

**Jones, M. H. (1988, August). The effects of a small examinee sample size on the precision of measurement for tests developed by four different item-selection strategies. Dissertation. Florida State University, Tallahassee, Florida.**

## Abstract

This study sought to determine the best item selection procedure to use in mastery testing situations where small examinee samples exist. The item selection procedures evaluated were (a) modified classical, utilizing p-values and point biserial correlations; (b) criterion referenced, utilizing phi coefficients; (c) domain sampling, utilizing random item selection; and (d) Rasch model, utilizing item logits (b-values).

A computer program was used to create a data matrix containing the simulated responses of 1000 examinees to 240 items. From this matrix 12 sets of data were drawn. Each data set contained the responses of 50 randomly selected examinees and their associated responses to all 240 items. Three data sets were used for computing test results and or item statistics for each item selection procedure.

Item selection procedures were evaluated in terms of test information, standard error of the estimate, misclassification rate, and accuracy of the domain percentage correct score. Test information associated with each item was computed according to a three parameter

item response theory (IRT) model.

The results show that the classical and criterion referenced procedures were effective at selecting items which, for a given cut-off score, would be identified by a three parameter model as having high information. However, since ability scores for a three parameter model can not be generated from sample sizes of 50 there is no way to effectively utilize the items identified with high information. Indeed, all of the item selection procedures (Rasch model included), which used statistical indices as a bases for item selection, produced biased estimates of the domain percentage correct score. As a result of the biased domain score estimates all of the "optimal" item selection procedures produced generally higher misclassification rates and less accurate domain score estimates than the random item selection procedure. For this reason the random item selection procedure was recommended for mastery tests in which small examinee samples exist.

## Table of Contents

Chapter 1.	
Introduction	
Introduction .....	1
Purpose of Study.....	9
Chapter 2	
Review of the Literature .....	11
Mastery Testing .....	11
Reliability of Mastery Decisions .....	12
Test Size and Classification Accuracy .....	15
Item and Test Information Curves .....	16
Using Traditional Item Statistics to Focus Measurement Information .....	20
Strengths and Weaknesses of Conventional and IRT Optimal Item Selection Strategies .....	27
Interpretation of Domain Score Estimates .....	30
Studies Comparing IRT to Other Test Development Procedures .....	38
Chapter 3	
Methodology .....	48
Introduction .....	48
Definitions of Item Selection Techniques .....	50
Establishing Item Selection Criteria For The Modified Classical Technique .....	51
Data Generating Program .....	53
Program For Selecting Item and Subject Samples .....	54
Exam Item Pool .....	54
Test Length and Subject Sample Size .....	55
Test Item Selection .....	57
Dependent Measures .....	60
Procedures For Judging The Results .....	63
Chapter 4	
Results .....	66
Test Information and Standard Errors of Estimates .....	66
Misclassification Rates .....	74
Accuracy of Domain Score Estimates .....	88
Chapter 5	
Discussion .....	93
Overview.....	93

Measurement Precision For Small Sample Conditions .....	94
Measurement Precision For Large Sample Conditions.....	97
Accuracy of Domain Score Estimates and Classification Accuracy .....	98
Should Traditional Item Selection Procedures be Used to Simulate IRT Item Selection Procedures .....	98
Maximizing The Percentage Correct Domain Score Accuracy.....	102
Chapter 6	
Conclusions and Suggestions for Future Research .....	103
References .....	107
Appendix 1 .....	111
Appendix 2 .....	117

## Chapter I

### Introduction

The purpose of this study was to compare various item selection methods in terms of their effectiveness in maximizing test information at a cutoff score under conditions involving small examinee sample sizes (i.e.,  $N = 50$ ).

In recent years there has been a proliferation of mastery testing programs in education, professional licensure, and personnel selection. In each of these areas, the mastery/nonmastery classifications derived from test scores have serious impact on the lives of examinees. For this reason it is critical that classifications be as accurate as possible.

The primary means of achieving accurate classifications is to insure that the standard error of estimate (SEE) around the cutoff point is as low as possible. This is accomplished by selecting test questions which concentrate measurement information around the cutoff score of interest (Novick, 1968; Birnbaum, 1968). There are many item selection methods discussed in the literature that can be used to focus measurement information. Most of the techniques described are based

on a statistical index that assists the user in identifying items which are optimal for the purpose of providing measurement information. Further, each statistical index is associated with one of the four major test development approaches and/or theoretical measurement models commonly used today. These theoretical models and/or test development approaches are; item response theory, classical, criterion referenced, and domain referenced. One item selection technique associated with each of these test development approaches and/or theoretical models is investigated in the present study.

The objective of reducing the SEE around a cutoff can be easily achieved if an item response theory (IRT) approach is used for test development and scoring. In the simplest IRT model, a one parameter model, items are selected which have difficulty values close to the ability level (theta level) of the cutoff score. To maximize measurement information at a cutoff score, a test should be composed of test items that have difficulty values that are identical to the ability level of the cutoff score. However, it is unrealistic to generate item pools sufficiently large to enable the composition of tests made up entirely of items with the same difficulty value. Therefore, one typically selects items that are as close as possible to the cutoff point.

Item response models are well adapted to the task of

item selection at the cutoff because they have, unlike classical models, item statistics that are reported on the same scale as examinee abilities. The disadvantages of an IRT approach are that as the test length and sample size decreases so does the precision of item parameter estimates (see, e.g. Hambleton & Cook, 1982). However, parameter estimates derived from small samples might provide useful information for test construction purposes even though they are not accurate enough for score generating purposes.

