Comparison of three scoring procedures within context-dependent item sets

Pamela Ann Henderson
University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation
https://digitalscholarship.unlv.edu/rtds/234

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Comparison of three scoring procedures within context-dependent item sets

Henderson, Pamela Ann, M.S.
University of Nevada, Las Vegas, 1992
COMPARISON OF THREE SCORING PROCEDURES WITHIN CONTEXT-DEPENDENT ITEM SETS

by

Pamela A. Henderson

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Educational Psychology

Department of Counseling and Educational Psychology
University of Nevada, Las Vegas
December, 1992
The Thesis of Pamela A. Henderson for the degree of Educational Psychology is approved.

Chairperson, Kevin D. Crehan, Ph.D.

Examining Committee Member, Eunsook Hong, Ph.D.

Examining Committee Member, Peggy G. Perkins, Ph.D.

Graduate Faculty Representative, Evangelos A. Yfantis, Ph.D.

Dean of the Graduate College, Ronald W. Smith, Ph.D.

University of Nevada, Las Vegas
December, 1992
ABSTRACT

It has been theorized and evidenced that traditional reliabilities calculated on tests consisting of context-dependent item sets yield inflated estimates. However, the degree of inflated reliability for different scoring techniques has not been observed. The present study scored three context-dependent item sets as stand-alone items and as separate item sets using three different scoring techniques; number-right, polyweighting and the three-parameter IRT logistic model. Double cross-validation was performed and included in the comparisons. Differences in reliability estimates for the stand-alone and item set treatment were determined for each scoring procedure and compared. These three scoring procedures were also compared to determine which procedure yielded the highest reliability and validity estimates and precision of scoring within the item set treatment. The context-dependent item sets consisted of 10, 20 and 11 items and were randomly chosen from a 300-item licensure exam. Coefficient alpha reliabilities were calculated for the three context-dependent item sets treated as 41 stand-alone items. The context-dependent item sets treated as three separate item sets were combined and coefficient alpha calculated
with a linear combination reliability formula which took into consideration the variances and covariances among the three item sets. Correlations between the 41 items treated as three separate item sets and the remaining items in the 300-item pool was determined to estimate the concurrent criterion-related validity. The average absolute differences between standard scores on the separate item set treatment and the remaining items in the pool indicated scoring precision of the three scoring procedures. Average absolute differences were determined and compared within five ordered score groups based on 20, 40, 60 and 80 percentile ranks to evaluate scoring precision throughout the scale. The findings of this study were not consistent with previous research in that only the polytomous scoring technique yielded inflated reliability estimates. In every comparison, number-right scoring and polyweighting were similar and outperformed the three-parameter IRT ability estimation model. These findings indicate the need for future research on scoring of context-dependent item sets with different data and with comparisons among additional scoring techniques.
# TABLE OF CONTENTS

ABSTRACT...............................................................iii

LIST OF TABLES..........................................................vi

ACKNOWLEDGEMENTS......................................................vii

CHAPTER 1  INTRODUCTION.............................................1
            Polytomous Scoring........................................9
            Item Response Theory.....................................11
            Reliability...............................................13
            Validity..................................................14
            Research Purpose.........................................15
            Research Questions.......................................16

CHAPTER 2 REVIEW OF LITERATURE.................................17
            Context-dependent Item Sets.............................17
            Polytomous Scoring........................................19
            Option-weighting..........................................21
            Dichotomous IRT Scoring..................................25

CHAPTER 3 METHOD....................................................29

CHAPTER 4 RESULTS...................................................32
            Table 1...................................................32
            Table 2...................................................34
            Table 3...................................................35
            Table 4...................................................37
            Figure 1..................................................38

CHAPTER 5 DISCUSSION.................................................40

APPENDIX A.............................................................46

BIBLIOGRAPHY..........................................................47
LIST OF TABLES

TABLE 1: Means and Standard Deviation for Three Scoring Procedures with Items Treated as Sets and as Stand-Alones for Both Groups....32

TABLE 2: Coefficient Alpha Reliabilities for Three Scoring Procedures with Items Treated as Sets and as Stand-Alones for Both Groups....34

TABLE 3: Standard Error of Measurement for Three Scoring Procedures with Items Treated as Sets and as Stand-Alones for Both Groups (Expressed as Proportion of Standard Deviation).................................35

TABLE 4: Average Absolute Differences (SD) Between Item Set Totals (N=41) and Remaining 259 Items for Each Scoring Method Within Five Ordered Score Groups.................................37

FIGURE 1: Average Absolute Differences.........................38
ACKNOWLEDGEMENTS

I would like to extend a very sincere thank you to Dr. Kevin D. Crehan, my mentor and committee chair, to whom I will always be indebted. Without his knowledge and expertise in the field of measurement, his professionalism, extreme patience and confidence in me this project would have not been possible. A special thanks to Dr. Eunsook Hong who not only taught me a lot about statistics and measurement but also a lot about myself. I would also like to extend a sincere thanks to Dr. Peggy Perkins for getting me into this "mess" and always being there with concern and kindness. For her expertise in editing, I would like to thank Dr. Judith Costa. I would also like to thank Dr. Evangelos Yfantis for being a part of this undertaking.

Others I would like to thank are Mona Lauber and Colleen Cascio who were there through sickness and in health and always made me smile. It is with deep appreciation that I thank, Aaron P. Rucker, my true friend who stood by my side in this and every endeavor. Finally, I would like to thank Dillion Brooks Henderson, my son, who gave me the courage and strength to keep giving when I thought I had nothing else to give.
CHAPTER 1

INTRODUCTION

A frequently expressed concern and criticism of objectively scored tests is their limitation to the measurement of low-level thinking skills. Research conducted by Hoffman (1964), Morgenstern and Renner (1984), Stiggins, Griswold and Skikelund (1989) and Baker (1991) have found that the content of most objectively scored tests is at the level of recall. Multiple-choice testing has continually been criticized for emphasizing low-level learning such as the recall of facts. Consequently, traditional multiple-choice items are thought to be incapable of measuring high-level thinking and thus, are deficient in adequately estimating true ability. Ebel (1951) reported context-dependent item sets as a new and promising objective test format. Additionally, Haladyna (1990) has proposed the context-dependent item set as an alternative capable of measuring high-level thinking.

Context-dependent item sets consist of a set of objective items based on an introductory stimulus, which presents information or describes, illustrates, or defines a situation. The introductory data may be in the form of written materials, tables, charts, graphs, maps, or
pictures. The series of related test items may be in any type of test item format, including the essay and completion, but the multiple-choice or true-false item format is usually the format of choice (Gronlund, 1990). The context-dependent item set has been referred to by other terms such as interpretive exercise (Ebel, 1951), superitems (Cureton, 1965), item bundles (Rosenbaum, 1988), narrative exam (Terry, 1980), application test (Szeberenyi and Tigy, 1987), testlet (Wainer and Kiely, 1987) and profile items. The current study will use the term "context-dependent item set" to refer to this type of test item format.

