International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:7, No:2, 2013

Use of Item Response Theory in Medical Surgical Nursing Achievement Examination

Rita C. Ramos

***Abstract*—**Medical Surgical Nursing is one of the major subjects in nursing. This study examined the validity and reliability of the achievement examination utilizing the Classical Test Theory and Item Response Theory. The study answered the following objectives specifically : (a) To establish the validity and reliability of the achievement examination utilizing Classical Test Theory and Item Response Theory ; (b) To determine the dimensionality measure of items and (c) to compare the item difficulty and item discrimination of the Medical Surgical Nursing Achievement examination using Classical Test Theory (CTT) and Item Response Theory (IRT). The developed instrument was administered to fourth year nursing students (N= 136) of a private university in Manila. The findings yielded the following results: The achievement examination is reliable both using CTT and IRT. The findings indicate person and item statistics from two frameworks are quite alike. The achievement examination formed a unidimensional construct.

***Keywords*—**Achievement Examination, Item Response Theory, Medical Surgical, Nursing.

I. Introduction

MEDICAL Surgical Nursing covers greater percentage in the total number of hours in the nursing curriculum. This is also parallel to the licensure examination's test framework wherein it covers two major subjects. Thus; it suggests that these concepts are vital in the entry competency of a nurse as a professional. Additionally, it serves several advantages in the licensure examination. Statistics shows that there has been increasing percentage of failing among test takes in the Nurse Licensure Examination in the Philippines. More than a fourth (37.3%) of the test takers did make it in Nurse Licensure Examination [10]. This is the lowest ever noted percentage in the local licensure examination. Nursing Achievement Examination for Medical Surgical Nursing is aimed to diagnose, prepare and enhance nursing graduate prior to licensure examination. The Nursing Achievement Examination for Medical- Surgical items were developed with the framework of Classical Test Theory. The items are developed based on objectives and number of hours allocated. Furthermore, it can be concluded that all items are representative of concepts in medical surgical nursing. It has been noted that no local literature of similar study found. Health Education System, Inc in United States developed a computerized examination which made use to evaluate the achievement outcomes after the Bachelor of Science in Nursing [6]. This was conceptualized utilizing the framework of Classical Test Theory. This is aimed to assess the level of preparedness of graduate nurse before taking NCLEX. There have been several empirical studies supported the predictive ability of HESI [3], [7]. Thus, it can be concluded conceptualization and administration of standardized examination prior to licensure examination can likely improve and diagnose the weaknesses and preparedness of the graduate nurse [7]. The results accounted from the previous studies can be a springboard for the local setting to develop test items parallel to the nurse competencies. The importance of validity and reliability of this achievement examination is of great importance since this aimed to prepare nurse for licensure examination. Several studies also made used both CTT and ICT in test development and found out limitations and advantages of both theory. [5], [4], [9]. However this study is aimed to apply both the Classical Test Theory and Item Response Theory. This study attempts to establish validity and reliability of the achievement examination using the two frameworks (CTT and IRT).

Objectives

1. To establish the validity and reliability of the instrument using CTT and IRT framework
2. To determine the dimensionality measure of items.
3. To compare the item difficulty and item discrimination of the Medical Surgical Nursing Achievement Examination using CTT and IRT.

R. C. Ramos is with University of the Philippines Open University, Philippines (phone: 6349-5366010; e-mail: rita.ramos@upou.edu.ph).

II. Methodology

A. Participants

There were 137 4th year nursing students from a private university in Manila. These were the entire graduating students of the said institution. They were selected in the basis of the completion of all subjects in Medical Surgical Nursing.

B. Measure

Nursing Achievement Test for Medical Surgical Nursing is comprised of 219 items of multiple choices to measure the medical surgical nursing components in nursing .This was developed and assembled based on the test framework from the Board of Nursing. The framework enumerated the five major subjects of the licensure examination which stated as follows : Nursing Practice I Fundamentals of Nursing, Nursing Practice II (Community Health Nursing, Maternal Child Nursing), Nursing Practice III (Client in Pain, Perioperative Care and Alterations in Human Functioning) , Nursing Practice IV (Alterations in Human Function, Client in Biologic Crisis and Emergency / Disaster Nursing) and Nursing Practice V (Disturbances in Perception and Coordination and Maladaptive Behviors). The Medical Surgical Nursing Concepts are limited only on Nursing Practice 3, 4 and half of portion of Nursing Practice 5 (Only

