Differential Item Functioning on The Myers-Briggs Type Indicator

by

Stuart Elliott Greenberg

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

R. J. Harvey

Roseanne Foti

Neil Hauenstein

Sigrid Gustafson

Steve Markham

October, 1992 [1993]

Blacksburg, VA

ABSTRACT

Differential Item Functioning
on The Myers-Briggs Type Indicator

by

Stuart Elliott Greenberg

Committee: R. J. Harvey (Chair), Roseanne Foti, Sigrid Gustafson, Neil Hauenstein, Steve Markham

(abstract)

Differential item functioning on the Myers-Briggs Type Indicator (MBTI) was examined in regard to gender. The Myers-Briggs has a differential scoring system for males and females on its thinking/feeling subscale. This scoring system preserves the 60 % thinking male and 30 % thinking female proportion that is implied by the Jungian theory underlying the Indicator. The MBTI's authors contended that the sex-based differential scoring system corrects items that subjects at a certain level of a latent trait either incorrectly endorse or leave blank.

This reasoning is the classical definition of differential item functioning (DIF); consequently, the non differentially scored items should exhibit DIF. If these items do not show DIF, then there would be no reason to use a differential scoring system.

Although the Indicator has been in use for several decades, no rigorous item response theory (IRT) item-level analysis of the Indicator has been undertaken. IRT analysis

allows for mean differences in subgroups to occur, independent of the question of DIF. Linn and Harnisch's (1981) pseudo-IRT analysis was chosen to test for the presence of DIF in the MBTI items because it is best for tests of relatively small length. The Myers-Briggs subscales range from 22 to 26 items, which is relatively small by IRT standards.

IRT analyses conducted on N=1887 subjects indicated that no items on the thinking/feeling subscale showed evidence of DIF. Out of 94 items, only one extraversion/introversion item and one judging/perception item showed evidence of DIF; no Thinking/Feeling items showed DIF. It is recommended that sex-based differential MBTI scoring be abandoned, and that the distribution of type in the population be examined in future studies.

Acknowledgments

The document you now hold in your hands could not have been completed without the support of many people. The first person I must thank is R.J. Harvey, who is my advisor and friend. His help was eagerly sought and even more eagerly given. The second person I must thank is Neil Hauenstein, who is my friend and advisor. His door was always open, even when it was closed. I also must thank my outstanding committee members, past and present. I wish to thank Roseanne Foti, who I am still afraid of, but in a good way. I also must thank Sigrid Gustafson for her wit, wisdom and unbelievably understanding nature. Dr. Steve Markham helped me by always asking the question I dreaded the most, a talent shared by Dr. Lee Wolfle. Dr. Joe Sgro was of invaluable help, by reminding me that not too long ago people did their analysis by hand. I also wish to thank Danny Axsom who not only helped me with my research, but with my teaching.

I would be remiss if I did not thank many of my fellow graduate students. My class will not be soon forgotten at Virginia Tech. I was Dr. Bob Brill's shadow throughout all of our courses and I learned much from him. Marta Carter was not a class mate, but became a true friend and confidant. Dean Stamoulis was a steady friend and even

steadier source of feedback. Dean also was an excellent entertainment safety facilitator. I also must thank many other students for their friendship and support: Bill Murry, Major Justin Rueb, USAF, Dr. Monnie Bittle, Dr. Lance Becker, Carissa Luch, Mark Cowgill, Sharon Flinder, Bruce Wayne, Darren Ritzer, Dwayne Norris, Maureen Walsh, Amy Russell and the conglomerate entity that appeared whenever the last four individuals were in a group.

There are also other individuals who provided much needed support during my graduate education. Gayle Kennedy, who was much more than a secretary, was like a mom away from home. Jonathan Thayer always was willing to listen. Dr. Jeff Klawsky, Dr. Bill Byrne and Dr. Dennis Armstrong from United Airlines taught me the importance of details.

Of course, the biggest support came from my family. My mother and father, Marsha and Ed Greenberg, provided unceasing support, guidance and motivation. This document would have not been possible without them. My sister, Aileen, and my brother-in-law, Jeff Postorino, would always take the time to cheer me up. Lastly, I would like to thank my Grandma Jeanette who always knew I could do this and anything else I wanted to do.