Gronlund (1990) projects that the context-dependent item set has the ability to measure many complex learning outcomes that are based on higher order thinking skills such as understanding, reasoning, critical thinking, scientific thinking and creative thinking, and problem solving. Szeberenyi and Tigy (1987) in a twelve-year study in the medical biology field, conclude that context-dependent item sets assess higher levels of the cognitive domain such as interpretation and problem solving. They also state that factual knowledge is necessary but not sufficient to solve context-dependent item set tests. With context-dependent item sets, sufficient background information is presented in the common set of data in a systematic and objectively scorable manner. Thus, understanding, thinking, problem
solving or multi-step thinking can be observed with very little anxiety about recall or the need for additional information. Terry (1980) added that the emphasis is on application rather than memorization. Therefore, cognitive behaviors and test items are combined, serving as evidence that a test format may influence what and how the learner learns (Haladyna, 1992). Context-dependent item sets are efficient for measuring specific aspects of the problem-solving process, thus, providing a diagnostic view of an examinee's problem solving skills rather than a more holistic view as in essay questions. Additionally, this structure is an advantage over the essay format in that examinees cannot redefine the problem so as to demonstrate those thinking skills in which they are most proficient. The series of objective items forces them to use only the mental processes dictated. Hence, separate aspects of problem solving can be measured with objective scoring procedures (Gronlund, 1990).

Although construction of context-dependent item sets is more difficult and time consuming than other types of objective tests, context-dependent item sets have the advantage over "stand-alone" item formats of measuring higher-level cognition; in addition, can be scored objectively as opposed to the essay test format, which is more subjectively scored. Also, the context-dependent item set is more efficient, allowing a broader sampling of
content than the essay for a fixed testing period. Objective scoring and greater efficiency lead to better precision of measurement for the context-dependent item set. To alleviate construction difficulties, Haladyna (1988) has recently introduced item-generating algorithms which provide a means for producing many items in an automated way while allowing the item writer freedom in developing the introductory passage.

However, one major problem with context-dependent item sets exists. Traditionally, test structure has been ignored and context-dependent item sets have been scored as stand-alone items rather than separate item sets. This results in inflated reliability estimates due to the correlated errors within item sets in classical test theory or violation of the local independence assumption in item response theory. Therefore, there is not fine discrimination among levels of performance as reliability estimates might indicate.

This concern has long been recognized. Kelley (1924) commented that the Spearman-Brown reliability coefficient would be too large when two or more exercises contain common features. Guilford (1936), Thorndike (1951), and Anastasi (1961) warned about inflated reliability estimates when a group of items dealing with a single problem is scored as separate items. Cureton (1965) argued that errors of measurement in classical test theory are correlated for
items within context-dependent item sets. Yen (1992) stated that context-dependent item set scores result in less test information than do scores based on separate items for two reasons. First, items within context-dependent item sets are indeed locally dependent; therefore, treating these items separately exaggerates the information these items contribute to the total. Also, a context-dependent item set analysis results in greater SEM than a separate item analysis due to information lost when the items are combined into a context-dependent item set, disregarding the pattern of item responses. However, she stated this greater SEM is more realistic.

According to Wainer and Lewis (1990), an examinee who gets one item from the context-dependent item set correct tends to respond correctly to the other items in the context-dependent item set. Response to one item within a context-dependent item set may govern ones responses to other items as comprehension of the introductory passage influences one's comprehension or ability to respond to the entire series of related items. Therefore, local independence in item response theory is threatened, creating a dilemma concerning how to score results (Haladyna, 1992).

Rosenbaum's (1988) Theorem 1 states that there can be unidimensionality---therefore local independence---between context-dependent item sets. There is a single latent variable underlying the general proclivity to respond
correctly to questions following passages on various topics in the same domain (Thissen, Steinberg & Mooney, 1989). The entire test represents a relatively unitary domain and the context-dependent item sets with different content or topics represent related aspects of this larger domain (Haladyna, 1990). It is the common factor between passages that we wish to measure.

According to Yen (1992), context-dependent item sets are a most efficient way to relate performance on one set of items to the other items in the test. They provide more accurate information because they involve dependent items which is consistent with the purpose of authentic tests.

Wainer and Lewis (1990) outlined three approaches to the context-dependent item sets potential violation of local independence. One was to use only a single item for each passage; however, this was deemed to result in a cost inefficiency. A second approach was to ignore the interdependencies among the related items, fit a dichotomous IRT model and hope for the best. Thirdly, and preferably, is the context-dependent item set approach, which treats the passage and its related items as a single unit.

The second and third approaches were examined by Thissen, Steinberg, and Mooney (1989). They expanded Wainer and Kiely's (1987) concept of the context-dependent item set used in Computerized Adaptive Testing in the analysis of a reading comprehension test consisting of four passages with
22 related items. In a dichotomous/polytomous IRT model comparison, they found that scoring at the item level with a three-parameter logistic model was erroneous, as local independence was violated. This resulted in overestimation of the precision of measurement as characterized by the test information function. The item level reliability was .08 (12-13%) higher than the context-dependent item set reliability. They proposed scoring with context-dependent item set response models where the focus is still at the level of the individual item but within a context-dependent item set format. The context-dependent item sets, as separate item sets, are locally independent; therefore, unidimensional scoring is appropriate.

More recently, Sireci, Thissen, and Wainer (1991) found that failing to consider the dependencies within four context-dependent item sets resulted in a 10-15% overestimation of reliability. This would equate to doubling the test length to eight passages to achieve this higher reliability.

Yen (1992) also preferred scoring context-dependent item sets as separate item sets, thereby creating subscores as opposed to traditional stand-alone or individual item level scoring.

Unfortunately, there has not been very much research on scoring context-dependent item sets as separate item sets versus treating them as consisting of stand-alone items.
Nor has there been much research on scaling, item analysis, validity and reliability of different scoring methods within the context-dependent item sets treated as stand-alone or separate item sets. Since context-dependent item sets are being widely used in standardized competency, certification, and licensure testing the best possible scoring method should be determined.

Lord (1980), Yen (1984b), and Yen and Candell (1991) stated there is more information in a score that considers the pattern of item responses than in one based on number correct. However, when context-dependent item sets are considered a unit, a greater SEM may occur due to loss of useful information in the pattern of item responses (Yen, 1992).

Ward (1985) projects that future item response theory models should parameterize sets of questions as opposed to individual questions to deal with problems of local item dependence.

Newly proposed scoring procedures may provide more diagnostic and summative information than obtained from other formats. The problem with local dependence within item sets may possibly be overcome with the use of polytomous scoring (Haladyna, 1990). Polytomous scoring estimates partial knowledge, not just whether an examinee got the item right or wrong. Wainer & Kiely (1987) and Thissen, Steinberg, and Mooney (1989) illustrated the
context-dependent item set as a measure of complex thinking with scoring enhanced by polytomous models.