The problem of adequate sample size presents a rather serious obstacle for many test development applications because, while there is typically an adequate domain of content from which to develop a large pool of items, there is not always a large pool of examinees from which to establish accurate item parameter estimates. This is the case in many of the low examinee incidence areas in education, professional licensure, and personnel testing. In these low incidence areas the number of examinees tested per year may be less than 100.

Lord (1983) has shown that the one parameter Rasch model (Rasch, 1960) is slightly superior to other IRT models for the purpose of estimating examinees' true score for sample sizes between 100 to 200. However, Lord offers no recommendations about what IRT model to use, if any, when examinee samples are below 100. The literature is

silent with regard to whether the Rasch model may still be useful for purposes of optimal item selection even though ability estimates might be unacceptable.

A second model for use in identifying optimal items for inclusion in tests is the classical model, which includes item statistics such as item difficulty,  $p$ , and item discrimination,  $r$ . Unfortunately, with classical statistics there is not a connection between the scales for persons and items. Hambleton and de Gruijter (1983) argue that for this reason classical item statistics are "inappropriate" for selecting optimal items from a given pool. However, it is believed that the term inappropriate used by these authors, may be too strong, and that "non optimal" might be a better descriptor. That is, if a test developer does not have access to IRT computer programs then the use of classical item statistics would not be entirely inappropriate. Under most circumstances a classical approach will allow for more optimal item selection, for the purpose of focusing information around the cutoff, than simply randomly selecting items. Indeed, Hambleton and de Gruijter stated later that: "classical item discrimination values have some usefulness in item selection. In general, items with high item test score correlations (discriminations) were more useful than items with low correlation values" (p.356).

*Non optimal*

A third strategy for optimal item selection would be

to use one of the discrimination indices associated with criterion reference testing (CRT). Shannon and Cliver (1987) review four criterion-referenced item selection procedures: (a) phi- the phi correlation between examinees' item and dichotomized test performance outcomes. (b) B-index- the difference between item difficulties (i.e., proportion correct) of examinees passing and examinees failing the test. (c) phi/phi max- a modification of the phi index proposed by Cureton (1959) as a solution to the restriction in range of phi resulting from differences in marginal totals. This value is the phi coefficient divided by the maximum phi coefficient. (d) agreement statistic- the probability of agreement between outcomes on a given item and outcomes on a test as a whole. In this study the authors used item information functions derived through item response theory formulas as the standard by which each CRT procedure was evaluated.

Item information functions (IIFs) were first proposed by Birnbaum (1968) and can be interpreted as the amount of measurement information provided by an item at a specific ability (theta) level. The item information functions are particularly useful for measuring optimal item selection strategies because they offer measures of item discrimination power at cut-off scores. Hambleton and de Gruijter (1983) consider IIFs superior to conventional indices for this reason.

The results of the study by Shannon and Cliver (1987) showed that some of the CRT statistics investigated produced high correlations with the item information function (IIF) used in item response theory. These results indicate that CRT discrimination indices may be reasonable substitutes for IIF's when circumstances do not allow the use of IRT procedures. Of the conventional statistics compared, the phi coefficient produced the highest median rank correlation with IIFs. This result suggests that the phi coefficient may be the CRT discrimination index of choice for optimal item selection.

Several studies, Hambleton and Cook (1979) and Hambleton and Arrasmith (1987), investigated the amount of measurement precision that can be gained through an IRT based optimal item selection strategy relative to a classically based approach. A weakness of these studies was the failure to use tests of realistic size (i.e., 50 items or more) for purposes of comparing the various optimal item selection procedures. Instead, studies of relatively short length, 20 to 30 items were used. It is reasonable to assume that most applied testing situations do not require limiting test sizes to 20 or 30 items. At present there are not any studies comparing optimal item selection strategies in which realistic test sizes were used.

Another factor that has not been studied, which

affects the measurement accuracy of optimal item selection strategies, is examinee sample size. This is a particularly important area of investigation because it is known that the stability of IRT item parameter estimates decreases as the subject sample size decreases.

Therefore, it is also important to determine how an IRT item selection strategy functions in relation to more conventional item selection techniques (e.g., classical and criterion referenced) in situations where a small (i.e.,  $< 100$ ) sample size exists.

Information gleaned from studies involving realistic test sizes and small examinee samples is needed to complete the research base regarding the relative performance of optimal item selection strategies. Further, the results will clearly be of assistance to test developers who are weighing the advantages and disadvantages of switching from one test development method to another.

A fourth item selection model, currently being used, is the domain sampling model. This model requires that one simply select items in a random or stratified random fashion from the content domain. In this approach items are primarily selected to represent important classes of behavior that make up the domain. Many researchers (Popham and Husek, 1969; Hambleton, Swaminathan, Algina and Coulson, 1978; Nitko, 1984) have stated that in domain

referencing the application of item statistics for selecting an optimal set of items would theoretically weaken the interpretability of the domain score. By selecting this model for item selection the test developer takes an implicit position that accurate domain representation is more important than the measurement accuracy which is gained by optimal item selection. Such a position raises the question of how much measurement accuracy is gained by optimal item selection verses random item selection.

Several studies (de Gruijter & Hambleton, 1983; Hambleton & de Gruijter, 1983; Haladyna & Roid, 1983) investigated this question in terms of test information at a cut-off and misclassification rates for tests of various sizes (i.e., 8-20 items). Tests for these studies were constructed through both, optimal (IRT) item selection and through random item selection. Although relatively small tests were used, the subject sample sizes used (i.e.,  $N > 100$ ) were large enough to achieve relatively stable Rasch parameter estimates. These studies did not address the question of how the misclassification rate, domain score accuracy, SEE and test information of a random item selection strategy compares to an optimal IRT strategy under circumstances where examinee samples are smaller than 100. Because domain sampling models are frequently used in settings where examinee sample sizes are less than

100 this research question needs to be addressed.