Table of Contents

List of Tables

List of Figures

Differential Item Functioning on The Myers-Briggs Type Indicator

Overview

The Myers-Briggs Type Indicator (MBTI) is a personality scale that is several decades old. Although there is evidence that males and females may respond differentially to the Thinking-Feeling subscale of the indicator (i.e., the Thinking-Feeling subscale is differently scored for males and females), no rigorous differential item functioning studies have been undertaken. The test constructors used measures of item popularity and prediction ratios to examine gender effects and determine the differential scoring system. Much of the primary research on gender was based on inferences about Jungian theory made by the original researchers. Other researchers examined gender issues at the level of analysis of type, not at the item level.

The present study has examined these gender issues on an item level. Because the Myers Briggs Type Indicator is widely used in counseling, academic, and industry settings, and because there is strong evidence of differential item functioning for males versus females on at least one of the four scales, it was necessary to examine the possible psychometric deficiencies of this instrument.

There are several methods of determining if an item is biased. Classical test theory methods include distractor

analysis, the ANOVA approach, the modified delta plot, and the chi squared method; item response theory methods include Lord's chi squared method and Raju's area under the characteristic curve method. IRT methods require the metric linking of parameter estimates. There are two types of linking methods: non iterative and iterative. Non iterative methods (Divgi, 1985) link the estimates in one step. However, biased items in an item pool may violate the assumption of unidimensionality underlying IRT models. One way to combat this potential threat lies in methods using iterative linking. Iterative methods include Lord's purification method, Divgi's iterative linking method, the modified Lord purification method, and the iterative parameter linking and theta scale purification method (Park & Lautenschlager, 1990).

Item response theory is the preferred differential item functioning detection technique because item and person parameters are invariant; that is, they are not dependent on the sample in which they are calibrated. However, item response theory methods require a large number of items and a very large sample size in order to provide stable parameter estimates. Of course, stable parameter estimates are necessary for applications such as item bias research (Hulin, Drasgow & Parsons, 1983). When one has small sample sizes or a small number of items, approximations to IRT that

take these factors into account have been suggested (Linn & Harnisch, 1981; Hambleton, Swaminathan & Rogers, 1991). The pseudo-IRT method has been shown to the most accurate of these approximations (Shepard, Camilli & Williams, 1985).

The pseudo-IRT method was used to investigate the differential functioning of items on the MBTI in terms of males versus females. This method was used to examine the items that have been shown to function differently for males and females in the scoring system developed by the constructors of the MBTI (i.e., the Thinking-feeling (TF) scale); in addition, the three remaining scales that are not scored differentially for males versus females were also examined to test for sex based differential item functioning (DIF). Based on the predictions of the MBTI's authors, it was hypothesized that many of the items in the TF scale would demonstrate significant DIF; conversely, if the MBTI's authors were correct, the items in the remaining three scales would not demonstrate significant DIF. However, only two items were found that exhibited DIF. These items were on the extraversion-introversion (EI) and judging-perception (JP) subscales. This lack of item level evidence for differential item functioning between males and females suggests that the differential scoring system based on gender for the Thinking-Feeling subscale may not be necessary.

The Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator is used in counselling, academic, and industry settings. The MBTI has been developed over a period of several decades; in the context of the current study, the indicator will be described in terms of the overall stages of its development as well as in terms of specific gender issues.

Development of the Indicator

The Myers-Briggs Type indicator is a personality questionnaire that was developed to test Jung's theory of psychological types and to present the information in a usable form (McCaulley, 1990). Jung's typology has four major concepts, and all of these were incorporated into the type indicator. One of the basic concepts is Extraversion (E) or Introversion (I): The extraverted individual is predicted to seek out the external world, while the introvert is more concerned with the inner workings of their mind. A second concept is Sensing perception (S) or Intuitive perception (N): The sensing individual is hypothesized to be more concerned with real, observable events (McCaulley, 1990, p. 382), whereas the intuitive individual is seen to be more focused on symbolic meanings and what may occur in the future. The third concept is Thinking (T) or Feeling (F): Thinking individuals are aid to use logical examination of causal relationships to make