**Polytomous Scoring**

Polytomous scoring assumes that distractors are unequally informative; therefore, they can and should be weighted differentially. However, polytomous scoring can only be effective if the distractors are of high quality and operate as intended.

Levine and Drasgow (1983) and Thissen, Steinberg and Fitzpatrick (1989) found that there is a systematic relationship between the correct and incorrect answers in multiple choice items.

Samejima (1984) indicated that most items have distractors that are differentially informative. Thus, the existence of patterned relationships between distractors and total test score is well documented.

Recent polytomous IRT models introduced by Bock (1972), Samejima (1979), Sympson (1981, 1983, 1986, 1989), Thissen and Steinberg (1984), and Thissen, Steinberg, and Mooney (1989) have provided an impetus for using distractors as a basis for improving the scoring of test results. The use of innovative test formats such as the context-dependent item set seems to be compatible with polytomous scoring methods (Haladyna, 1992). Unfortunately, computer programs currently available can handle only very small item sets.
With the increase in high-stakes testing such as licensure exams and the use of context-dependent item sets, there is a need for scoring methods that will increase reliability. Small increases in reliability can be extremely important.

Guttman (1941) proposed a scoring strategy, max-alpha, for multiple-choice tests based on a theory that maximized coefficient alpha. Response categories were given a value equal to the mean of the total test score for all examinees choosing that response category or option. This initial weight is called the option mean. Guttman's procedure iterates these weights until a criterion of stabilization in alpha is attained.

Polyweighting, introduced by Sympson (1988), is a polytomous scoring method which derives from max-alpha but replaces option mean with percentile rank in the initial weighting of options. As with max-alpha the initial weights are iterated to a criterion of stability in alpha. An examinee's "polyscore" is equal to the mean of the polyweights of the categories chosen by the examinee. Items and options are weighted according to their difficulty. Thus, polyweighting results in the examinee being given more credit for correct answers to difficult items and less credit for correct answers to easy items. Also, the examinee is penalized more heavily when choosing a response other than the keyed response to easy items than when
choosing a distractor in more difficult items. This is in contrast to the traditional number-correct dichotomous scoring.

The use of percentile ranks as option weights is analogous to equipercentile equating and eliminates the sample dependence of polyweights (Sympson, 1988). Another advantage to polyweighting over max-alpha is that scoring weights are bounded so that an examinee can never receive more credit for an incorrect response than for a correct response. Unlike polytomous IRT models, polyweighting can be applied to small samples and does not require assumptions of local independence, unidimensionality, "latent" abilities, or the mathematical form of the regression of item responses on unobservable factors.

Item Response Theory

Item response theory is concerned with whether an examinee got each individual item correct or not rather than with the raw test score. An examinee's item response vector and item parameter estimates are used to estimate his/her unknown ability parameter. At each ability level there will be a certain probability that an examinee with that ability will give a correct answer to that item. For each item there is an S-shaped curve called the item characteristic curve, or ICC, which describes the relationship between the probability of a correct response to an item and the ability
scale (Baker, 1985). The discrimination parameter, "a," describes how well an item can differentiate between examinees having ability below the item location and those having abilities above the item location. The difficulty parameter, "b," is a location index which locates the point on the ability scale where the item functions best. It is the point on the ability scale where 50% of the examinees will get the item correct if "c" = 0. The parameter "c" is the probability of getting the item correct by guessing alone. The value of the guessing parameter does not vary as a function of the ability level. In addition to estimation of the "a" and "b" parameters, the three-parameter model includes the guessing parameter "c."

All IRT models have the assumptions of local independence and unidimensionality. When all the ability dimensions influencing performance have been considered, local independence is obtained. After taking examinees' abilities into account, no relationship exists between the examinees' responses to different items. If the complete latent space consists of only one ability, the assumption of unidimensionality is met.

In the traditional number-correct dichotomous scoring of multiple-choice test items, all correct answers are given a weight of one and all incorrect answers given a weight of zero. The three-parameter item response model is dichotomously scored, with all incorrect answers treated
equally, but the correct answer is differentially weighted as a function of theta or ability score. The same correct response to a given item will receive different credit for examinees with different ability. Thus, in the three-parameter model, examinees with the same raw score could obtain different ability estimates because they answered different items correctly (Baker, 1985). Small weights are assigned to items difficult for the ability level and larger weights to easier items for the ability level. This, in effect, nullifies guessing.

Dichotomous scoring, whether in a traditional sense or in the three-parameter logistic item response model, implies that all distractors are equally informative and grouped together in a single incorrect category, and they are therefore weighted equally. This is true, even though the correct response in the three-parameter IRT model may be weighted differentially.

Reliability

When coefficient alpha is high, examinees have performed consistently across items within a test, and the test is said to have "item homogeneity." In other words, the test measures the same type of performance or represents the same content domain. This results in more true score variance.

Polytomous scoring seems to have the effect of
purifying the trait, making it more homogeneous, therefore more reliable (Haladyna & Symson, 1988). Patnaik and Traub (1973) found inter-item correlations of weighted scores were higher on the average than the inter-item correlations of number-correct scores.

Validity

The most important concern in testing is validity. Consequently, any study of test scoring should address this issue (Haladyna, 1988). In achievement testing, decisions are made about the examinee's levels of information or misinformation. Thus, scoring methods that reflect such levels should enhance test validity. A scoring method needs to be chosen with the perspective of test validity in mind (Jaradat and Tollefson, 1988).

Haladyna & Symson (1988) report that the effect of polytomous scoring on concurrent validity is hard to interpret. They found that polytomous scoring tends to purify the trait measure of the test, resulting in reduced predictive and concurrent correlations. Research has found that when the criteria and predictor are similar, polytomous scoring is favored, but when they are dissimilar this is not the case (Haladyna and Symson, 1988).

Kemerer and Wahlstrom (1985) stated that the context-dependent item set has excellent predictive validity because the choices can be structured to reflect actual
decisions which must be made by the examinee in the workplace. Thissen, Steinberg, & Mooney (1989) found that concurrent validity of context-dependent item sets exceeded that of "stand-alone" items when correlated with an external concurrent validity criterion.

Research Purpose

More research on scoring of context-dependent item sets to improve reliability, precision of scoring and validity should be conducted (Haladya, 1990). Therefore, this study is designed to compare among three scoring methods for context-dependent item sets treated as separate item sets or stand-alone items. The multiple-choice scoring and scaling methods employed in this study will be number-right scoring, polytomous scoring, and item response theory ability estimation. The related items in the current study are of the multiple-choice format. (See Appendix A.)

The difference in reliability estimates for the items treated as separate item sets and as stand-alone items will be evaluated and compared for each scoring method. The three scoring methods will then be compared to determine which yields the highest internal consistency reliability and concurrent criterion-related validity estimates within the item set treatment. Precision of scoring for each method will be determined by an average absolute difference analysis.
Research Questions

As the above discussion indicates, there are a number of areas to be considered in making comparisons between number-right scoring, polytomous scoring and IRT ability estimation as applied to context-dependent item sets. Presented below are the research questions to be investigated:

Question 1- Do context-dependent item sets scored as separate item sets differ on reliability as compared to treating these same items as stand-alone multiple-choice items?

