstrom, & Nordstrom, 2012) to identify possible predictors influencing nursing competence revealed critical thinking was an essential component which contributed substantially towards making judgements based on evidence. Nursing of today expects professionals to use evidence based approaches (Polit & Beck, 2012) to care and demands nurses to master skills to interpret research findings accurately and effectively.

2.1.1.3 Challenges in assessment of nursing aptitude

While the nursing literature challenges novice researchers to include cognitive and non-cognitive constructs in assessment of aptitude, the characteristics of current nursing students pose a challenge. The intent of students joining nursing programmes these days are different more so are their learning habits (Bhasan, 2010). Today’s nursing students doubt their readiness to college (Pardue & Morgan, 2008). While studies in India (Andrade, 2012;
Andrade et al., 2013a, Andrade et al., 2013b; Andrade et al., 2013c) revealed that majority of nursing entrants chose nursing primarily because of secondary gains of being in the profession, a study in Mauritius reported the choice of nursing was to seek stability in a government job (Hollup, 2012).

Bhasan (2010) quotes the view of American Association of Colleges of Nursing, which says, nursing work for the future is complex and challenging, hence would require a nurse worth far different skills than what is expected today. Demands of nursing profession (Andrade & Mayya, 2013a) are many however the boundaries of practice are not clear unlike that of other established professions like medicine, teaching and engineering. Outcomes of interventions in nursing are less predictable because human beings are unique and there are no laws governing human behaviour (Black & Jacobs, 1993). Literature search in nursing unveiled the reality of lack of a single discipline specific meta-theory in nursing on which to conceptualize the concept of nursing aptitude and to base the items of the tool. Development of a test to assess nursing aptitude amidst the limitations within the profession is a challenge moreover such a venture is less welcomed by nurse academicians in India owing to acute shortage of qualified nurses globally and nationally.

The review above revealed that assessment of nursing aptitude had been the interest of few researchers in developing countries. While standardized tests are available in developed countries, India is in the look-out for similar valid tests for its setting. The review revealed the concept of nursing aptitude is broad and is elusive for measurement. Orientation to the construct of aptitude and aptitude testing as well as the review on essential elements to be included in NAT provided a basis for development of a blueprint for test construction.
2.1.2 Nursing aptitude and academic performance

Institutions may select candidates for professional programmes on the basis of recommendations or referrals, assessment of interest in the course, performance in an interview, tests of aptitude, personality, intelligence and the like or merely on the basis of Grade Point Averages (GPA). Studies conducted to assess predictive property of these tests on academic success revealed efficacy of each one of them in few occasions. Interviews (Aldrich, 1987; Foley & Hijazi, 2013) for selection of candidates for professional programmes predicted academic performance though there were likelihood of false negatives. A few aptitude tests which predicted academic success are described in this section.

In an integrated review (Pitt et al., 2012) authors reviewed studies which used NET, ACT, NLN Admission test, Mosby Assess Test and TEAS and reported that the scores of these tests had a positive correlation with the academic success in preadmission nursing programs. Similarly Rhodes (2008) presented a large review on predictors of academic success in which prediction of academic success was attributed to NET and TEAS over the traditional use of Grade Point Average (GPA) score. Rhodes explored cognitive, non-cognitive and demographic variables to predict academic success and found Science GPA, reading and comprehension and math scores of NET predicted early academic success. While demographic and non-cognitive attributes did not predict Baccalaureate (BSN) program completion, reading comprehension, math scores and previous degree did predict program completion.

A few studies used cumulative GPA to predict academic success. Newton and Moore (2009) reported that TEAS (math, science, reading and English aptitude) was able to predict
first semester nursing program success than GPA and concluded that early academic achievement depends on admitting academically strong candidates into the program. TEAS provided additional data on students’ core knowledge beyond GPA. In another study in which she compared the fall and winter cohort of students with regard to their TEAS and pre nursing GPA, the scores differed significantly between groups. TEAS appeared to be the predictor of success in nursing major (Newton, Smith, & Moore, 2007). Hernandez (2011) also found composite TEAS and scores of Science and reading subscales of TEAS were predictive of student graduation and NCLEX RN licensure examination than GPA scores. However author recommended weightage for both GPA and TEAS as preadmission criteria because GPA also could predict some outcome variables in the study.

A study (Newton et al., 2009) held with a purpose to identify the relationship between BSN students’ math aptitude and ability to pass medication calculation test in single attempt revealed a weak positive relationship ($r = .264 \ P = .003$). The study also revealed as math aptitude decreased, number of attempts to pass math calculation test increased ($r = - .336 \ P=.001$). Study observed that students with strong overall nursing aptitude scores were more likely to succeed in the medication calculation assessment. Based on the findings, authors recommended assessment of core academic skills of students on entry to BSN programme.