Of the studies reviewed thus far, in which comparisons were made of the misclassification rates for various item selection procedures, the domain score estimates for the IRT procedures were typically derived using the test characteristic curve method. This method described by Hambleton and Swaminathan (1984) produces a domain score estimate that should appropriately be interpreted relative to the latent ability domain being measured by the IRT procedure.

This IRT domain score estimate can be contrasted with another version of the domain score associated with the domain sampling model. In this model, the domain score estimate, taken from the observed score, is an unbiased estimate of the domain score when the observed score is based on a random sample of items from the domain. Since none of the studies reviewed utilized the latter definition of a domain score, little is known with regard to how IRT procedures compare with other item selection procedures in terms of the accuracy of observed scores (i.e. domain score estimates) to domain scores.

#### Purpose of Study

The present research study focuses on settings involving low examinee incidence (i.e.,  $N$ 's < 100) similar to those encountered in teacher content area certification, professional licensure, and personnel

testing. The test size and the size of the item pool used for the study are: (a) a test of reasonable size (e.g., 50 items), (b) an item pool of 240 items.

The following questions were addressed:

1. What range of p-values and point biserial correlations should be used for selecting items to focus information at a specified cut-off score with a given underlying ability distribution?
2. Given a small examinee sample size ( $n = 50$ ), what are the differences in the SEE, test information functions, accuracy of domain score estimates and classification accuracy for test development strategies such as classical, criterion referenced, domain sampling, and the Rasch model?
3. How do the results found in number 2 (above) compare with results derived from tests in which the item statistics, used to select items, were derived through the use of large examinee samples.
4. How does focusing test information through optimal item selection affect the misclassification rate and accuracy of domain percentage correct score estimates?

## Chapter II

### Review of the Literature

This study results from the need to know more about how to increase measurement information in mastery test using conventional item statistics. Additionally, there is a need to know which item selection procedures, IRT, classical, CRT, or domain sampling, are best to use in a situation where small samples exist. This study begins with a review of the literature dealing with the following topics: mastery testing, reliability of classification decisions, test size and classification accuracy, using conventional statistics to focus measurement information, using criterion referenced discrimination indices to focus measurement information, item and test information curves, strengths and weaknesses of conventional and IRT optimal item selection strategies, and interpretation of domain score estimates. This was followed by a more detailed review of studies comparing IRT to other test development procedures.

#### Mastery Testing

As Berk (1983) points out it is not uncommon to find terms like domain-referenced test, objectives-referenced

test, competency-based test, proficiency test, mastery test, and criterion-referenced test used interchangeably in the literature. Because there may be some question as to the meaning of the term mastery testing the definition used for the purposes of this study is as follows:

"Mastery testing is a subtype of criterion-referenced testing in which only one score is important, a cut-off or mastery score above which one is regarded as successful and below which he is regarded as failing. The other score levels don't tell us what the person can do except that he is not above the mastery score. We will call test of this type mastery tests to keep them separate from the more informative but much more difficult to create criterion-referenced tests" (Hills, 1981, p. 97).

The general term "criterion-referenced testing" which encompasses mastery testing was defined as: " a test deliberately constructed to yield measurements, that are directly interpretable in terms of specified performance standards" (Glaser & Nitko, 1971, p. 653).

#### Reliability of Mastery Decisions

There are more than a dozen different statistics for measuring reliability of criterion referenced tests. Each falls into one of three categories of reliability, (a) threshold loss, (b) squared-error loss, or (c) domain score estimation. Each index has its own associated advantages and disadvantages and appropriate applications.

The threshold loss function assumes that a dichotomous qualitative classification of mastery or nonmastery of an objective based on a cutoff score exists. It usually assumes the losses associated with all misclassifications (false mastery and false nonmastery) are equally serious regardless of size. An example of this type of function is the  $P^0$  index (Hambleton & Novick, 1973). This index reflects the percentage of the examinees taking a test who were correctly classified (i.e., pass or fail).

The squared error loss approach is based on the squared deviations of individual scores from the cutoff score. The squared deviations reflect the degree of mastery or nonmastery along the score continuum. In contrast to the threshold loss approach, this approach deals with the consistency of measurement, and assumes that losses associated with false mastery and nonmastery decisions are not equally serious. The statistics associated with this approach are most useful in instructional situations where the teacher is concerned with the degree of mastery-nonmastery for each student along a score continuum. In licensure and certification tests typically there is not a concern with degrees of mastery-nonmastery, however, the statistic does offer another perspective on the score precision.

The domain score estimation statistics deal with

estimating the proportion of items within a domain which are known independent of any cut off score or mastery standard. Domain score adequacy statistics can be broken down into two categories, individual specific and group specific. The individual specific statistics are estimates of standard error that can be used to form confidence intervals around each individuals observed score. The group specific statistics are averages of the individual specific statistics calculated over all individuals tested.