Question 2- Do differences exist on internal consistency reliability among number-right scoring, Sympsom's polyweighting, and the three-parameter IRT ability estimation within the item set treatment?

Question 3- Do differences exist on average absolute differences among number-right scoring, Sympsom's polyweighting, and the three-parameter IRT ability estimation?

Question 4- Do differences exist on concurrent criterion-related validity among number-right scoring, Sympsom's polyweighting, and the three-parameter IRT ability estimation within the item set treatment?
CHAPTER 2

REVIEW OF LITERATURE

Three studies have compared reliability and validity estimates of a context-dependent item set treated as stand-alone items and as separate item sets or subtests (Sireci, Thissen, and Wainer, 1991; Thissen, Steinberg, and Mooney, 1989; and Yen, 1992).

Context-dependent Item Sets

Thissen, Steinberg, and Mooney (1989) in an item set and stand-alone comparison used the standard three-parameter logistic model for the test as 22 items and a constrained nominal model for the test as four passages. They found that the test scored as individual items violated the assumption of local independence; therefore, the information curve, or reliability, was overestimated or deceptive, and scoring at the item level yields misinformation. This was caused by the three-parameter logistic model's inability to deal with the excess intra-passage correlations among the items. The concurrent validity of the item set scores was equal to or greater than that of items scored individually.

Sireci, Thissen, and Wainer (1991) also used the conventional three-parameter logistic model for comparing
context-dependent item sets scored as individual items to those same items scored as a unit using Bock's (1972) nominal model. They, too, found that failing to take into consideration the dependencies caused by having four sets of items relating to a common passage resulted in a 10-15% overestimate of reliability.

Yen (1992) compared two tests scaled three ways. In the first scaling all of the items were treated as separate units. The second scaling compared items within context-dependent item sets with stand-alone items. The third scaling involved only context-dependent item sets but the items chosen were all as locally independent as possible. The information from the separately scaled items was twice as great as the information from the locally dependent items treated as a unit. In the locally independent context-dependent item set analysis, results showed little effect on information. In other words, it is not the context-dependent item set format that results in less information.

Albanese and Jacobs (1990), although not comparing different treatments of the context-dependent item set, assessed the reliability and validity of 33 multiple-choice items consisting of stimulus material. In a test-retest alternate-form reliability study they found good consistency across context-dependent item sets. They also found support for construct validity.
Kemerer and Wahlstrom (1985) designed an instrument consisting of context-dependent item sets based on previous essay questions. Content experts chose distractors which distracted from the correct answer to approximately the same degree. This resulted in a test which discriminated very effectively between those who knew the correct answer and those who were guessing. Twelve individuals ranging in ability from very high to very low were selected. On every question the high ability examinees outperformed the ones perceived to have low ability. They concluded the context-dependent item set instrument was very credible in that the position ranking of the examinees was significant beyond the chance level.

Polytomous Scoring

It has only been during the last 40 years that the possibility of differentially weighting the incorrect options of an item has been examined. The question of concern is: Can weighted response options of items on ability and achievement tests improve reliability and validity? There is clear evidence supporting the existence of partial information. Throughout the previous research it has been demonstrated that wrong answers are not randomly distributed.

Levine & Drasgow (1983) found incorrect option choice to be related to estimated ability for many items. Wrong
answers definitely have psychological and statistical meaning (Powell, 1968). Partial knowledge, which is not considered in conventional scoring, often serves as a basis for responding. This inability of the conventional method to discriminate between partial information and complete information or misinformation is a disadvantage (Abu-Sayf, 1979). Therefore, employing proper scoring techniques may increase reliability (Coombs, Milholland, & Womer, 1956; and Dressel and Schmid, 1953). Researchers have investigated the hypothesis that dichotomous scoring of multiple-choice items does not encompass the full information available in the responses concerning a person's ability (Bock, 1972; Claudy, 1978; Hambleton, Roberts, and Traub, 1970; Hendrickson, 1971; Nedelsky, 1954; Sabers & White, 1969; Waters, 1976); among others. The majority of these authors have found that quantifying the degree of incorrectness of the answer adds an additional source of information about the person's ability. Raffeld (1975) supported the belief that a Guttman-weighted objective test can have psychometric properties that are superior to those of dichotomous items. To overcome this perceived deficiency of dichotomous scoring, polytomous techniques such as response weighting, answer until correct (Frary, 1982), confidence weighting (Dressel & Schmid, 1953; Michael, 1968; Abu-Sayf and Diamond, 1976), elimination scoring (Coombs, Milholland & Womer, 1956; Hakstian & Kansup
polytomous IRT models (Bock, 1972; Changas & Samejima, 1984; DeAyala & Koch, 1987; Levine & Drasgow, 1983; Smith, 1987; Symponent, 1983, 1986; Thissen, 1975; Thissen and Steinberg, 1982), distractor identification (Austin, 1981); among others, have been developed. The most popular of these methods has been response or option weighting. Stanley and Wang (1970) reported that response option weighting might be effective in increasing reliability and validity since it considers the variance from the incorrect responses.

Option-weighting

There are two types of option weighting, theoretical a priori and empirical. The a priori method assigns weights to response alternatives based on degree of correctness determined by judges or a theory. Empirical option weighting uses the response data from a sample of examinees to determine the response weights.

The following studies have compared empirical or a priori option weighting scores with unweighted scores. Guilford, Lovell, and Williams (1943) empirically weighted responses to achievement test items and found an insignificant gain of .02 in reliability and validity coefficients in favor of the weighted scores. Research conducted by Nedelsky (1954) was the first examination of a priori differential weighting of incorrect responses.
Unique weights were assigned to the correct response and the least correct distractor while all other distractors were assigned weights of zero. He found a slight gain in reliability and validity, as did Lord (1965). The landmark study of Davis and Fifer (1959) employed modified empirical option weighting and significantly raised the cross-validated comparable-forms reliability of a 45-item arithmetic-reasoning test from .68 to .76, without lowering its validity. Patnaik and Traub (1973) found a priori weighted scores more reliable but less valid when correlated with mean school achievement. Jacobs and Vanderventer (1968) used facet analysis (Guttman & Schlesinger, 1967) in an a priori method of keying the response alternatives which showed a moderate degree of test-retest reliability, and concurrent and predictive validity. Sabers and White (1969) also found a slight increase in reliability and predictive validity due to empirical option weighting. Again, Haladyna (1988) found empirical option weighting to be superior to conventional scoring in terms of reliability and validity. Hambleton, Roberts, and Traub (1970) in an a priori option weighting study found predictive validity to be increased; however, reliability was decreased. Conversely, Hendrickson (1971) reported using Guttman's (1941) max-alpha technique in a large scale, doubly cross-validated study, results of a substantial increase in internal consistency. Reilly & Jackson (1973) also used Guttman's (1941) max-alpha...
technique and found that tests scored with the empirical weights were more reliable than formula scores but had less predictive validity when correlated with undergraduate GPA.