A hold on English language skills was a predictor of completion of nursing programme. (Knauss & Wilson, 2013; Pitt et al., 2012). Francis (2013) in his report too emphasized nurses to possess good English language skills (Lintern, 2013). In a study (Newton et al., 2010) to find the relationship between reading aptitude, English aptitude and medication calculation assessment test, reading aptitude accounted for 7.3% of variance in
medication calculation assessment test and English aptitude was not related significantly ($r = .135, P = .32$). There was weak positive relationship between reading aptitude and writing ability and a strong positive relationship between English aptitude and writing ability. However in a multi-linguistic country like India, where majority of the patients speak local languages the decision on level of proficiency required of English language would require a considerable thought (Andrade & Mayya, 2013a).

A descriptive study in Philippines (Ong, Oalompon, & Banico, 2012) assessed predictors of nursing licensure examination with nursing aptitude test, IQ test, composite score of math, science, reading and English tests, college GPA and pre-board examination performance. Though all variables showed a significant relationship with licensure examination scores, regression analysis revealed that pre-board examination performance and college GPA could only predict the success in licensure examination. The author found that the scores of all variables were above average except for the pre-board examination scores and hence alerted education institutions to be careful while promoting students to next academic year if they do not exhibit requisite competencies. On the contrary a pilot study held by the researchers on nursing aptitude and academic performance in India revealed that nursing aptitude test scores were predictive of academic performance than the PCBE scores. The mean PCBE scores were higher than the nursing aptitude scores (Andrade, Mayya, & George, 2014).

Bhasan (2010) recollected the view of California Post-Secondary Education Commission which attributed the problem of student attrition in nursing courses, to admission of minimally qualified students. She adds, in India too, nursing teachers observe problems in
academic skills including reading and comprehension, arithmetic ability especially ability of students to work with fractions. She highlights the need for guiding students and their parents in India to take decision on continuation of the course rather than forcing or motivating to continue when one is incapable to perform the expected skills.

Higher attrition rates in nursing occur in first year of nursing program. A study among 157 students of Associate Degree in Nursing in USA (Knauss & Wilson, 2013) revealed that HESI Admission Assessment test scores were positively, strongly and significantly correlated with Nursing - I (accounted for 27% of variance) and Nursing - II (accounted for 21% of variance) grades. The HESI Admission Assessment test was inclusive of math, reading, general knowledge and vocabulary and grammar subscales and each was significantly correlated with final course grades. The author recommended assessment of applicant’s vocabulary and general knowledge as a measure of student’s ability to succeed in a nursing program. Similar studies with HESI Admission tests were also predictive of academic success (Underwood, Williams, Lee, & Brunnert, 2013; Chen & Voyles, 2013) among students of Bachelor of Science in Nursing and Associate Degree in Nursing

A study in Philippines (Navarro, Vitamog, Tierra, & Gonzalez, 2011) assessed academic performance, nursing aptitude and college admission test to predict success in Nursing Board examination. Multiple regression analysis revealed all three variables were effective predictors. The researchers recommended the eligibility to nursing programmes as, a score of 85% or higher in college admission test, a pass score in nursing aptitude test and ability to handle the mental demands of the profession.
The efficacy of standardized aptitude tests in professional institutions and organizations (except nursing in India) in India and abroad promoted their use even today. The review affirms the property of prediction of aptitude tests in terms of academic success. The reviewed literature is a resource and motivation to nursing professionals in India who are yet to be convinced of use of aptitude tests in nursing. There is paucity of research in this area in India and hence considerable work would be essential.

### 2.1.3 Nursing aptitude and Clinical performance

Excellence in clinical performance is the hallmark of “good” nursing care and reflects aptitude, cognition, affect and skills. Over the years attempts have been made to measure clinical nursing performance in the patient care unit of staff and students (Krichbaum et al., 1994). The preliminary tools comprised of nursing attributes which were rated by faculty using judgement. After 1962, evaluation tools comprised of items developed based on criterion referenced standards namely knowledge, amount of guidance required, time spent in teaching, demonstration of dexterity and coordination, and were observable and measurable. In the last two decades attention is paid to rating performance on patient care behaviours. Scales to evaluate clinical performance of students are in place in nursing schools and colleges in India. However validity and reliability of many such scales are not reported.

Evaluation of clinical performance reflects knowledge, skills and attitudes of the one being evaluated (Cassidy, 2009). An integrative review by Wilkinson (2013) identified that appraisal of registered nurses included areas of basic care and assessment; communication, coaching and teaching; personal and professional development; knowledge of principles of care including legal and ethical issues; interpersonal relationship, teamwork, leadership and
management; critical thinking and research aptitude problem solving and nursing process. Such elements are also a part of student clinical evaluation in training institutions however the concern in student or staff evaluation in nursing is the agreement regarding benchmarks for nursing. There are different interpretations of competence and competence assessment in nursing (Cassidy, 2009).