rational decisions, whereas feeling individuals are supposed to examine the various alternative solutions before rationally deciding on the final solution (McCaulley, 1990, p. 382). Thinking types prevail among males and feeling types prevail among females (Hall & Nordby, 1973). The final concept in Jung's typology, and in the Myers-Briggs Type Indicator is that of Judgment (J) or Perception (P): Individuals who rely on judgment typically organize and plan events, while making quick decisions (McCaulley, 1990, p. 382); in contrast, individuals who rely on perception are seen to view the world as an ever-changing place and take advantage of opportunities as they appear (McCaulley, 1990, p. 382). This last concept was touched upon by Jung, but it was more fully developed by the MBTI authors.

Each person is predicted to prefer one member of each pair. Based on their subscale scores, examines are assigned to one of 16 "type" groups. The MBTI authors developed descriptions that fit each of the sixteen types. (McCaulley, 1990, p. 383).

The Myers-Briggs Type Indicator was developed over a period of several decades. Myers and Briggs were followers of Jung and examined his work for the most important indicators of type. They researched Jungian theory and observed individuals with regard to their type differences for two decades before starting work on the actual MBTI.

There were three assumptions underlying the development of the Indicator: a) preferences for Jungian types actually exist, and measurement of these preferences should be possible for individuals with either strong or weak preferences for the various types (Myers & McCaulley, 1985, p. 140); b) a self report questionnaire can adequately measure these preferences; and c) these preferences exist as dichotomies and each pole of the dichotomy as important as the other (Myers & McCaulley, 1985, p. 140).

This third assumption (equality of the poles) required that Myers and Briggs construct items that did not have alternatives that were split into socially desirable / non socially desirable dichotomies. Items that were not overly extreme to either pole were constructed; extreme items would not be helpful in discriminating between the two poles. Items concerned everyday behaviors: instead of directly asking the examinees about their preferences, these preferences were inferred by requiring a choice between these everyday behaviors. This inferential process was used because it was feared that self-report of actual preferences would not be accurate (Myers & McCaulley, 1985, p. 141).

The item format used in the Myers-Briggs Type Indicator is forced choice. The questions were first tested on associates of the researchers; the types of these associates were previously determined by the MBTI's authors

through years of observation. This process took place from 1942 through 1944, and produced Forms A and B of the MBTI. These forms consisted of the same items, the only difference being the ordering of items. Items that did not receive a response by 60% or more examinees of the appropriate type were eliminated (Myers & McCaulley, 1985, p. 142). Examination of the responses to these items demonstrated that different responses were more or less popular; a differential weighting procedure was developed to compensate for these popularity differences. The weighting system was also incorporated to allow for "better differentiation of individuals scoring near zero (Myers & McCaulley, 1985, p. 141).

For Form C, the researchers developed a prediction ratio to determine which items were more popular. For example, the prediction ratio for response A on item 1 on the E scale is given as the percentage of E subjects giving the A response on item E1 divided by the percentage of E subjects with the A response on item E1 plus the percentage of I subjects giving the A response on the E1 item. This ratio includes only cases in which the scale difference between the two poles is 2 points or greater (Myers & McCaulley, 1985, p. 146), and indicates how well the response to that item fit the other item responses for that preference. Item responses that have a prediction ratio above .62 for the

pole in question are given a weight of 1; item responses that have a prediction ratio above .72 for the pole in question, as well as having a prediction ratio for the other pole below .50, are given a weight of 2 (Myers & McCaulley, 1985, p. 146-148).

Also on Form C, questions were eliminated that demonstrated high item total correlations with multiple scales. Items were eliminated that had low prediction ratios, and one new item was added. Several items with unclear wording were revised (Myers & McCaulley, 1985, p. 142).

During 1956 through 1958, Form D was developed. All of the Form C questions that passed the previous screening process remained, and new items that had gone through a similar procedure as the original items were added. These new items consisted of word pairs: the researchers had noticed that word pairs tended to work quite well (Myers & McCaulley, 1985, p. 143). This was the first time that children, undergraduates and graduate students were used as validation samples.