Willson (1982) maximized coefficient alpha when using the Serlin-Kaiser procedure of empirical option weighting. This suggests that test reliability and homogeneity can be increased by weighting options empirically. This homogeneity may account for the decrease in validity when the criterion is factorially heterogenous. Waters (1975) and Downey's (1976) findings, using empirical option-weighting techniques, were consistent with those reporting increases in reliability and decreases in validity.

Echternacht (1976) compared a priori and empirical option weighting with each other and with conventional scoring with respect to internal consistency and concurrent validity. Results favored empirical option weighting over a priori option weighting for reliability and validity. Razel & Eylon (1987) also compared a priori and empirical option weighting with conventional scoring. Results indicated that option weighting was superior to conventional scoring due to increases in reliability and validity. Empirical option weighting, was again the most efficient scoring method. Bejar and Weiss (1977) found a priori weights yielded consistently more reliable scores than two empirical option weighting methods. Also, the two empirical option weighting
methods performed less reliably than did conventional scoring at low homogeneity levels. As a result they found differential option weighting to be a function of inter-item correlations.

In a recent study, Haladyna (1989) found that two types of empirical option weighting methods—modified reciprocal averages (Guttman, 1942) and point-biserial—yielded higher reliability and content validity indexes; however, both polytomous scoring methods were inferior to conventional scoring when the size and direction of error were considered. They also were inferior to conventional scoring in terms of predictive validity. These findings affirmed the earlier consensus that polytomous scoring has the tendency to make the trait more homogeneous, therefore more internally consistent. However, this results in poor predictive validity when the criterion is dissimilar to the predictor.

Haladyna's (1989) study has several limitations which should be considered before generalizing the results. First, a modified reciprocal averages was used which proved to not be as effective as previously thought. Also, the three test forms consisted of only 10, 20, and 30 items from a total test of merely 127 items. Consequently, a new linear procedure, polyweighting, introduced by Sympson (1988), was projected to be more effective in both small and large testing programs. Thus, polyweighting is the
empirical option weighting method of choice in this study.

Sympson & Haladyna (1988) evaluated the effect of polyweighting on the reliability and domain validity of test scores from a 200-item test on 1100 examinees in the medical profession. Polyweighting resulted in higher internal consistency and domain-related validity than did conventional scoring.

Option weighting methods, such as polyweighting, have been classified as linear polytomous scoring methods by Haladyna (1989), as opposed to nonlinear methods involving trace lines or option characteristic curves for each option in item response models. The item response models in this study will be dichotomously scored.

**Dichotomous IRT Scoring**

In view of the limited amount of research done on IRT scoring models this study has opted to employ the three-parameter logistic model. This IRT model scores items dichotomously but applies differential weights to items and ability levels. Shea, Norcini, and Webster (1988) found that the three-parameter IRT model fit well to questions that measure medical knowledge with an emphasis on those medical licensure exams that measure more sophisticated measures of cognitive skills. This could be extended to context-dependent item sets, which are used in many licensure exams such as those in the medical profession.
Ferris (1967) found that using item weights computed by the three-parameter model does not improve predictive validity. However, correlations between the weighted scores and the raw criterion scores were higher than those between the unweighted scores and the raw criterion scores. Thus, it may be beneficial to use the three-parameter logistic model to determine indices of difficulty of test items for improving predictive validity. It is also useful in item analysis to pinpoint items that have no real effect on the intertest correlation.

Lord (1968) applied the three-parameter model to Scholastic Aptitude Test Verbal scores and reported encouraging results in favor of this IRT model. He stated there was a need for weighted scores at low ability levels due to random guessing. However, number-right scores were more efficient than weighted scores at high ability levels. Therefore, he concluded there were limitations to this model but that it seemed promising.

Yen (1983), in a CTB/McGraw-Hill test development procedure, converted traditional standardization procedures to IRT procedures. Number-right estimates were found to have higher observed score standard deviations, higher standard errors of measurement and lower split-half correlations. Yen (1984) compared number-correct and item response pattern scoring using the three-parameter model. She found that the number-correct trait estimates were
similar to those estimated by the three-parameter model. However, standard errors were higher for number-correct trait estimates with both empirical and simulated data, especially for low-scoring examinees. She stated that with the three-parameter model, "noise" due to guessing is reduced, resulting in more accurate and stable scores. Thus, the greatest difference between the two scoring methods was for low-scoring examinees. The decrease in the standard error due to item response pattern scoring is equivalent to a substantial increase in items, especially for low-scoring examinees.

Huynh and Casteel (1987), using Bock's (1972) multi-nominal model, also found item response scoring to be similar to number-correct scoring.

Thissen (1976), in a dichotomous IRT (Birnbaum 1968) and polytomous IRT (Bock, 1972), comparison found that for the lower half of the ability range, the polytomous IRT model provided from one-third more to nearly twice the information of the dichotomous IRT model. However, there was no substantial difference for the upper half of the ability range. The overall reliability of the test was approximately the same for both scoring methods. This indicates that increases in reliability shown by using linear polytomous methods may come from the lower part of the ability range. This is due to using information in wrong responses which are more prevalent in the lower
ability range.

Drasgow, Levine, Williams, McLaughlin, and Candell (1989) found the test information function of a polytomous IRT model was larger than a three-parameter logistic information function for low to moderately high ability levels. This was consistent with earlier studies.

Yen and Candell (1991) compared number-correct scoring with the three-parameter IRT model within five score metrics. The three-parameter IRT model was more reliable in every metric.
CHAPTER 3

METHOD

Data for this study were obtained from a 300-item licensure exam with scores on 3000 examinees. The mean score for the entire population of examinees was 113.645, with a standard deviation of 13.929 and a standard error of measurement of 4.947. The original raw score distribution, scored number-right, was slightly negatively skewed, -0.524, with the coefficient of kurtosis -.161. Coefficient alpha for the 300 items was .874. The average item difficulty was approximately .75.

In the present study, double cross-validation was applied to an equal size, random split of the 3000 examinees. Three context-dependent item sets with a total of 41 items were selected for analysis. The item sets contained 10, 20, and 11 items, respectively. The procedure was to score each of these three context-dependent item sets as stand-alone items and as separate item sets. Each test treatment was scored using traditional number-right, Sympson's polyweighting, and the three-parameter IRT scoring methods. Programs used for scoring were "Iteman" for number-right, "Poly" for polyweighting and "Ascal" for the three-parameter IRT ability estimation.
Total scores for each examinee were determined for each test treatment and each scoring method. Stand-alone scores were summed to produce a total score for each examinee for the entire 41 items. Items treated as three item sets produced three separate scores for each examinee. These scores were then linearly combined, (Nunnally, 1967) to produce one composite score for each examinee.