Several factors within the environment of evaluation should be meticulously dealt with during the process of evaluation to reduce bias or errors in evaluation. Authors of an integrated review (Pitt et al., 2012), draw attention to a study from Hong Kong among third and fourth year nursing students, which found higher levels of anxiety resulting in more errors during clinical evaluation. The presence of faculty (Nell, Kristen, & Sharon, 2008), relationship between student and a teacher (Lambert, Lambert, & Petrini, 2004), varying acuity level of patients admitted to the unit, qualification and experience of faculty evaluating the performance of students (Andrade, 2014) and the trend of grade inflations in nursing (Scanlan & Care, 2004) are few threats to development of a tool to evaluate clinical performance.

There is a paucity of research related to relationship between nursing aptitude and clinical performance. A study in USA (Krichbaum et al., 1994) explored the predictors of clinical performance among GPA, college aptitude (American College Test - ACT), age at admission, college credit earned prior to entry, moral reasoning and ethical theory test. Age, moral reasoning and academic experience (college credit earned) had relationship to clinical performance. ACT had higher correlation compared to GPA, however during logistic regression analysis both did not predict a relationship with clinical performance.
Clinical performance of nurses and nursing students is evaluated differently in different contexts in different organizations. The tools or criteria used for evaluation differ from institution to institution. However the investigators of this study developed a clinical evaluation proforma for evaluating the nursing care planned and implemented by first year nursing students based on the objectives of Nursing Foundation course. The reliability (Intra Class Coefficient) of the clinical evaluation proforma was .764. The scores of NAT INDIA, PCBE and clinical evaluation were correlated (n=41) which revealed no significant correlation of clinical evaluation with NAT INDIA or PCBE (Andrade, 2014).

The paucity of research in India and abroad on assessment of relationship between NAT and clinical performance is primarily related to lack of clarity on competence and its assessment (Cassidy, 2009) as well as universal agreement on benchmark for evaluation in nursing. There is scope for research related to development of tools to evaluate clinical performance and predictors of clinical nursing performance (Pitt et al., 2012).

2.2 PROCESS OF TOOL DEVELOPMENT

The process of tool development is logical, systematic, structured, iterative (Rattray & Jones, 2007) labour intensive and lengthy process demanding fair amount of statistical sophistication. It is because, the developed measure requires to internalize the characteristics namely, conceptual clarity, reliability, validity, responsiveness or sensitivity, interpretability and adaptability to language and culture (Maloney & Chaiken, 1999).

Steps and methods used for development of a measure are almost similar in several studies. The methods recommended by proponents of scale development and used by a few researchers are briefed in this section under the following headings.
2.2.1 Steps of tool development

2.2.2 Choice of a measure

2.2.3 Measures of validity

2.2.4 Standard setting using Receiver Operating Characteristic Method

2.2.5 Table of norms

2.2.1 Steps of tool development

An expert committee (National Research Council, Committee on US Naturalization Test Redesign, 2004) of Board of Testing and Assessment (BOTA) recommends that test development process should be open, honest, transparent and accountable. Naturalization tests are administered to large number of people, to allocate social benefits and thus tests should be fair, valid and reliable. The development of test instrument requires a sound scientific basis and the process should be documented systematically. Thus a detailed plan for its construction should be created with help from an advisory panel and reviewed by an oversight committee. The steps for test construction recommended by the committee were: 1) Specification of the purpose of the test and the inferences to be drawn 2) Development of content frame works 3) Building test specifications 4) Creating potential test items and scoring rubrics 5) Review and pilot test items 6) Evaluation of quality of items 7) Assembly of test forms and 8) Setting passing scores.

Psychologists (McClelland, 1973) are critical about the qualities of tests and the methods of development of a measure. McClelland suggested six guidelines for a novice researcher on test development namely, 1) The best testing is criterion sampling (If one is assessed for his ability to drive a car, he/she should be given a driving test. 2) Test should be
designed to reflect changes in what the individual has learned. 3) How to improve on the characteristic tested should be made public and explicit, 4) Test should assess competencies involved in clusters of life outcomes, 5) Test should involve operant as well as respondent behavior, 6) Test should sample operant thought patterns to get generalizability to various action outcomes.