Disturbances in Perception and Coordination). Test objectives for each major category and its subsets were derived from course syllabi. There were all 11 topics derived from three major categories (Nursing Practice 3, 4 & 5). Major Nursing subjects have both theoretical and clinical (Related Learning Experience) components. Such as the concept of Disturbances in Oxygenation which has 72 hours (See Appendix C). The time referred to (72 hours) is a combination of time spent both in lecture and clinical. The total hours summed up to 408 hours in the 11 subtopics (Nursing Practice 3, 4 and 5). The allocation of the number of items per subtopic was computed by dividing the total hours (408) with the number of hours per subtopic. The total item computed for each subtopic were distributed according to five domains of New Bloom's Taxonomy specifically recall, understanding, applying, analyzing, evaluation and creativity .. The subject area for each subtopic were reviewed and noted to ensure all topics included. Thus, the final structure and draft of the achievement examination was arranged according to the main three parts namely: nursing practice III (100 items), nursing practice 1V (100 items) and nursing practice V (19 items).

C. Procedure

Letter of permission was sent and approved by the Dean of the College of Nursing. The examination was held in an auditorium to accommodate all participants (137). This required help of a research assistant to ensure inquiries and instructions to be well disseminated in this large class size. The participants were instructed based on the purpose and structure of the examination. They were given two hours to accomplish the examination. However it was stressed that they can leave the place even before the allotted time and stay beyond three hours in case they go beyond then. Each participant was asked to note in the examination paper the time they started and ended with the achievement examination. It yielded a mean 1 hour and 48 minutes for 100 students with a standard deviation of 15minutes. There was more than a fourth (26.5 %) who did not write and note down time started and ended. However, majority (73.5 %) complied with the instructions. The minimum no of hours was 55 minutes and maximum 2 hours and 15 minutes. Those who failed to write down were not accounted. There were a few who remained beyond two hours. The examination paper were checked and encoded for further analysis.

D. Data Analysis

Results from the achievement examination were encoded. Total scores ranged from 96 to 155 (SD = 13.0) out of a maximum of 219 (mean = 126, median 125)... The reliability of the achievement examination utilizing the framework of Classical Test Theory was analyzed using SPSS version 11.5 (Chicago, Illinois). Item difficulty and discrimination were computed and analyzed according to formula.WINSTEP version 3.69 [11] was used to assess the following: unidimensionality, hierarchichal ordering of items, person reliability and separation, and item reliability and separation.

III. Results

This presented the utilization of two frameworks namely Classical Test Theory and Item Response Theory on the following aspects: reliability, item discrimination and item difficulty.

A. Comparison of Internal Consistencies

The achievement examination yielded a cronbach alpha of .7546. The following results for reliability were achieved when the examination was divided into three main subparts (from BON) framework: Test III - .7526 (100 items) Test IV - .6029 (100 items) and Test V - .7761 (19 items). The ranges of the Cronbach's alpha of the entire test examination when separated into three parts ranges from .6029 to .7761 respectively.It can be inferred that regardless it is accounted as one dimension or three parts, the resulting cronbach ALPA apparently similar across different forms.

The following findings were derived from the achievement examination using Rasch model on the following inquiries: unidimensionality, hierachical ordering of items, person reliability and item reliability and separation. The Real MRSE = 5.71 and Model RMSE = 5.74. This yields .99 which means the close the value of the coefficient to 1.0, the more closely the data approximate unidimensionality.

The sample yielded a person reliability coefficient of .76. This implies that items are working well together to consistently reproduce a participant's score. The IRT based person fit assessment involves the assessment of stability of an examinee's item response pattern with a set of estimated IRT parameters [2]. The sample produced a person separation statistic of 1.78. The strata formula was used to determine the number of distinct ability strata (HP= (4GP + 1)/3). Thus, it resulted to strata equaled 2.70. This implies that the sample can be grouped and separated into three distinct ability groups. This reflects the possibility of separating the sample into divergent performance levels. The item separation was 5.71 when computed using the stated above formula ((HP= (4GP + 1)/3) resulted to 7.94. Findings suggested that the test items can be categorized into eight subgroups.