The two treatment formats, item set and stand-alone, scored with the three procedures, were compared to determine the difference in coefficient alpha. For the stand-alone treatment, coefficient alpha was calculated based on total scores from the entire 41 items. Reliability for the item set treatment was calculated by a linear combination reliability formula which takes into account the variance of each item set and the covariances among them.

Average absolute differences were calculated to assess the relative precision among the scoring methods. The average absolute difference between the 41 items treated as three separate item sets and the remaining 259 (250 for number-right and IRT) items was determined for each scoring method. Scores from each scoring method were converted to standard "z" scores and examinees placed in five ordered score groups based on the 20th, 40th, 60th and 80th percentile ranks within the separate item set treatments. For each ordered score group the average absolute difference between the "z" score on the linear combination of item sets
and the "z" score on the remaining items scored with each method was calculated to determine the relative precision of the three scoring methods.

The three context-dependent item sets treated as separate item sets were used as predictors in a concurrent criterion-related validity study. The remaining 259 items in the test (250 for IRT and number-right) were used as the criterion. Total scores from the linear combination of the separate item sets were obtained for each scoring method and correlated with the criterion scored conjunctively.
CHAPTER 4

RESULTS

Table 1 contains means and standard deviations for each scoring method by group consisting of 1500 examinees each.

Table 1
Means & Standard Deviations for Three Scoring Methods with Items Treated as Sets and as Stand-Alones for Both Groups

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Test Format</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MEAN</td>
<td>SD</td>
</tr>
<tr>
<td>Number Right</td>
<td>Stand-alone</td>
<td>30.7</td>
<td>4.21</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>7.6</td>
<td>1.56</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>14.5</td>
<td>2.64</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>8.5</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>30.7</td>
<td>4.21</td>
</tr>
<tr>
<td>Poly</td>
<td>Stand-alone</td>
<td>50.2</td>
<td>2.05</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>51.3</td>
<td>4.61</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>50.2</td>
<td>2.93</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>50.8</td>
<td>3.68</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>152.2</td>
<td>8.36</td>
</tr>
<tr>
<td>3-Parameter IRT</td>
<td>Stand-alone</td>
<td>-39.1</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>-8.9</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>-14.5</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>-11.7</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>-35.2</td>
<td>3.96</td>
</tr>
<tr>
<td>Poly</td>
<td>Stand-alone</td>
<td>49.7</td>
<td>2.02</td>
</tr>
<tr>
<td>Cross-Validations</td>
<td>Item Set 1</td>
<td>51.2</td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>50.0</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>50.4</td>
<td>3.86</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>151.6</td>
<td>8.38</td>
</tr>
<tr>
<td>3-Parameter IRT</td>
<td>Stand-alone</td>
<td>-41.9</td>
<td>4.88</td>
</tr>
<tr>
<td>Cross-Validations</td>
<td>Item Set 1</td>
<td>-9.3</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>-16.0</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>-12.5</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>-37.9</td>
<td>4.37</td>
</tr>
</tbody>
</table>

32
Observed differences in means and standard deviations between group one and group two were negligible for number-right and polyweighting. However, for IRT the difference in means varied by approximately 7/10 of a standard deviation. For items scored with IRT and treated as separate item sets and stand-alones, the standard deviations were about the same, yet the means differed by almost a standard deviation for both groups in the original sample. For cross-validations scored with IRT, the average difference in means was approximately 1/2 of a standard deviation. Separate item set and stand-alone means for the cross-validations differed by about 8/10 of a standard deviation for group one and by 9/10 of a standard deviation for group two.

Table 2 presents coefficient alpha reliabilities for each scoring method and group for items scored as stand-alone and item set. For number-right scoring and IRT ability estimation, coefficient alphas for the two test treatments were basically equivalent for both groups, .61 and .48 for the first group and .62 and .52 for the second group, respectively. Coefficient alphas for the IRT cross-validations were also fairly consistent for both treatments across both groups, .52 and .49. Poly, however, showed a considerable difference (.073 and .064; .053 and .052) in coefficient alpha for both groups of the original sample and cross-validations when items were scored as stand-alones rather than as separate item sets. On average, number-right
alpha reliabilities were .623, poly reliabilities about .628, and IRT reliabilities approximately .502 when treated as linearly combined item sets. Reliabilities of the cross-validated scores for poly were slightly higher than number-right alphas for group one.

Table 2
Coefficient Alpha Reliabilities for Three Scoring Methods with Items Treated as Sets and Stand-alone for Both Groups

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Test Format</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Right</td>
<td>Stand-alone</td>
<td>.616</td>
<td>.626</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>.375</td>
<td>.330</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>.490</td>
<td>.488</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>.270</td>
<td>.330</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.619</td>
<td>.628</td>
</tr>
<tr>
<td>Poly</td>
<td>Stand-alone</td>
<td>.698</td>
<td>.696</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>.410</td>
<td>.385</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>.562</td>
<td>.572</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>.349</td>
<td>.383</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.625</td>
<td>.632</td>
</tr>
<tr>
<td>3-Parameter IRT</td>
<td>Stand-alone</td>
<td>.484</td>
<td>.522</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>.224</td>
<td>.250</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>.364</td>
<td>.407</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>.191</td>
<td>.236</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.483</td>
<td>.521</td>
</tr>
<tr>
<td>Poly</td>
<td>Stand-alone</td>
<td>.673</td>
<td>.672</td>
</tr>
<tr>
<td>Cross-Validations</td>
<td>Item Set 1</td>
<td>.396</td>
<td>.355</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>.542</td>
<td>.543</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>.297</td>
<td>.350</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.620</td>
<td>.620</td>
</tr>
<tr>
<td>3-Parameter IRT</td>
<td>Stand-alone</td>
<td>.515</td>
<td>.489</td>
</tr>
<tr>
<td>Cross-Validations</td>
<td>Item Set 1</td>
<td>.235</td>
<td>.237</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>.404</td>
<td>.369</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>.216</td>
<td>.221</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>.522</td>
<td>.491</td>
</tr>
</tbody>
</table>

The difference in reliability between number-right scoring and poly of .005 equates to lengthening a linearly
combined test scored number-right with .623 reliability by a negligible 2%. However, the difference in reliability of .126 between linearly combined item sets scored with poly and IRT ability estimation results in lengthening an IRT scored test by 67% to attain a reliability of .628.

Standard errors of measurement (SEM), expressed as a proportion of a standard deviation, are reported in Table 3.