In a workshop (Farh, 2007) on survey instrument development, the expert highlighted major three requirements for a tool development namely, 1) Definition of the construct and developing a conceptual meaning. If domain clarity is low then scale developer should proceed to collect behavioural incidents, classify behaviours into categories, form dimensions from categories and define the domain. 2) Development or choice of a measure consistent with the definition. 3) Logical analysis and empirical testing. Further the expert introduced six steps of tool development namely, 1) Item generation. 2) Questionnaire administration 3) Initial item reduction. 4) Confirmatory factor analysis. 5) Convergent and discriminant validation and 6) Replication. These steps are on par with the stages recommended by a reviewer of Scale development (Hinkin, 1995) namely item generation, scale development and scale evaluation. DeVellis (1991) in his book on scale development outlined the general guidelines for development of measurement scales namely 1) Clarity of phenomenon to be measured 2) Generation of item pool 3) Choice of a measure (format of measurement) 4) Review of initial item pool by experts 5) Validation of items 6) Administration of items to a development sample 7) Evaluation of items and 8) Optimization of scale length.

Polit and Beck (2012) in their book on nursing research, described five stages in the process of development of self-report scales namely 1) Conceptualization and generation of
items; 2) Evaluation of items through internal and external review; 3) Preliminary assessment by administering to a development sample; 4) Analysis of scale development data: item analysis or factor analysis; and 5) Scale refinement and validation including establishment of cut off points, norms and the manual. Further Polit explained the relevance of translating scales into other languages and the process of validation of the translated tool.

Most testsSCALE construction studies in nursing too described the steps used for test/tool construction. Bu X and Wu Y B (2008) used BU and Jezewski’s theory of patient advocacy and Fishbein and Aizen’s definition of attitude to guide the development of tool on attitude toward patient advocacy. The purpose of the research was to develop and test psychometric properties of the instrument to measure nurses’ attitudes toward patient advocacy. Steps used in their research were: 1) Defining the construct of attitude toward patient advocacy, 2) Generating items and establishment of content validity 3) Survey studies to examine test retest reliability and preliminary internal consistency reliability and 4) Test of construct validity and internal consistency reliability.

In a study (Arranz, et al., 2004) to develop a haemophilia disease specific quality of life questionnaire, the author used both qualitative and quantitative methods. In phase 1, the author held semi-structured interviews with patients and care providers and qualitative analyses was used to identify domains of assessment. The item pool was generated. In phase 2, the developed items were evaluated by experts with respect to acceptance, relevance and comprehensibility. On the basis of suggestions and rating, items were reformulated. In phase 3, the scale was tested and subjected to psychometric analysis.
The steps of tool validation vary slightly when an existing tool is accepted for a different population or cultures wherein the language used is not that of the original tool. A study (Lopes, Azeredo, & Rodrigues, 2013) to evaluate the psychometric properties of the scale on interpersonal communication assessment followed the steps of translation, back translation, comparison of both versions (original and back translated) and evidence of the scales validity in the new context. The study selected bilingual translator to translate the tool, which was then validated by a panel of bilingual experts with experience in the renowned scientific and academic community. The authors condensed the 54 item tool to 23 item test through factor analysis to make it user-friendly. Test tool was found to possess very good temporal stability (test retest reliability (r) of .74) as well as internal consistency measure (α) of .84 and had high levels of general correlation with the scores of original version.

When developing questionnaires the tool developer should pay attention to three issues namely: 1) What will the questionnaire measure? 2) What type of scale could be used and 3) How to generate items for the questionnaire? (Rattray & Jones, 2007). Questionnaires are widely used in medical and nursing education research. In an effort to introduce medical researchers to the process of development of questionnaires (Artino et al., 2014), for the purpose of programme evaluation and research, authors described systematic, seven step design process namely: 1) Conduct literature review. 2) Conduct interviews or focus group discussion. 3) Synthesize literature and focus group/ interviews. 4) Develop items. 5) Conduct expert validation. 6) Conduct cognitive interviews and 7) Pilot test.

Review on steps of tool development reveal that the process of tool development demands meticulous attention to each step of test construction without which interpretation of
results of study would be misleading. Researcher should strive to work with patience to meet
the norms and standards of tool development.

2.2.2 Choice of a measure

Testing cannot completely assess one’s talent or competency, however it provides a
relative measure of strength or weakness in relation to others (Clifton & Schriner, 2010). Measurement of a known or defined variable is based on structured predefined assessment
criteria. Forced choice rating scale, extended or restricted response and supply type items, Multiple Choice Questions (MCQ), visual analogue scale and checklist or self- report scales are often used in education programmes. Self-report scales used could be of frequency, Thurston, Rasch, Guttman, Mokken, or Likert type (Rattray & Jones, 2007). Most of the measurement scales in education programmes are questionnaires to assess knowledge; attitudes, belief, intension or satisfaction; cognition; emotion and behaviour (Rattray & Jones, 2007). Questionnaires are handy, quick to complete, relatively economical and are easy to analyse, however are not the method of choice when the knowledge of subject area is little.