Since true domain scores are rarely available, one of the procedures mentioned above is typically needed for making estimates of the accuracy of observed scores. However, progress has been made in specifying content domains to the extent that some finite item domains are specifiable (Roid & Haladyna, 1982). In this regard, researchers can simulate large but finite study domains through the use of computers. From a finite domain, smaller sub-samples of simulated item responses can be drawn to produce domain score estimates. By using this technique the scores on the total domain (i.e., domain scores) can be calculated as well as differences between population and sample domain scores. For example, the misclassification can be obtained by simply summing the instances where examinees' classification (pass/fail), based on their observed score, differs from their true

classification based on their domain score. Similarly the accuracy of the observed scores (expressed as percentage correct) can be evaluated by summing the absolute deviations of observed scores from the domain score, and then taking an average.

#### Test Size and Classification Accuracy

When using a mastery test the matter of determining a test length is directly related to the number of classification errors that are tolerable to test users. Obviously very low probabilities of misclassification can be achieved if the size of the test is very large. However, most applied mastery testing situations involve the reality of specified time parameters which set the upper limit on test size. Practical experience suggests that the lengths of most of the tests in teacher certification, personnel selection, and occupational licensing, fall within the range of 50 to 200 items. Very few tests within these realms will contain less than 50 items simply because of the difficulty of generating acceptable test reliabilities with fewer than 50 items. However, all of the studies (e.g., Hambleton & Cook, 1979; Hambleton & de Gruijter, 1983) investigating optimal item selection methods for mastery testing have utilized tests of less than 30 items. It is believed that future studies should utilize test sizes that are more realistic in terms of applied testing situations.

Item and Test Information Curves

Birnbaum (1968) defined the notion of information as a quantity inversely proportional to the squared length of the confidence interval around an estimate of an examinee's ability.

Generally, tests vary from one another in terms of where information is focused within the test. In this regard the test information curve within a single test varies with the ability level measured by the test. Because the information varies from one point on the test scale to the next, it has been suggested that test information curves should replace the use of classical reliability estimates and standard errors of measurement in test score interpretation.

Stated mathematically, the test information curve appears as follows:

$$I(0) = \frac{\sum_{g=1}^n P_g'^2}{\sum_{g=1}^n P_g Q_g}$$

In this expression the amount of information at an ability level is expressed as  $I(0)$  and  $P_g$  is the probability of a correct answer to item  $g$  by an examinee with ability level 0;  $Q_g$  is equal to  $1 - P_g$ ; and  $P_g'$  is the

slope of the item characteristic curve at ability level 0.

The quantity which is summed in the equation presented represents the information that item  $g$  contributes to the total information at a given ability level. The plot of the information contributed by an item at all ability levels is referred to as the item information curve. When the information curves are summed and plotted for all test items, the resulting plot is called the test information curve. The item information curves, and the resulting test information curves allow one to determine the accuracy with which each ability level is estimated. Specifically, one can directly measure test items for a given ability level by simply measuring the height of the test information at a given ability level. With the two and three parameter models, the item information curve depends on the slope of the particular item characteristic curve and the conditional variance of test scores at each ability level. However, in the one parameter model the items are considered to be equally discriminating, and therefore the height of the test information curve at a cut-off is dependent on the number of items with difficulty values close to the cut-off. The test information curve and the particular type of test information it provides is particularly useful in the construction of mastery tests where measurement information must be focused. Many excellent discussions

of information curves can be found (Hambleton & Swaminathan, 1985; Lord, 1980; Lord, 1977; Wright, 1977; Birnbaum, 1968).

The procedure for using the test information curve to focus measurement information is fairly simple and involves four basic steps (Lord, 1977):

1. Describe the shape of the desired test information function. Lord (1977) calls this the target information function.
2. Select items with item information functions that will fill up the hard to fill areas under the target information function.
3. After each item is added to the test, calculate the test information function for the selected test items.
4. Continue selecting test items until the test information function approximates the target information function to a satisfactory degree.

It should be noted that items which are optimal (i.e., highest in information) for a given cut-off can be generally identified or estimated without using the item information data. The relationships of item parameters and information have been shown by many researchers (see e.g., Hambleton & Swaminathan, 1986). Although the procedures for item selection will vary for the one, two and three parameter models the basic concepts are the same. In all three models the b-value estimates the point

on the ability scale where an item is maximally discriminating. However, for the two and three parameter models the addition of the a-value provides for an estimate of the amount of discrimination which is focused at the b-value. For example, in a two parameter model, to choose items which are optimal for a given cut-off ability one would select items with b-values at that ability and which have the highest a-values. The c-value which is added with the three parameter model tends to distort the properties of the a-value and the b-value when the c-value rises above zero. However, the item selection procedures would remain basically the same.

A function related to the test information is the standard error of the ability estimation (SEE). This function is equal to  $1/\text{square root of the information}$ . The SEE is the expected standard deviation of errors of estimated ability. For example, if one were to give a test to a group of examinees with identical 0's, and use the test to estimate their 0's, the standard deviation of those estimates would be the SEE.

Since  $I(\theta)$  varies along the  $\theta$  scale, so will the SEE. As the information increases, the SEE decreases. This concept of SEE in IRT (Hambleton & Swaminathan, 1983) is a more viable alternative to the classical function:

$$\sigma_e = \sigma_x [P_{xx} (1-P_{xx}')]^{1/2}$$

This function represents the standard errors averaged over the ability levels. Samejima (1977), concludes that the act of averaging errors and assuming the independence of true and error scores is unreasonable, and that it's coefficient, and the classical standard error of measurement are unpalatable.