<table>
<thead>
<tr>
<th>Scoring Method</th>
<th>Test Format</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Right</td>
<td>Stand-alone</td>
<td>2.61 (.619)</td>
<td>2.56 (.613)</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>1.24 (.792)</td>
<td>1.23 (.821)</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>1.89 (.714)</td>
<td>1.85 (.713)</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>1.29 (.858)</td>
<td>1.24 (.816)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2.59 (.615)</td>
<td>2.54 (.609)</td>
</tr>
<tr>
<td>Poly</td>
<td>Stand-alone</td>
<td>1.12 (.546)</td>
<td>1.11 (.549)</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>3.54 (.767)</td>
<td>3.42 (.782)</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>1.93 (.658)</td>
<td>1.89 (.651)</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>2.96 (.804)</td>
<td>3.08 (.783)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5.01 (.599)</td>
<td>4.98 (.600)</td>
</tr>
<tr>
<td>3-Parameter IRT</td>
<td>Stand-alone</td>
<td>3.18 (.717)</td>
<td>3.26 (.691)</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>1.36 (.884)</td>
<td>1.35 (.878)</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>1.98 (.794)</td>
<td>2.05 (.769)</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>1.52 (.901)</td>
<td>1.51 (.869)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2.84 (.719)</td>
<td>2.88 (.691)</td>
</tr>
<tr>
<td>Poly Cross-Validations</td>
<td>Stand-alone</td>
<td>1.16 (.574)</td>
<td>1.08 (.568)</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>3.51 (.776)</td>
<td>3.48 (.801)</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>1.95 (.674)</td>
<td>1.87 (.675)</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>3.23 (.836)</td>
<td>2.87 (.803)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5.14 (.613)</td>
<td>4.86 (.613)</td>
</tr>
<tr>
<td>3-Parameter IRT Cross-Validations</td>
<td>Stand-alone</td>
<td>3.40 (.697)</td>
<td>3.08 (.715)</td>
</tr>
<tr>
<td></td>
<td>Item Set 1</td>
<td>1.44 (.871)</td>
<td>1.29 (.870)</td>
</tr>
<tr>
<td></td>
<td>Item Set 2</td>
<td>2.13 (.772)</td>
<td>1.93 (.791)</td>
</tr>
<tr>
<td></td>
<td>Item Set 3</td>
<td>1.59 (.885)</td>
<td>1.46 (.884)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3.02 (.691)</td>
<td>2.75 (.714)</td>
</tr>
</tbody>
</table>
Overall, when compared with number-right scoring and IRT ability estimation, poly performed consistently better. When item sets were treated as separate items and linearly combined, the average SEM for polyweighting was .599 as compared to an average SEM of .612 for number-right and .705 for IRT ability estimation. This translates to 2% less error variance for polyweighting over number-right scoring and 15% less error variance for polyweighting over IRT ability estimation for the original samples and 13% for the cross-validations. Thus, polyweighting appears to be the more accurate measure, representing more true score variance. Number-right scoring when compared with IRT ability estimation resulted in a SEM that was on average 12% lower when item sets were linearly combined.

Table 4 outlines results of the average absolute difference analysis. The five ordered score groups are based on the 20th, 40th, 60th and 80th percentile ranks within the separate item set treatments. The first ordered score group is comprised of those examinees scoring at or below the 20th percentile rank. Each successive ordered score group represents higher ranking examinees culminating in the fifth ordered score group which consists of examinees scoring at or above the 80th percentile rank. Number-right scoring and polyweighting were more precise than IRT, with an average absolute difference of .57 and .58 as compared to .64, respectively.
Table 4  
Average Absolute Differences (SD) Between Item Set Totals (N=41) and Remaining 259* Items for Each Scoring Method Within Five Ordered Score Groups

<table>
<thead>
<tr>
<th>ORDERED SCORE GROUPS</th>
<th>SCORING METHOD</th>
<th>FIRST</th>
<th>SECOND</th>
<th>THIRD</th>
<th>FOURTH</th>
<th>FIFTH</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR (G1)</td>
<td>.73(.51)</td>
<td>.56(.39)</td>
<td>.52(.40)</td>
<td>.48(.37)</td>
<td>.51(.42)</td>
<td>.57(.44)</td>
</tr>
<tr>
<td></td>
<td>NR (G2)</td>
<td>.77(.58)</td>
<td>.60(.43)</td>
<td>.52(.41)</td>
<td>.51(.38)</td>
<td>.49(.42)</td>
<td>.57(.46)</td>
</tr>
<tr>
<td></td>
<td>POLY (G1)</td>
<td>.66(.46)</td>
<td>.54(.42)</td>
<td>.48(.39)</td>
<td>.52(.43)</td>
<td>.59(.47)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>POLY (G2)</td>
<td>.58(.44)</td>
<td>.54(.44)</td>
<td>.47(.38)</td>
<td>.47(.40)</td>
<td>.57(.47)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IRT (G1)</td>
<td>.51(.46)</td>
<td>.62(.42)</td>
<td>.64(.47)</td>
<td>.88(.59)</td>
<td>.66(.51)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IRT (G2)</td>
<td>.55(.47)</td>
<td>.59(.50)</td>
<td>.63(.45)</td>
<td>.86(.62)</td>
<td>.63(.51)</td>
<td></td>
</tr>
</tbody>
</table>

CROSS-VALIDATIONS

|                      | POLY (G1)      | .74(.56)| .66(.47)| .52(.40)| .47(.38)| .52(.42)| .58(.46) |
|                      | POLY (G2)      | .79(.56)| .57(.44)| .55(.43)| .48(.39)| .48(.40)| .57(.46) |
|                      | IRT (G1)       | .49(.43)| .52(.46)| .60(.44)| .61(.47)| .88(.57)| .62(.50) |
|                      | IRT (G2)       | .57(.47)| .56(.48)| .59(.48)| .66(.45)| .90(.63)| .65(.52) |

1250 items on number-right and IRT due to program limitations

For number-right scoring and polyweighting, score differences decrease at approximately the same rate resulting in the highest precision evident for the fourth and fifth ordered score groups, between which there was no discernible difference. In other words, number-right scoring and polyweighting were most precise above the 60th percentile rank and the least precise in the first ordered score group or below the 20th percentile rank for both samples and cross-validations. For IRT ability estimation the opposite was true. IRT precision was most noticeable in the first and second ordered score group, i.e., below the
Average Absolute Differences
For Three Scoring Methods

![Graph showing average absolute differences for three scoring methods: Number Right, POLY, and IRT. The x-axis represents percentiles from 1-19% to 80-99%, and the y-axis ranges from 0 to 1. The graph illustrates the differences across different percentiles.]
40th percentile rank. IRT ability estimation resulted in score differences that generally increased over the five ordered score groups, however, the fifth ordered score group showed a marked increase for both samples and cross-validations.

Number-right scoring and polyweighting were better predictors of the criterion than IRT ability estimation, averaging .735 and .72 across both groups as compared to .665, respectively. Linear combination item sets scored with IRT ability estimation would have to be lengthened 62% to match the validity achieved through polyweighting and 89% to match that of number-right validity. Cross-validation correlations were consistent with the original sample correlations.