A number of scales are used in educational institutions to measure the desired outcome. Goal of any educational institution is to produce competent graduates to serve the population. An appropriate form and or format of a measure are essential to ensure higher standard in educational assessment. Nursing has wide subject knowledge however most of the knowledge is from related bio-psycho-social sciences. MCQ items demand critical thinking skills and hence have been used in nursing examinations in India and abroad.

MCQ items can be used to assess skills of higher order thinking namely evaluation, synthesis, application, and analysis. Most of the Licensure examinations in nursing use MCQ
items. Clark and Watson (1995) quote and justify the analysis of Comrey (1988) about MCQ items, ‘MCQ formats are more reliable, give more stable results, easier to administer and score, cover large content area in a shorter time span and produce better scales’. One should possess sound subject knowledge to respond to MCQ items. MCQs demand application of more than one concept to answer the questions (Su, Osisek, Montgomery, & Pellar, 2009) and thus if framed well, can differentiate students with knowledge and those who require additional training.

However, construction of MCQ items is time consuming and demands greater competency and subject knowledge. While generating items, one should pay attention to the number of alternatives to be included. Thought also should be given to coverage of full range of possible alternatives and the order or position of given alternatives. Alternatives given should be mutually exclusive and collectively plausible.

2.2.3 Measure of validity

A measure is valid when it measures what it is supposed to measure. Scale validity is concerned with whether the variable is the underlying cause of item variation. The adequacy of a scale as a measure of a specific variable is an issue of validity and is assessed in terms of item sampling adequacy (content validity), empirical association with some criterion (criterion related validity) and extent to which it behaves as it is supposed to behave (construct validity).

Face validity or content validity of a scale involves use of experts (Rattray & Jones, 2007). Selection of expert should be on the basis of knowledge of the construct being measured and experience however their willingness and availability is also a concern. There is
no consensus on number of experts however literature is suggestive of 6-10. More experts (up to 20) generate a clearer consensus (Artino et al., 2014) on representativeness, clarity, relevance and distribution of items. In a review (Hardesty & Beardon, 2004) authors opined that the items retained after face validity correlate fairly well in some occasions with those that would have survived an exhaustive process of scale development. The authors recommend retention of an item with at least a minimum of 75% agreement. The authors tested three decision rules for retention of item and found that sum-score rule outperformed the complete agreement rule.

Measurement of target and target of measurement are important for optimal scale development (Clark & Watson, 1995). Face validity is an initial step however is not sufficient by itself. The goal of a scale development is to maximize validity than reliability. Hinkin (1995) expressed, scales might show good internal but may lack content validity due to multidimensionality or inappropriate representation of construct under consideration. A scale constructed with a focus of uni-dimensionality will identify a more homogeneous group. A focus on discriminant and convergent validity will help identification of what exactly a scale measures and what not. Clear conceptualization of the construct/concept and correlation of a measure with related or dissimilar measures is equally important to justify validity of the measure.

Construct validity or the relationship between the items of a questionnaire and the underlying conceptual structure can be assessed through the statistical procedure of factor analysis. Conceptual clarity is essential for confirming relationship between items and the variable measured. DeVellis recommends use of a theory to bring clarity to the construct
being measured which is distinct, specific and sensitive. A study (Roberts, 2008) to develop an instrument to aid college students to assess their interest and self-efficacy towards nursing used the social cognitive theory and career psychology of Betz to construct a 48 item Likert scale. During the validation process the tool revealed significant difference in nursing and non-nursing items. However, the differences in responses of racial and ethnicity groups called for further evaluation of the tool with a larger sample.

A psychometric analysis (Siego, Wong, Keanne, & Grumbach, 2008) on measurement of attributes for success in nursing programs is an appreciable effort of tool development and validation based on a conceptual model of academic success. The researchers strived to assure reliability and validity of the tool. The investigators did a thorough literature search of existing tools and the other relevant literature. The developed tool was validated by content expert and then was pretested for its readability, consistency and comprehension. The tool was pilot tested among two groups of college students with diverse ethnic backgrounds before subjecting, the more than 100 item tool to a validation sample. The authors reported deletion of items which did not load on the construct of subscales.

One of the ways to bring conceptual clarity to the construct being measured is to differentiate what it is from what it is not. For e.g. aptitude, interest, intelligence and achievement are dissimilar constructs (Mangal, 2008). A person may have interest in a profession but may not have the aptitude for the chosen profession or vice versa may also be true. At times, a candidate might choose to seek admission to a course, without interest or aptitude, wherein the reason for the choice could be a social one. A career search instrument (Roberts, 2008) developed to assess interest and self-efficacy perceptions for the career of
nursing, among pre-nursing students was commendable to determine, if nursing was a viable career for aspirants of nursing however not for testing nursing aptitude.