Using Traditional Item Statistics to Focus Measurement Information

Richardson (1936) showed that if one wants to differentiate examinees below a given ability level from those above it, without making distinctions among examinees in the two groups, all the items in the test should be of a difficulty level such that they are marked correctly by half the examinees at the ability level of interest. In other words if a test developer wants to build an exam to discriminate between examinees capable of passing items of thirty percent difficulty he/she would build a test composed of items which have a thirty percent difficulty level. Since it will be unlikely to have enough items of precisely the desired level, the test developer will have to use some items above and below the desired difficulty level. Other authors (e.g., Davis, 1961; Henrysson, 1971) have also discussed the subject of item selection to focus measurement information. The

basic procedure offered by these sources is to use p-values to select items that focus information at the area of interest. Then, of the items falling in the area of interest selecting those with the best discrimination (e.g., high point biserial correlations).

To date there have not been any studies investigating what specific ranges of p-values and point biserials to use in order to maximize the measurement information at a given cut score. Hambleton and de Gruijter (1983) point out that there is not a relationship between the underlying scale of the domain scores and the p-values. They conclude that, for this reason, classical statistics are inappropriate as an optimal item selection method. However in two other studies, Hambleton and Cook (1979) and Hambleton and Arrasmith (1987), the authors used a specified range of p-values and point biserial correlations to select test items. Although it was not discussed by the authors the apparent intent of using a particular range of p-values was to focus test information at a particular cut-off point. Intuition would suggest that there is some relationship between the p-value and the domain ability score albeit not a mathematically direct relationship.

In the next section of this study a mathematical relationship will be shown between another conventional item statistic, the phi coefficient and the IRT test

information function at the cut-off score. Although a similar relationship with test information can not be established for the p-value and point biserial correlation, these conventional statistics are related to other IRT statistical concepts.

Schmidt (1977) shows that the b parameter is related to the p-value in the following way:

$$b = \frac{yz(1-c) \text{ KR-20}}{d \text{ pq}}$$

where

d = d-value, the point biserial item-test correlation

p = p-value, the proportion of examinees correctly answering the item

$$q = 1-p$$

K.R. 20 = Kuder-Richardson formula 20 reliability

y = the height of the N(0,1) curve at the z score that cuts p' proportion of the area under the N(0,1,) frequency function

z = the z-score that cuts off p' proportion in the upper portion of the area under the N(0,1) frequency function

c = the c-value (item pseudo chance level)

$$p' = \frac{p - c}{1 - c}$$

Schmidt shows that the a parameter is related to the point biserial correlation in the following complex way:

$$a = \frac{d \cdot pq}{(KR-20)(1-c)^2 y^2 - d^2 pq}$$

These formulas demonstrate that there is a mathematical relationship between conventional statistical measures of difficulty (p-value) and discrimination (point biserial correlation) and the IRT corollaries of difficulty (b-value) and discrimination (a-value). Given these mathematical relationships it is believed that an empirical relationship (e.g., correlational) between item information and these statistical measures can be shown. If this assumption is true then the p-value and point biserial could be used in a manner similar to the b-value and the a-value in identifying items that will focus information at a given cut-off score.

The theoretical work by Richardson (1936) clearly shows that measurement information can be focused through the use of classical statistics to make item selections. The question that remains is whether there are ways to maximize the effectiveness of these item statistics for the purpose of focusing information at a cut score.

Because there has not been any research into ways to maximize the effectiveness of using p-values and point biserial values for the purpose of maximizing test information it is believed that research regarding this question is warranted.

While the relation of p-values and point biserials to the item information function is at present unknown, there are other traditional item statistics in which an empirical or mathematical relationship has been shown (Harris & Subkoviak, 1986; Van der Linden, 1981). A study in 1987 by Shannon and Cliver evaluated four conventional item discrimination indices, phi, B index, phi over phi max, and the agreement statistic, using item information functions (IIF) as criterion of effectiveness. Of the indices reviewed the one with the highest rank correlation, with the IIF's (median  $r=.96$ ) was the phi coefficient. The authors postulated this finding based on the fact that the phi coefficient is approximately proportional to the IIF. This approximate relationship was explained by first relating IIF to the B index and the

phi coefficient.

$$\text{IIF approximately } \frac{B^2}{P_i Q_i} \text{ and}$$

Where  $P_i$  = the proportion of examinees who answer item  $i$  correctly and  $Q_i = 1 - P_i$ .

$$\text{IIF approximately } 0$$

In this context  $B$  represents a proportional relation.

This was accomplished by delineating the covariance relations between the binary item and test scores expressed as,  $u_i$  and  $u_T$  :  $\text{Cov}(u_i, u_T) = P_{iT} - P_i P_T$ . The variance of  $u_T = P_T Q_T$ , so  $B$  can be expressed as:

$$B = \frac{\text{cov}(u_i, u_T)}{\text{var}(u_T)}$$

This equation represents the slope of the regression line predicting  $u_i$  from the binary ability measure  $u_T$ .

The numerator of the information function in the equation below is defined by Birnbaum (1968) as the squared derivative of the item response function with respect to ability.

$$I[O, u_i] = \frac{[P'_i(O)]^2}{P_i(O) [1 - P_i(O)]}$$

Since  $P_i(O)$  is the regression of the item score on the continuous ability measure  $O$ , the numerator in the equation above is the squared slope of the item ability regression. Therefore the squared slope for the information can be replaced with  $B^2$  and  $P_i$  can be substituted as an estimate of  $P_i(O)$ .

The IIF can then be related to phi because phi and  $B$  are related:

$$\frac{B^2}{P_i Q_i} = \frac{O^2}{P_T Q_T}$$

Then because  $P_T Q_T$  is constant for all items in a test:

$$\frac{B^2}{P_i Q_i} = O^2$$

therefore: the IIF approximately  $\sigma^2$ .