Stand-alone correlations for both polyweighting and IRT were larger than item set correlations. Stand-alone correlations with the criterion were .76 for polyweighting and .71 for IRT as compared to polyweighting item set correlations of .71 and .66 for IRT item set correlations with the criterion.

Correlations among the scoring methods on the criterion were very high, approximately .98. On average, correlations among the three item sets were .321 for polyweighting, .234 for IRT ability estimation, and .294 for number-right scoring.
CHAPTER 5

DISCUSSION

Several writers have raised caution concerning the treatment of context-dependent item sets in estimating reliability (Anastasi, 1961; Guilford, 1936; Kelley, 1924; Thorndike 1951). Additionally, previous research has concluded that treating context-dependent item sets as stand-alone items rather than as a set results in an inflated reliability estimate due to correlated errors or local dependence (Thissen, Steinberg and Mooney, 1989; Sireci, Thissen and Wainer, 1991; Yen, 1992). The present study was designed to investigate the degree of inflated reliability among three very different scoring models, number-right, polyweighting and three-parameter IRT. This study, in evaluating these three different scoring methods, found inflated reliability only for the polytomous scoring procedure, Sympson's polyweighting. Reliability estimates for number-right scoring and the three-parameter IRT model were very similar for stand-alone and item set treatment, in contrast to previous research. In fact, for number-right scoring and cross-validations for IRT, item set alpha reliabilities were slightly higher. Previous research has
also indicated that treating context-dependent item sets as stand-alones will also result in an underestimate of the standard error of measurement as information in the pattern of item responses is lost (Yen, 1992). Again, only the results for polyweighting were consistent with previous research.

This study extended previous research by comparing three different scoring methods rather than by using only IRT models to analyze the different treatment of context-dependent item sets.

When treating context-dependent item sets as separate item sets, number-right scoring yielded a slightly lower coefficient alpha than did polyweighting. However, number-right scoring had a greater standard error of measurement than polyweighting. Polyweighting resulted in a considerably higher coefficient alpha and lower standard error of measurement than IRT ability estimation. What was surprising was that number-right scoring also yielded a higher coefficient alpha and lower standard error of measurement than did IRT ability estimation.

Consistent with previous research (Lord, 1968; Yen, 1983, 1984), IRT ability estimation was more precise in the lower ordered score groups, i.e., below the 40th percentile, than number-right scoring, as evidenced by the smaller average absolute differences. As the three-parameter IRT model compensates for guessing by making allowance for the
possibility that low-ability examinees may guess, scores are therefore more stable in the lower end of the distribution where guessing is most prevalent. In other words, IRT item scoring weights at low ability levels take into consideration the general discriminating power of the item and the amount of random guessing present at these low ability levels. This is in contrast to number-right scoring and polyweighting, where guessing is not taken into consideration, resulting in less precision in the lower ordered score groups. As expected, there was a large difference in precision in the first ordered score group between IRT, polyweighting, and number-right scoring; but the greatest difference was in the fifth or highest ordered score group where polyweighting and number-right scoring were most precise. Number-right scoring proficiency in the fifth ordered score group was consistent with Lord's (1968) conclusion that number-right scores were more efficient where little random guessing occurs. The finding of polyweighting to be less precise in the lower half of the test score distribution was congruent with previous research (Crehan and Haladyna, submitted for publication). However, generalizations of these results are limited to the sample characteristics and test data used in this study. In particular, a limitation on generalizability is indicated by the average item difficulty of .75. Findings may differ for a test with different characteristics, e.g., higher or lower
average item difficulty or higher or lower test score variability.

The concurrent criterion-related validity of the items scored as stand-alones, with polyweighting and IRT ability estimation, was higher than that of item set correlations. This is in contrast to a previous conclusion (Thissen, Steinberg and Mooney, 1989) where the item set approach was higher than the stand-alone correlation, although only slightly, .01. Among the three scoring procedures evaluated, number-right scoring performed only slightly better than polyweighting in the correlative analysis. Polyweighting performed fairly well compared with number-right scoring, as the criterion was similar to the predictor. Both polyweighting and number-right scoring were better predictors of the criterion than was IRT ability estimation.

Previously, the three-parameter IRT model was found to fit well to complex measures of cognitive skills such as context-dependent item sets. Unfortunately, this was not the case in the present study. Polyweighting and number-right scoring outperformed IRT ability estimation on every criterion observed.

This study, therefore, raises some questions and directions for future research on the statistical analysis of context-dependent item sets. Primary among these questions is selection of the most appropriate estimate of
reliability. Coefficient alpha was used in this study since it was readily estimated for all three scoring methods. However, Sireci, Thissen and Wainer (1991), using IRT models, calculated marginal reliability, which is computed from the information function and describes the reliability of estimates of theta. Secondly, more research is needed to determine which scoring technique maximizes estimated reliability. Replication of this study for score distributions with different characteristics would be helpful. Thirdly, do scoring techniques differ on important validity criteria? Thus, these issues of reliability and validity need to be further explored with new "classical" polytomous and IRT polytomous models and computer programs that have been introduced recently (Thissen, Steinberg, and Fitzpatrick, 1989).

The current philosophy on test development endorses tests comprised of larger tasks, such as context-dependent item sets, which better emulate real world tasks. Most licensing and certification testing programs have developed tests which utilize context-dependent item sets, since this item type presents tasks which are more commensurate with job descriptions than are stand-alone items. Since important decisions are made via testing programs, it is imperative to employ test analysis and scoring procedures which maximize the precision of measurement around the point in the distribution where the decision is made, i.e., the
passing score. It seems clear that polytomous scoring models have much promise in this regard. However, the research on these scoring models as applied to tests employing context-dependent item sets is scant. As the awareness of the potential for improved measurement through the use of polytomous scoring increases, the knowledge base is sure to broaden.
Appendix A

Read the following comments a mother made about day care. Then answer the questions that follow by circling the letter of the best answer.

"Children go to day care out of necessity, not preference. Day care cannot take the place of a mother's love, attention and affection. Day care can only meet children's basic needs such as food, shelter and comfort. And these needs are met at the day care workers discretion. Consequently, there is the potential for neglect. Children's self-esteem would benefit if mothers could stay home with their children."

1. Which one of the following unstated assumptions is this mother making?
   A. Children go to day care out of necessity.
   B. Day care workers play with children.
   *C. Day care makes no contribution to children's self-esteem.
   D. Day care cannot take the place of a mother's love, attention and affection.

2. Which one of the following child care providers is this mother talking about?
   A. Nanny or Au pair
   B. Housekeeper
   C. Familial babysitter
   *D. Commercial day care centers

3. Which one of the following statements is most essential to the final conclusion?
   *A. A mother's love, attention and affection is beneficial to a child's self-esteem.
   B. Day care cannot take the place of a mother.
   C. Day care usually meets children's basic needs.
   D. Children do not go to day care out of preference.
Bibliography


Crehan, K. D. & Haladyna, T. M. Submitted for publication. A Comparison of Several Linear Polychotomous Scoring Methods.