A reliable tool may not be valid however a valid tool is reliable. The valid tool must demonstrate properties of the tool and predict an outcome. David’s Battery of Differential Abilities (DBDA) – Revised version adapted for Indian population is used by psychologists. The subscale of Verbal ability (VA) of DBDA is proven to be a predictor of academic success and VA with Numerical ability seemed to be a good measure of general learning ability (Vohra, 1994). A study (Daley, Kirkpatrick, Frazier, Chung, & Moser, 2003) with an aim to compare the effectiveness of two standardized tests namely Mosby Assess test and HESI Exit Examination in terms of properties, revealed HESI Exit test had better acceptable sensitivity, specificity, positive and negative predictive value and test efficiency compared to Mosby Assess Test.

McClelland, (1973) expressed dissatisfaction on the decisions made from the findings of the intelligence measurement scales, the validity of which were questionable. His study of a group of people with poor and good performance in the academics revealed that there was no difference in the achievement in life between groups except for the difference in the type of professional schools to which they sought admission to. The factors affecting validity of a test are reliability, item difficulty level, item discrimination index, readability and different forms of bias. These are briefly described below.

2.2.3.1 Measure of reliability

Scale reliability is the proportion of variance attributable to the true score of the latent variable (DeVillis, 1991) and is assessed in terms of internal consistency (homogeneity of
items of a scale), temporal stability (constancy of scores from one occasion to the other) or equivalence (relationship between parallel forms of test). The reliability score ranges between 0 and 1. While Nunnally suggests (DeVilllis, 1991) a value of .7 as a lower acceptable bound for alpha, Beck (1999) explains the need for at least a .8 for comparison of groups. DeVellis asserts that an alpha of .85 would be perfectly adequate for comparing groups however, in individual assessments where important decisions rest, higher standards (.9) are deemed essential. The scale developer should specify alpha statistic for separate domains within a questionnaire than reporting for the entire questionnaire (Rattray & Jones, 2007).

There are variations in the recommendations related to required sample size for a tool development venture to estimate reliability. While Hinkin (1995) recommends a minimum sample of 150, Clark and Watson (1995) recommend pilot testing of the scale on a representative sample of at least 300. The reviewers (Floyd & Widaman, 1995) on factor analysis recommend 5-10 samples per item. The traditional thought of larger the sample the better is the reliability is devalued by proponents of scale development. However, the sample size estimation in tool development studies which do not employ factor analysis consider factors namely, homogeneity of population, research design, effect size, participant cooperation and attrition, sensitivity of measures and subgroup analysis (Polit & Beck, 2012).

Higher the test reliability coefficient, smaller will be the error of measurement and greater will be the dependability, the predictive value of the instrument and the confidence attached to the judgements. Clark and Watson (1995) recommend scale developers to strive for a scale reliability of .8. They recommend revision of the scale if reliability value falls below .8, either by writing items if the scale is brief or by eliminating the weaker items from a large one.
Reliability of the instrument can also be expressed in terms of Standard Error of Measurement (SEM), which is dependent upon the standard deviation of the distribution of obtained scores and the coefficient of reliability of the test from which the distribution of scores was obtained (DeVillis, 1991). The estimated SEM provides a rough indication of the amount of error to expect in tests of different length. According to Paul Diederich (Grondlund, 1977), the expected standard error for a test of 48-89 items is four and 90-109 is five. SEM is more useful than reliability coefficient as it is directly applicable to interpretation of individual scores.

A low reliability coefficient indicates items are unrelated to each other. The most common reasons for low reliability are that the items are either too easy or difficult, unclear or poorly worded which fail to discriminate or do not test a unified body of content (McGahee & Ball, 2009). A high reliability coefficient indicates that individual items of the test pull together and those students who did well overall are likely to answer the questions correctly.

In nursing examinations a reliability coefficient of greater than .5 could be considered a good coefficient because most nursing tests cover multiple areas of concept and topics (McGahee & Ball, 2009). The reliability of a scale could be improved by administering longer tests, or by testing on a heterogeneous group of students or by attempting to change the items where 70% - 80% of students answer correctly (McGahee & Ball, 2009).

2.2.3.2 Item development and Item analysis

Farh (2007) outlined four steps of item development namely 1) Clarity of item domain. 2) Development 3) Refinement and 4) Empirical testing. Items can be generated with support from reviewed literature or in consultation with experts. The number of items needed
to adequately assess the construct is not clear however scales containing six to ten items will usually suffice to capture the essence of the construct (Artino et al., 2014).