The mathematically direct and empirically strong relationship between the phi coefficient and the IIF suggest that the phi coefficient would perform well for the purpose of selecting optimal test items. However, at present there have not been any studies confirming that tests developed through the use of the phi coefficient are comparable in measurement precision to those developed through IRT procedures. Specifically, there have not been any studies that compared the classification accuracy, standard error of the estimate and test information of tests produced through the use of IRT and phi coefficients.

#### Strengths and Weaknesses of Conventional and IRT Optimal Item Selection Strategies

The concepts of strong and weak are relative and their use must be accompanied by a reference point from which a comparison can be made. In the case of IRT the reference point for comparison is the standard testing technology, sometimes referred to as classical theory, which dominated the applied world of testing through the 1970's and perhaps still dominates the 1980's. Briefly, the weaknesses of conventional item statistics commonly cited are, group dependent item statistics, test dependent

ability estimates, and a single statistic for representing measurement error existing in a test (Lord & Novick, 1968; Marco, 1977; Hambleton, Swaminathan, Cook, Eignor & Gifford, 1979). The use of an IRT model allows the user to avoid these pitfalls by generating item and sample free statistics and measures of the precision of ability estimation at different ability levels.

IRT is especially useful for this task because the contribution of each item to the test information function can be determined independently of the other test items. With classical testing technology, the exact contribution of any item to test reliability or the error of measurement can not be determined independently of all the other items in the test. However, the concept of focusing test information at a cut off point is not alien to the classical measurement philosophy. Indeed, the use of p-values and discrimination index were being used to focus test information long before IRT was popular (Richardson, 1936).

The benefits of using an IRT approach do not come without some disadvantages. The four most commonly cited disadvantages according to Hambleton and Swaminathan (1985) are; meeting dimensionality assumptions, identifying the model that best fits the data, meeting the needs for large samples, using complicated computer programs, securing highly trained technical staff to

interpret the results, and explaining results to lay individuals.

The disadvantages associated with an IRT approach to optimal item selection in general do not apply to the statistical procedures utilized in the classical and criterion referenced approaches. Indeed, it is the disadvantages associated with the IRT approach that makes the more conventional statistical approaches appealing. The following are four advantages that conventional (i.e., classical and criterion referenced) approaches offer which are conceptually parallel to the four disadvantages associated with the IRT approach. (a) The conventional approaches do not require complex statistical analysis to prove the data is unidimensional. (b) The conventional model that is chosen is determined by the purposes of the test rather than the nature of the candidate response data. For example, from an IRT perspective if candidate response data indicates that extensive guessing is occurring then a three parameter model is warranted. (c) Small sample sizes do not present as serious a problem for conventional approaches. The point biserial and phi coefficient could be calculated for three bivariate pairs of numbers. In contrast, it is recommended (Lord, 1980) that at least 1000 subjects and 30 items be used for the IRT computer program LOGIST for calculating item and ability parameters. (d) The general public and test developers

have more familiarity with conventional procedures for computing item statistics and total scores.

#### Interpretation of Domain Score Estimates

In 1969 Popham and Husek published an article titled "Implications of Criterion Referenced Measurement". In this article they discussed the distinctions between norm referenced and criterion referenced approaches to testing. The authors point out that one of the central differences is the use of score variability. For the norm referenced procedure, score variability is very important because the test constructor wants to be able to evaluate an examinee's performance in relation to all other examinees. In this case the amount and the type of discriminations an item makes becomes important. Therefore, certain statistical indices are used (e.g. point biserial correlations) in order to evaluate the discrimination power each item exhibits.

In the case of criterion referenced measurement, test are typically constructed by selecting a subset of items from a large pool of items that represent a domain of task or instructional material. Each items' importance to the measurement accuracy of the test is determined by the importance of the instructional material or task it represents and not the amount and the type of discriminations it produces. Therefore the use of item discrimination indices, to increase score variability, may

reduce the interpretability of the test scores. This would occur if the items chosen through item analysis were not representative of the larger item group (domain). This position is supported by other researchers, (e.g. Hambleton; Swaminathan; Algina & Coulson, 1978; Berk, 1980).

Early in the 1980's researchers (Hambleton & deGrujter, 1983; Haladyna & Roid, 1983) began investigating how IRT procedures might be used to evaluate item performance on criterion referenced tests. They compared several traditional item discrimination indices to item response theory item information indices for the purpose of focusing measurement information at a cut-off score. As a result of their research findings, these researchers proposed the use of IRT measures of item information to improve measurement accuracy.

The question that arises is how can IRT procedures, and more specifically IRT procedures which depend on score variability, be used for selecting test items without destroying the interpretability of the test scores. A review of the literature reveals that apparently this question has yet to be directly addressed and debated. Perhaps the question has been lost in the confusion that has typically surrounded criterion referenced testing and the related ill-defined terminology. It is believed, however, that the reasoning behind the suggested use of

item information to make item selections is based on the definition of a "domain score" as it is normally used in IRT applications to criterion referenced tests.