The old and new items were put into Forms E and F; Form F also contains unscored experimental items. Additionally, preference scores were used in these forms instead of percentages to determine the strength of preferences (Myers & McCaulley, 1985, p. 144). Form F was published by the

Educational Testing Service in 1962 and was used in large scale data collections. Form F became the standard form of the MBTI in the early 1970's (Myers & McCaulley, 1985, p. 144).

In the mid 1970's, a new standardization program was initiated. Item test correlation and prediction ratio results were found to be similar to the original standardization sample for adults as well as younger examinees. The standardization prompted new scoring weights for the TF scale, and also led to the development of Form G. Form G eliminated 38 of the non scored research items; one non-research item was added and two non-research items were dropped. A few ambiguous items were revised. Scored items were placed first in the sequence of items, ranked in terms of their predictive power. Form F (1977) and Form G are deemed "essentially interchangeable" by Myers and McCaulley (p. 144, 1985). A self scoring, shorter version of the MBTI (form AV) was also developed.

In summary, the development of the Myers Briggs Type Indicator has unfolded over many years. It is firmly based in Jungian personality theory. The establishment of the questionnaire assumed that preferences for the Jungian types actually existed, these preferences could be measured by self report measures and that these preferences existed as dichotomies. Although several forms of the MBTI have been

developed, and each form was a revision of a previous form, it is somewhat surprising to note that no rigorous psychometric item analyses have been performed. In effect, the analysis methods that were used relied solely on the inferences and intuitions of the original MBTI researchers.

Gender Issues and The Myers-Briggs Type Indicator

Several researchers have explored possible gender effects on the Myers Briggs Type Indicator. The research, in general, has been at the subscale or type level; no gender research beyond the initial development of the indicator has been at the item level. This section will review the research in this area.

Carskadon (1977) examined test-retest reliability for males and females on Form F of the MBTI. Continuous TF scale test-retest reliabilities differed for males and females, but were not statistically significant. On Form G of the MBTI (Carskadon, 1979), a difference in reliabilities was found on the continuous TF scales (.48 for 32 males and .87 for 24 females, $p<.001$). In a replication, Carskadon (1982) again found a difference on continuous TF scores between males (.91, n=24) and females (.56, n=24) on Form G.

Padgett, Cook, Nunley and Carskadon (1982) addressed the relationship between androgyny and the 16 types determined by MBTI. Guzie and Guzie (1984) proposed that there are masculine and feminine archetypes that complement the 16

types derived from the MBTI. However, Guzie and Guzie only addressed the relationship between types and archetypes, and did not explore these issues on an item level.

Myers and McCaulley (1985) reviewed the gender research on the MBTI, reporting the results of several item analysis studies. These item analyses used prediction ratios and item popularity measures for males versus females. At first, different scoring systems were used for males and females on all scales. By Form F, the different scoring systems were only used for the TF scale. It was believed that the differential responses on the TF scale did not mirror true preference differences, but instead mirrored social desirability issues. Apparently, it was not socially desirable for T females to report their T preference, so they marked an F preference instead (Myers & McCaulley, 1985, p. 148). In the 1970's restandardization procedure, Feeling responses were more popular and reweighted to have a smaller weight; Thinking responses were less popular in this standardization, and weightings for these responses were increased. The primary concern behind this reweighting appears to have been to maintain the appropriate proportion of Thinking and Feeling types in the standardization sample.

Stokes (1987a) surveyed a variety of experts in the fields of type research and gender research concerning the experts' opinions about the match between gender and type;

the level of analysis remained on the type level. Stokes (1987b) reviewed several empirical studies of gender and type at the type level, and also examined the "gender nuances" in the instrument and scoring (p. 47). Stokes noted the different weightings on the TF scale for males and females, and also noted that less strength of preference (on a 1 to 9 scale) is given to the TF scale than the other scales (p. 47). Specifically, males tend to prefer F less and females tend to prefer T less when examined in relation to the other scales. Stokes noted that the MBTI manual explains this phenomenon by claiming that these small preferences may be due to the differential social pressure placed on males and females (p. 47).