Knowledge of the specific phenomenon is the most important consideration in item generation. DeVellis (1991) asserts that the phenomenon measured in social sciences is often derived from a theory. The scale developer should strive for a balance between the theoretical construct of interest and the methods of measurement used to operationalize them, because poor measurement exert a limit on the validity of the conclusions drawn from the study. When a variable of interest is not measurable or validated, items could be generated atheoretically (for e.g. Public opinion), the findings of which could be used later to develop a theory. Such scales are used in an applied context to quantify ability or performance relative to some normative standard. Thus attention to the process of scale development from the very beginning is very essential to improve the validity of the scale.

Generation of item pool should include all relevant items pertaining to the construct under study from all alternative theories possible (Clark & Watson, 1995). Initial pool should be broader and comprehensive moreover have adequate sampling (sufficient representation) of items from each content area. Construction of items subscale-wise, can also be thought of in this stage if appropriate.

The decision on whether to form a scale or subscale is very essential especially when new scales are formulated. It makes no sense to combine unrelated items or subscales into a single overall score (Clark & Watson, 1995) instead scale developer should ensure that items within a scale define a single factor. If they do not, then it is right time to decide to split the scale into subscales. While splitting scales the scale developer should ensure that intra scale
item correlation must be higher than inter-scale item correlation. If inter-scale correlation is zero then the subscales remain distinct and there is no point in combining the scores of subscales into an overall score. If criterion is not met, subscale formation should be abandoned.

Meticulous attention needs to be paid during item generation phase. Type and order of items may bias response pattern. Controversial and emotive items at the beginning could be avoided. Presence of both positively and negatively worded items may reduce acquiescent response bias (Rattray & Jones, 2007). Hinkin (1995) reports many scale developers avoid negatively worded items for the risk of introducing systematic error. The distractors selected for MCQ items however should be attractive and equally competitive. If one distractor is not chosen, means the test item is not performing at all. Item construction should be fair and must also pay equal attention to gender, language, culture and other contextual factors (Monica, 2005) including test administration time. Readability and comprehension is to be given a thought when framing items.

Item analysis statistically examines the quality of test items and of the scale as a whole (McGahee & Ball, 2009). Item discrimination indices can be estimated through the difference between the proportion of high scorers and low scorers who selected the correct option or correlation between each item in a test to the total score wherein higher the correlation, better is the item. Higher the average item discrimination indices on the test, more reliable the test as a whole will be. There are no universal guidelines on desirable item discrimination index on a nursing examination however as a general rule, item discrimination of below .2 is a poor question; between .2 to .3 are fair and can be improved upon; and between .4 to .7 are
considered good. However each question needs to be evaluated in terms of purpose of test and purpose of individual item (McGahee & Ball, 2009).

Guilbert (1976) justifies that a discrimination index of .25 and above is a good question and .35 and above is an excellent question. Clark and Watson (1995) recommend an inter item correlation of .15 to .5, wherein they specify an inter item correlation of .15 to .2 for a broader higher order construct and .4 to .5 for a narrower construct. If a question is so critical to the knowledge base of the students, wherein a teacher expects 100% of students to answer it correctly, a point biserial of zero may be the goal (McGahee & Ball, 2009).

For estimation of item difficulty and discriminatory indices, the following steps are usually recommended: 1) Award of score to each student. 2) Ranking in order of merit. 3) Identification of groups of high scorers and low scorers, 27% each and 3) Computation of indices using the formula:

**Difficulty Index:** \( \frac{100(H+L)}{N} \)

**Discriminatory Index:** \( \frac{2(H-L)}{N} \)

Wherein: 
- \( H \): The number of students in the high scorer group who answered the item correctly
- \( L \): The number of students in the low scorer group who answered the items correctly
- \( N \): The total number of students in both the groups.

Item difficulty refers to the percentage of candidates who answered the items correctly. Higher the percentage easier the item is. Item lying between 30% and 70% is acceptable, in that range the discrimination has more chance to be high (Guilbert, 1976).
Frequency of endorsement is a function of difficulty of the item. If more than 80% or less than 20% of responses endorsed one response, then item should be considered for deletion (Rattray & Jones, 2007; Streiner & Norman, 2008). Items which are theoretically important can be retained even if they do not meet the recommended criteria (Rattray & Jones, 2007). A good item in other respect with a difficulty level of 19% would not be rejected in some cases. Test items measuring higher level outcomes are expected to have a lower level of difficulty index (Clifton & Schriner, 2010)

### 2.2.3.3 Readability

The language used while framing items should be simple, straightforward and appropriate to reading and understanding level of the target population. A respondent with modest education should be able to understand an item. It is wise to avoid complex, double barrelled, negatively worded statements with biased language (Artino et al., 2014; Clark & Watson, 1995). Items written should be preferably of vocabulary of target population and clear, unambiguous questions that match the item stem to the response anchor.