The feasibility of dividing it into three distinct ability groups may further be evaluated by considering the following: Nursing subjects (Theories and Clinical grades) into low performing, average performing and excellent performing students. Empirical studies espoused the contribution of nursing subjects to licensure examination results. The higher the nursing subjects grades likely higher scores in the achievement examination. The content of the achievement examination is a combination of theoretical and clinical components. The item separation resulted to eight subgroups. This is not far distant from the actual subtopics of 12 in the licensure examination. Therefore the 12 subtopics can be further examined whether there will be a need to converge some topics. However, further validation with experts in the nursing education to look into the different subtopics.

Item Reliability. The item reliability of the said achievement examination is .97. The SZTD resulted to .00.ZSTD less than 0 indicate greater predictability. It can be

concluded that the reliability results are both consistent with the CTT and IRT framework. The medical surgical nursing achievement examination is highly reliable.

TABLE I
DIFFERENCE OF CTT AND IRT ON INTERNAL CONSISTENCY MEASURES

IRT	
Person reliability	.76
Item reliability	.97
CTT	
Cronbach's Alpha	.7546

B. Comparison of Item Discrimination

1. Item Discrimination and Item Difficulty using CTT

Item difficulty was analyzed using Classical Test Theory. There were 139 items (63.47 %) considered as Average; 80 items (36.52) difficult. The 21 (26.25) out of 80 items considered difficult were all taken and lifted from test III. More than 73.75 % were from Nursing Practice IV (Test 4). This espoused that the following questions resulted to large difference of proportion between low and high group. Apparently, Nursing Practice 4 contained more difficult questions and very good items compared to Nursing Practice 3. Nursing Practice 4 (# 101 – 200))subsets are the following: Fluid and Electrolytes, Inflammatory and Infectious Disturbances, Disturbances in Immunologic Functioning, Disturbances in Cellular Functioning, Client in Biologic Crisis and Emergency and Disaster Nursing. This part contained more subtopics than the other two (nursing practice 3 - #1 - 100; nursing practice 5- #201- 219). This implied that further examination and evaluation of Test 3 need to consider. Majority of reasonably good items , marginal items and poor items were all from test 3 (# 1-100) with the following results : 16 items of average in difficulty in Test 3 (# 1-100) were considered reasonably good items ; 3 items of average in difficulty were considered marginal items and 4 items considered poor items.

Item discrimination. Resulted to the following: of the 219 items, 9 items were poor items (4.10 %), 8 items were marginal (3.65 %), 27 items (12.37 %) were reasonably good items, 11 items (5 %) were good items and 164 items (74.88 %) were very good items. Majority of the items were very good items. This suggested that the achievement examination for medical surgical nursing can distinguish high group and low group. The nine poor items and difficult were as follows : items # 106, 185, 92, 157, 6, 72, 117, 129 and 147...It implies that either both high group and low group scored low or high equally. Six of nine poor items came from test 4 (# 101 to 200) specifically items 106, 185, 157, 117, 129 and 147. The three poor items were from test III (1-100) namely items 6, 72 and 92. Item 106 (Which of the following clients is at risk of developing hyponatremia ?), item 117 (Type of hepatitis that are enteric borne and are endemic in countries with poor sanitation), item (Which of the following statements , if made by a 44 year old female would support a nursing diagnosis of knowledge deficit :early detection of breast cancer) item 147 (To which of the following nursing measures should a nurse give priority in the care of a patient who is receiving vincristine sulfate) , item 157 (Which of the four phases of emergency management is defined as " sustained action that reduces or eliminate long term risk to people and property from natural hazards and their effect ?) And item 185 (Should you checked for hemorrhage, how will you do it?). The need to inquire and clarify on the following items .The above mentioned items were all equally important in decision making skills of a nurse given those type of situations. This may be attributed that there are few instances of clinical exposure and homogenous sample for this study. These findings of items discrimination may lead to deletion of items or clarification of the nature of items for nursing students.