Specifically, IRT allows one to estimate a proportion correct score (domain score) on a large or infinite domain of items from a small subset of items which have been selected to provide optimal discrimination. Hambleton and Swaminathan, (1984) state the following:

When the test items included in the test area are a representative sample of test items from the domain of items measuring the ability, the associated test characteristic function transforms the ability score estimates into meaningful domain score estimates. A problem arises, however, if a non-representative sample of test items is drawn from a pool of test items measuring an ability of interest. Such a sample may be drawn to, for example, improve decision making accuracy in some region of interest on the ability scale. The test characteristic function derived from such a non-representative sample of test items does provide a way for converting ability score estimates to domain score estimates. While ability estimates do not depend upon the choice of items, the domain score estimates will be biased due to the non-representative selection of test items. However, if the test characteristic function for the total pool

random or stratified random selection of items from a larger pool of items. The score resulting is not completely unambiguous. If an examinee earned a score of 90 percent correct one does not know which items the examinee missed. However, if the items were a representative sample of the larger domain of items, representing the desired knowledge domain, then one does have an unbiased estimate of the percent of the knowledge domain the subject has mastered. It is with this second type of CRT test where an IRT derived domain ability score estimate would appear to diverge from the domain percentage correct score offered by Popham and Husek.

It is difficult to specify just what interpretation a user may give to a domain ability score estimate, derived from a typical licensure and or certification test. The ability dimension which the items define is one which is ethereal and difficult to interpret. If the items are drawn out of a single narrowly defined set of subject matter, for example, anatomy of the foot, then the ability dimension would probably have some relation to the level of the items in terms of some taxonomic schema like that of Gagne or Bloom. However, as the number of subject areas is increased then the meaning of the ability dimension becomes more complex because certain subject areas might be systematically more complex than other areas and by chance produce items with p-values all

centered at a particular ability level. If this occurs it would be difficult to imagine how an ability score could be directly interpretable in terms of any specified performance standards.

Many IRT theorists might argue that the second type of CRT tests would tend to violate unidimensionality assumptions underlying IRT theories, thus contraindicating the use of IRT procedures. However, it is likely that IRT procedures will continue to be used on CRT test of the second type and thus open the way for mis-interpretation of the domain score estimates.

The difference in the percentage correct and ability interpretations that can be given to a domain score has importance for the purposes of this study. Specifically, this study focuses on increasing measurement accuracy at a cut-off point in testing situations where small examinee samples exist (i.e.  $n = 50$ ). The small samples used eliminate from consideration the use of IRT estimates of domain abilities. Therefore, the domain score estimates must necessarily be derived from percentage correct observed scores. In this way information will be gleaned regarding the effect that focusing measurement information at a cut-off point has on the accuracy of a domain score estimate calculated from the percentage correct observed score. As previously mentioned, this can be accomplished because with simulated data a large subject by item poll

can be created and smaller samples of subjects and items can be drawn for estimating the known population parameters. The accuracy of the domain score estimate can be evaluated both in terms of classification accuracy (pass/fail), and absolute deviation from the domain score (i.e. score on the item pool).

The information gained from this analysis will be especially important in light of the study by Shannon and Cliver (1987). In their study the authors recommend the use of the phi coefficient to identify items that would have a high correlation with item information (from a three parameter IRT perspective) when IRT procedures are not feasible (i.e. small sample sizes). The authors recommend (personal conversation May 6, 1988) using a number correct score as the estimate of the domain score. Since a test characteristic curve would not be available to estimate domain ability score the domain score would have to be reported in terms of the percentage of the items in the domain which could be answered correctly. Once again the importance of the present study will be in determining what effect focusing measurement accuracy has on the classification accuracy of domain score estimates when the estimates are interpreted as an estimate of the percentage of items examinees would answer correctly if they were to given all items in the item pool.

Studies Comparing IRT to Other Test Development Procedures

Studies contrasting the measurement accuracy of an IRT item selection strategy to classical, domain referenced and other item selection strategies are limited. In 1979 Hambleton and Cook used simulated data for 200 items and 200 subjects to compare five item selection techniques. Tests were compared in terms of score information at five ability levels. The item selection strategies selected were (a) Random- items selected totally at random; (b) Standard- items with difficulties between .30 and .70 were selected. Of the items within this range, only items with the highest item discriminations parameters were chosen; (c) Middle difficulty- the thirty test items that provided the maximum amount of information at an ability level of 0.0 were selected from the pool; (d) Up and down- this method involved a three step process. First an item was selected that provided the maximum amount of information at an ability level of -1.0. Then an item was selected that provided the maximum information at 0.0. level. The third step was to select an item at +1.0 and then go to step one. This was repeated until thirty items were selected. (e) Maximum Information- involved averaging the information provided by each of the items in the pool across three ability levels 1.0, 0.0, and 1.0. The items

with the highest average across the three ability levels were selected.

In the results of this study, the random method, not surprisingly, produced the smallest amount of information at the ability levels of interest. The standard/classical method provided maximum information for abilities at the center of ability distribution (i.e., .0). A reflection of the roughly normal distributional shape of the item pool. Interestingly this approach provided as much information as the maximum information at the center of the distribution and at the upper levels of the ability levels of interest. In fact, the only method that presented information that surpassed the standard method for ability level 0.0 was the middle difficulty method. All procedures surpassed the classical method at levels below -1.0. The middle difficulty method in addition to providing the greatest amount of information at the cutoff also provided appreciable amounts of information at the two adjacent ability levels. The up and down method provided the least amount of information at 0.0, with the exception of the random method. However, this method surpassed all other methods at ability levels of -1.0 and +1.0. The "maximum information" method provided information at almost the same level as the up and down method at the levels of -1 and +1 but was equal to the classical and surpassed by the middle difficulty method

for the level of 0.0.

In summary, the random method, which would most closely correspond to the domain referenced approach, fared poorly at the theta levels studied. The classical approach did surprisingly well, equal to or better than the two IRT based approaches at the center level of 0.0 and was surpassed only by the middle difficulty procedure. The difference at the level of 0.0 was five test information points with a value of 40 for the middle approach and 35 for the classical approach.