Stokes also reported on the consistency information given in the MBTI manual. The MBTI manual (Myers & McCaulley, 1985) noted that on Form F, females demonstrated more disagreement between word pairs and phrase questions for the JP scale than do males. On Form G, females have less agreement both on the JP scale and the TF scale (p. 61). Stokes also reported information on type and gender frequencies in the general population.

Stokes concluded by suggesting that more research on the MBTI and gender is needed, calling particular attention to the need for research on the TF preferences. However, she cautioned that these irregularities may be due to cultural

confusion about gender, and therefore may be impossible to erase (p. 49).

Harris and Carskadon (1988) examined the old and new scoring weights for the MBTI TF scale, noting that the weights were changed to make the distribution of types in the general population match the distribution expected by the test developers. The distribution of types resulted in too many male Feeling types and too few male Thinking types (p. 54). After the MBTI was administered to 645 college students, the types obtained were compared to self reports of type by the students; results indicated that although the old and new scoring systems worked similarly for females, for males the old scoring system was more accurate at predicting the subjects' self report of type than the new scoring system. This is another example of study of gender and the MBTI at the type level - as opposed to the item level of analysis.

In conclusion, there have been several studies of gender and the MBTI; however, none have been at the item level. Of perhaps the greatest importance to this study, Harris and Carskadon (1988) compared old and new scoring weights on the MBTI, noting that the change in the scoring system was made because the distribution of types in standardization samples had changed in ways that the test author believed to be incorrect. They found the new system to be biased against

males. Unfortunately, their study was at the type level, so information about differential item functioning was not available. The current study has addressed the need for an item level analysis of bias.

Philosophical Issues in Differential Item Functioning

It has been established that rigorous item analysis of the MBTI in terms of gender has not been performed. The current concern is the choice of the appropriate item bias detection technique. The choice of an item bias strategy is a complex issue, as the different techniques all have advantages and disadvantages. A careful review of this area is necessary to make the most appropriate choice.

In terms of a general definition, Osterlind (1983) stated that a test item is unbiased "when the probability for success on the item is the same for equally able examinees of the same population regardless of their subgroup membership (p. 11)." When this probability of success differs for equally able members of the same population due to subgroup membership, the item is deemed to be biased. This concept of equally able examinees of the same population is of vital concern. Differences in mean performances on a given test does not necessarily indicate that any individual test item is biased. There may very well be differences in the distribution of the latent trait in each subpopulation.

There is evidence for bias when the level of the latent trait is held constant across subgroups and those subgroups respond differentially to the given item.

Bias is a separate issue from test fairness: Bias can be thought of in terms of the psychometric properties of the test itself, whereas fairness issues arise out of the use of the test for a specific purpose. Referring to the conceptual model in Figure 1, the issue of bias can be viewed from "society's" point of view or from "science's" point of view. "Society's" assessment of test bias is best determined by expert judges. These judges, who may be involved with any or all stages of the testing process, can determine if the items suffer from any potential causes of bias; for example, stereotypical representation of minority groups, or unfair familiarity with subject matter by any one subgroup (Tittle, 1982, p. 31). The main benefit of using these expert judgments is the increased perception that the testing instrument in question is fair. This increased perception relies on the dissemination of the results of the judgmental review. As has been stated, there is a difference between fairness and bias, so to study bias, the viewpoint of "science" will be utilized.

If the definition of bias is viewed from "science's" point of view, bias judgments can be assessed either before the test instrument has been developed, or after the

instrument has been developed. If bias is to be investigated before test construction, the appropriate research medium is experimentation. Scheimeiser (1982) gives three reasons for the use of the experimental approach to studying bias. The first is that this approach can provide a systematic plan for examining the assumptions and procedures that are inherent in the test construction process; one can determine if these methods produce differential treatment of subgroups, or if they can be changed to correct for any bias against subgroups (p. 66). The second reason is that the experimental approach can supplement the expert judgment procedure and the after-the-fact (or statistical) method of bias detection. The experimental approach can supplement the statistical approach by providing possible explanations as to the reasons why a particular item is seen to be biased. The third reason for the use of the experimental approach is that it has all the benefits of experimental research, including random assignment of subjects and control of extraneous variables. An example of the experimental approach to the study of bias can be found in a study of how different culturally specific reading tasks might have a differential impact on test examinees of different cultures (Scheimeiser, 1982).