2. Item Discrimination and Item Difficulty using Item Response Theory

Rasch analysis reflects the matching of ability with difficulty of item. The mean of the items was 1 logit below the sample. This signifies that the ability of the participants was higher than the all the difficult items. The independent variables are the person's score and the item's difficulty level. When combined additively, the item's difficulty is subtracted from person's ability. Thus, the relationship of this difference to item responses depends on which independent variable is modeled, log odds or probabilities. This can be implied if the trait level exceeds item difficulty, then the person is likely to answer it correct than incorrect. Based on the results, it can be inferred that as trait level exceeds item difficulty the student is more likely to answer correct answers than incorrect ones. For instance at the item mean (0.00) four items namely: 137, 168, 181, 215, 46 and 81. Those items represent same ability level. The rule mentioned that if trait level equals item difficulty which means the nursing student is as likely to succeed as to fail on those particular items. It can be inferred from the figure that if the nursing student responds correctly to item 187(CVP reading) which is at logit 1, the likelihood that he or she will respond correctly to item 64 (laryngectomy topic). The high probability that the nursing student will incorrectly respond to item 43 (Insulin) if she failed to answer correctly item 113.The result is also similar to the CTT wherein items 187 and 64 considered difficult and average respectively. The Rasch analysis is aimed to determine the relativity of ability with item difficulty. The infit is 1.00 and Average equals to 1.0. This indicates that the data for the items showed goodness of fit because the value was less than 1.5.

The mean of the person is 1 logits which higher than the mean of the items, indicating that this sample exceeds the difficulty of items as characterized by Fig. 1.

There were two potential item gaps (located between items 106, 185, 72 and 92; from 92 and 157; additionally from 157 and 197 and 6). The gaps are not significant considering that it is less than 2.00 logits.

The left side of the figure corresponds to the NAT- MSN participant's ability. It implies that symbols (like #) signify that the higher the participants with high ability in NAT. The items in the right side denote items arranged in hierarchy

(from difficult to less challenging). The less challenging items are located at the bottom whereas difficult ones are on topmost location. It is noteworthy to describe that the challenging and difficult questions found in fig. 1 specifically items # 106, 185, 72, 92 and 157 were also considered difficult in the CTT. Thus this will strengthen the revisiting of the following items for further validation.

As the items become more difficult the person with the highest scores is matched closed with the item. This would support the goodness of fit in the Rasch model.

FIGURE 1
PERSON - MAP - ITEM
<More>|<rare>

```
  4            +
               |
               | 106   185   72
               |
               |
               | 92
               |
  3             + 157
               |
               | 197   6
               |T 129   191
               | 117   147
               | 138   71
               | 69
  2             + 90
               | 164
               | 140   150   218
               | 107   108   126   198   2     88
               | 116   134   153   154   196   68    94
               |S 112   114   205   214   45
             . | 124   135   149   176   183   188   76
  1          ## T+ 109   143   187
       ######## | 128   141   146   177   180   202   206   209
                  211   51    59    67
         . ##### S| 115   133   144   159   162   163   204   216
                  96    97
      . ######### | 113   118   121   175   34    40    99
      ########## M| 145   18    189   200   208   22    25    38
                  47    5     65    8     91
############## | 131   139   182   184   194   195   199   50
      ########## | 130   148   161   169   192   212   23    24
                  42    64
  0        . #### S+M 137   168   181   215   46    81
           ### | 104   120   203   26    31    33    43
           ### T| 122   201   32    56    73    87
               | 105   119   136   14    160   207   58
               | 11    166   174   21    28    36    52    74
                 77
               | 152   179   186   20    217   3     54    80
                 82    83    9     93
               | 1     103   123   127   132   29    30    85
                 86
 -1             + 12    142   16    167   219   27    39    41
                 48    62
               | 10    100   125   19    213   35    57    98
               |S 101   110   13    173   44    49    55    84
                 95
               | 151   156   17    178   60
               | 165   172   4
               | 193   70
               | 37    61    63    7     75
 -2             + 158   170
               | 102   15    210   78
               |
               | 190   89
               |T 171   53    66
               | 79
               | 111
 -3             +
               | 155
               |
               |
               |
               |
 -4             +
```

Table II shows the mean and standard deviation of the Medical Surgical Nursing Achievement Examination when categorized into CTT and IRT. The CTT item difficulty is lower than the IRT difficulty index. Thus it implies that Rasch provides the lowest possible index.

TABLE II
DIFFERENCE OF CTT AND IRT ON ITEM DIFFICULTY

	Mean	SD
CTT difficulty	.282	.122
IRT difficulty	0.137854	.1209

Table III shows the mean and standard deviations of the Medical Surgical Nursing Achievement Examination when categorized into CTT and IRT. The CTT item discrimination value indicates very good items between high and low scores. IRT discriminates high and low scores into very good items.