The medium of instruction in nursing is English, however English is not the primary language in India. A few pilot studies held for the purpose of study revealed the need for reducing the readability level of the scale developed in English language. Moreover, a study conducted by the investigators among 92 first year B. Sc. Nursing students of a Nursing college in the first week of admission to course in 2012 (Andrade et al., 2013c) to identify the verbal ability of the students joining the nursing course revealed that none had good verbal ability. 79.3% of them had poor and 20.7% had average verbal ability.
Klare (1976) suggested estimation of readability using Flesch Reading Ease (RE) formula. RE assesses the difficulty in reading a passage in English, moreover is a simple approach to assess readability (Klare, 1976). The formula has been used in scientific studies in USA and outside USA (Monica, 2005) though was tested originally in USA population.

The Formula is: \[ \text{RE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}) \]

Wherein: ASL: Average sentence length (number of words divided by the number of sentences)

ASW: Average number of syllables per word (number of syllables divided by the Number of words)

RE scores range between 0 to100. Higher the score of RE, easier the text is to read. The scores are classified as: Very easy (90-100); Easy (80-89); Fairly easy (70-79); Standard (60-69); Difficult (50-59); Very difficult (30-49); and Very confusing (0-29). A score of 60 - 70 is largely considered acceptable. RE is also estimated using Flesch Kincaid formula

\[ \text{RE} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.9 \]

which specifies the grade level. For e.g. A score of five means a fifth grader can understand. However, RE measures only surface characteristics of the text and does not take into account the circumstances in which reader may use the text (Monica, 2005).

2.2.4 Standard setting using Receiver Operating Characteristic Method

Receiver Operating Characteristic (ROC) curve provides useful information for the evaluation of tests and can be used to set standards even in medical education (Tavakol & Dennick, 2012). It is a plot of sensitivity (true positives) and 1- specificity (true negatives) against a series of cut off points. Area Under the Curve (AUC) is a measure of the inherent
validity of a diagnostic test (Kumar & Indrayan, 2011) and the best index of overall performance of a test. The maximum AUC is one. Larger the AUC better is the performance of a diagnostic test. Sensitivity and specificity of a measure is computed against an available gold standard. ROC displays sensitivity and specificity values for various levels of cut off points and decision for a cut off score should be taken in light of the context. ROC provides information which is usable to estimate the cut score that will create optimal differentiation between passing and failing students (Tavakol & Dennick, 2012).

Item difficulty influences the sensitivity and specificity of the test. In the context of achievement tests, sensitivity (True Positive Rate) refers to proportion of correct answers predicted by standard setters which are correctly answered by students and specificity refers to the proportion of incorrect answers which are incorrectly answered by students. One minus specificity is the False Positive Rate. The ability of the test to classify correct and incorrect answers that are answered by students correctly and incorrectly is represented by the AUC. AUC below .5 indicates test does not discriminate between passing and failing students. In order to classify students into pass and fail states, a cut score is calculated using test centred method. Students with marks less than cut scores are labelled as failures and greater than or equal to cut score are labelled as passers (Tavakol & Dennick, 2012).

There are several uses of ROC curve namely 1) Identification of pass and failures (cases and non-cases) correctly; 2) Prevention of misclassification of cases and non-cases; 3) Comparison of two or more tests; and 4) Identification or comparison of inter observer variability. Bias that are likely to affect ROC curve are validity of the gold standard, verification bias, selection bias, test review bias, inter-observer bias, comorbidity bias and interpretation bias.
2.2.5 Table of norms

Testing could be Norm Referenced (NRT) or Criterion Referenced (CRT) in psychological and education testing (Grondlund, 1977). NRT assumes that human characteristics follow a normal distribution and thus the range within which test scores fall correspond to the descriptor ranging from superior to deficient. The norms are usually expressed as percentile ranks, standard scores, age equivalent, gender equivalent or grade equivalent scores. Percentile ranks indicate an individual’s position relative to the norm group. For e.g. If a student obtains a percentile rank of 90 means the student’s performance is equal to or more than 90 percentage of the students in the class (Ford, 2009). Scale norms represent scores of pupil in reference groups however should not be viewed as desired standards.

The test manual should describe test’s norm group to ensure appropriateness of the test, adequacy of norms and applicability of the test to a given individual/group. The manual should describe (Shreemathi, 2001): 1) Representativeness of the group 2) Number of individuals in group 3) Relationship of norms to the purpose of test 4) Type of decisions to be made 5) Amount of error in the scale and 6) Extent to which score agrees with available data. NRTs assume statistical rigor and hence are valid. They provide information about the average performance in a locale (institution/university/state/nation), useful for comparison of students, to identify training needs and to determine standards for student progress (Ford, 2009).

Summary

This chapter summarized relevant literature on assessment of nursing aptitude and the process of tool development. Review guided this research especially to conceptualize the concept ‘nursing aptitude’ and instrumentation of NAT which is described in chapter III.