In light of this study comparing alternative item selection strategies, there are a few questions a researcher might ask which are relevant to the purposes of this study. They are:

1. Given the fact that at the center ability level the classical approach was only surpassed in information by one IRT based approach (i.e., middle difficulty), what would the difference in information have been had the test lengths been set at 50 or 60 items?

2. Assuming that the test lengths remained at 30 items, what would happen to the differences between the IRT based procedures and the classically based procedures if the parameter estimates had been based on sample sizes of less than 100?

In 1983 Hambleton and de Gruijter conducted a similar study with three primary objectives. First, to consider

the inappropriateness of classical item statistics in criterion referenced test item selection. Second, to clarify the IRT item selection procedure. Finally, to offer two examples that highlight the advantages of an IRT method.

The authors point out that the primary disadvantage to a classical approach is that the item difficulty index,  $p$ -value, is on a different scale from domain scores. To illustrate, the authors provide a simple example using three items at three different difficulty levels 1.0, .80, .60 and five groups of twenty subjects each. In the example the  $p$ -value at the correct cut-off level does not produce accurate domain score estimates. The example is effective in demonstrating that the domain score estimate and the  $p$ -values are not on the same scale. What the example fails to show with regard to classical statistics is that there is a relationship between IRT difficulty index ( $b$ -parameter) and classical difficulty index ( $p$ -value). Further, it does not show that there is a relationship between the IRT discrimination index ( $a$ -parameter) and the classical discrimination index (point biserial). Finally the authors fail to point out that although there is not an exact relationship between the domain score scale and the item difficulty scale the relationship is such that classical item statistics can be useful in selecting optimal items. In fact, the results

of a 1979 study by Hambleton and Cook, previously cited, show this quite clearly. The Hambleton and de Gruijter (1983) article does not mention the results of the 1979 study. In fact the authors did not generate probabilities of misclassifications for optimal item selection data for a classical approach. Instead the authors only compared a random item selection strategy with an IRT optimal item selection strategy using simulated data. As would be expected the results showed the one parameter IRT approach to be superior for tests of various sizes (i.e., 8-20 items) and at several different cut-off scores (i.e., .65, .75, and .80) when item pools are homogeneous with regard to discrimination. The conclusion one can derive from this study is that an IRT strategy is ideal when it is important to produce the shortest test possible which still meets certain preset criteria regarding misclassifications as derived from an ability score scale.

A similar study was conducted by the same authors in the same year (1983). The study differed from the previous study in that simulated data was used and therefore the true probabilities of misclassification could be computed. This study differed also in that item parameters were estimated, thus making the study more realistic. The results were generally congruent with those of the previous study.

A third study was conducted in 1983 by Haladyna and

Roid, which was very similar to the other two studies reviewed. In this study the authors used a one parameter model (Rasch Model) to focus information at a cut-off score for a criterion referenced test on dental health. Once again a random sampling model was used for comparison purposes. This study differed from the other studies cited in that an additional measure of measurement accuracy was calculated for the domain score estimates derived. This measure was the average absolute deviation (AAD) of the domain score estimate from the domain score. The domain score for this study was defined as a large pool of items created for the study. The study utilized a ratio of the AAD to the SD of the deviations in order to compare the one parameter model to the random sampling model. The authors felt that through this procedure the relative accuracy of the domain score estimates from the Rasch and the random procedures could be judged. However, the present author believes that this type of comparison is inappropriate because of the conceptual differences that exist between a domain ability score and a domain percentage correct score.

In 1986, Hambleton and Arrasmith, carried the research in this area one step further by comparing several test development strategies based on items taken from a real test. Using procedures similar to some of the previous studies in this area, the researchers defined a

finite domain of items (item pool) and then developed a smaller test by selecting items from the pool. As with the previous studies the random method and the classical approach were compared to several IRT based strategies. However, the IRT approaches included a new approach called "content optimal."

The content optimal approach was created by the authors to deal with the problem of content validity which may arise when items are selected based upon statistical, rather than content consideration. For this procedure items were selected which provided maximum information at the cut-off score subject to the constraint that the final version of the exam must follow the content specifications approved by a content committee.

This study was the first to address the possibility of low content representation when using IRT procedures to focus information. However, the authors did not mention the reason for increasing the content representation, which is to increase the accuracy of the domain percentage correct score estimate. Indeed, there was no analysis of the differences between the percent correct observed scores (domain score estimates) and the domain percent correct scores for the finite population defined. As was the case in previous studies, the authors choose to evaluate classification accuracy in terms of an IRT domain score perspective.

The classical approach involved selecting items that had (a) p-values between .40 and .80; and (b) the highest available classical item discrimination index (point biserial correlations). Additionally, the exam had to be in compliance with the test blueprint.

For this study tests of only twenty items were used. The authors stated that the exam length was kept short to minimize the overlap with the criterion test. The criterion test contained 249 items distributed across 12 different sub-areas within the nursing field. Despite the large number of different subtests, suggesting the possibility of more than one dimension, all IRT models (i.e., one, two and three) fit well based on analysis of the standardized residuals. The authors chose to use the three parameter model for the study because it did produce a slightly better fit. The cut-off scores selected for this study were 65%, 70% and 75%. The distributional shape of the examinee population was not mentioned.

The findings of this study were similar to previous studies in that optimal exams provided three to four times more information than the random exams and resulted in practically significant improvement in classification accuracy based on domain ability scores. Classical exams produced decision accuracy comparable to IRT based exams when cut-off scores were near the center of the distribution. However, they fared less well when cut-off