TABLE III
DIFFERENCE OF CTT AND IRT DISCRIMINATION

	Mean	SD
CTT discrimination	0.564729	0.245878
IRT discrimination	1.023836	0.310584

IV. DISCUSSION

The utilization of two frameworks in testing intensified and strengthened the stability of the achievement examination. Though it has been known that" old rules 'were lifted to CTT and "new rules "to IRT [2]. One of which is the issue on reliability wherein CTT's assumption proposes longer test for more reliable test. IRT considered shorter test to be more reliable. Based on the light of findings, it has been established that both CTT and IRT yielded reliable results: .7546 and .97 respectively. Nevertheless it enhances the stability of items for the achievement examination. This will strengthen the consistency of this new achievement examination since the results of CTT and IRT has resemblance in term of reliability.

IRT has person reliability which is one of the limitations of CCT. Person reliability shows the consistency across participant's score. The result also may help to categorize abilities depending on the computed distinct ability strata [8]. The findings generated three distinct ability strata. Thereby IRT extends beyond what is known. Test items can be categorized into eight subgroups instead of 12 subtopics. This can be a springboard to revisit the subtopics of the said examination for improved ones.

The results of CTT and IRT are almost indistinct. Majority of the difficult items identified in CTT are synonymous with IRT results. This will balance the composition of items in the

improved version. This is similar in this study [1] wherein this can be the first phase for item writing. However, the IRT can predict the probability of each student to answer such item correctly or incorrectly based on the logit. Hence, it provides individual assessment instead of group.

ACKNOWLEDGMENT

R.C.Ramos thanks Naressia Ballena, Dean of Adamson College of Nursing for the support and assistance during data collection and Dr. Carlo Magno of De La Salle University Manila for his guidance during the course work.

REFERENCES

[1] B. A. Bhakta , " Using item response theory to explore the psychometric properties of extended matching questions examination in undegraduate medical education"*Biomed Central,vol,*pp. 1-13, March 2005.
[2] S. E. Embretsson& S.P. Reise , "The New Rules of Measurement " (Book style), " in *Item Response Theory for Psychologist, "* Mahwey, New Jersey: Lawrence Erlbaum Associates, Publishers, 2000, pp. 13-17.
[3] K.H. Frith,J.P. Sewell &D.J.Clark "Best Practices in NCLEX- RN Readiness Preparation for Bacclaureate Student Success, "*Computers, Informatics , Nursing,* vol.23, pp. 322 -329.September- October 2005.
[4] R.Hernandez , "Comparison of the Item Discrimination and Item Difficulty of the Quick- Mental Aptitude Test using CTT and IRT, " *The International Journal of Educational and Psychological Assessment,* vol. 1, pp. 12-18, April 2009.
[5] C.Magno , "Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data, "*The International Journal of Educational and Psychological Assessment* vol. 1, pp. 1-11, April 2009.
[6] S. C.Morrison, C. C. Adamson, A. A. Nibert, S.S. Hsia(2004). HESI Exams: An Overview of Reliability and Validity. *Computers in Nursing,* vol. 22,pp. 220-226, August 2004.
[7] A. T., Nibert, " A Third Study on Predicting NCLEX Success With the HESI Exit Exam, "*Computers in Nursing,vol. 19,* pp. 172-178, July 2001.
[8] J.Pomeranz , " Rasch Analysis as a Technique to Examine the Psychometric Properties of a Career Ability Placement Survey Substest, "*Rehabilitation Counselling Bulletin* ,vol. , pp. 251-259, July 2008.
[9] J.L. Silvestre- Tipay, " Item Response Theory and Classical Test Theory: An Empirical Comparison of Item/ Person Statistics in Biological Science Test, "*The International Journal of Educational and Psychological Assessment,* vol. 1, pp. 19-31, April 2009.
[10] Nursing Licensure Exam Results November (Periodicaltype) NLE. Civil Service and PRC Board Examintations, November 2009.
[11] J. M. Linacre , " User's Guide to Winsteps Rasch Computer Program 3.69.0 (Program Manual) , 2009, pp. 1- 485.