The experimental procedure, and to a lesser extent,

expert judgment, are most suited to studying bias "before the fact." After the test has been developed and administered to examinees, post hoc methods of bias detection must be used. These post hoc methods can be divided into three broad categories: a) techniques that examine the whole test, b) the related technique of factor analysis, and c) techniques that study item-level responses. Techniques that use the whole test also require some form of external criterion: in this case, the test is used to predict the external criterion for each of the subgroups of interest (Humphreys, 1986, p. 327) in a regression model. The groups are compared in terms of variance of errors of prediction, slopes of regression lines, and intercepts of regression lines (Humphreys, 1986, p. 327). There will be bias to the degree that there are differences in any one of these comparisons, but intercept differences are most likely. Factor analysis can also be used to examine item bias at the test level: If there is bias against any of the subgroups, the factor structure across subgroups would be expected to differ.

In conclusion, an item can be defined as biased when subgroup membership becomes a determining factor in the correct or incorrect response to a particular item by equally able subgroup members. Bias, which concerns the psychometric properties of a test, is different than test

fairness, which is concerned with the use of the test. Figure 1 summarizes the conceptualization of bias in testing. Bias can be decided by society through the means of expert judges; it can also be defined by science. If the route of science is taken, before the test is developed, experimentation can explore bias issues; however, most of the research options occur after the test is developed. If the whole test is the level of analysis and an external criterion is available, the regression model can be used to examine bias issues; factor analysis can also examine factor structure differences across various subgroups. However, bias research is not restricted to the test level of analysis, and there are a wide variety of methods to determine if a specific item is biased against a particular subgroup. Item bias methods typically use internal criteria to determine if bias exists. There are two approaches to this type of item bias research: the classical test theory approach and the item response theory approach. Each will be discussed in turn.

Classical Test Theory Methods

Distractor Analysis In distractor analysis, the examination focuses on the incorrect responses to an item. The premise is that if an item is biased, there will be differential preference for the various alternatives based on the examinee's subgroup membership (Osterlind, 1983).

Distractors are usually chosen to be plausible alternative responses. Research has shown that it is highly unlikely that distractor choice is determined purely by guessing (Lord, 1980). Distractor analysis is a relatively straightforward procedure. Once it has been determined that subgroups differentially respond to an item, the frequency of the chosen alternatives is examined. Distractor analysis is reserved for items in which there are polychotomous responses. Because the questionnaire of interest in the present study is a dichotomously scored device, distractor analysis is inappropriate in this circumstance.

ANOVA The ANOVA and related methods deal with dichotomously scored items, and could be used in the present circumstance. However, it will be shown that, for various reasons, most are inappropriate for the current research needs.

In the ANOVA approach, a test is administered to two or more groups and the interaction between group membership and scores on the individual items is examined. The groups are assumed to come from the same population.

The procedure follows the following format. For all items in all groups, a *p* value is computed. The *p* value is the percentage of individuals who correctly respond to the given item (Osterlind, 1983, p. 21). In a two group scenario, *p* values for one of the groups are oriented along the x axis of a graph, while the *p* values for the other

group are oriented along the y axis (Osterlind, 1983, p. 22). If a straight line (45 degrees, starting at the origin point) is drawn on this graph, p values that are equal for both groups will fall on this line. If this is not the case, the line will be a poor estimator of the data points, and the p values will differ between the groups (Osterlind, 1983, p. 22).

To test this theoretical relationship, an ANOVA procedure is run in which the p value is the dependent variable and group membership and items are the independent variables. If the ANOVA is not significant, it is assumed that there is no bias in the items. If there is a significant relationship between the two independent variables, bias is said to exist. Post hoc procedures may be used to determine exactly which items are biased.

The main problem with this method of item bias is that multiple sources of variance may cause the interaction to be significant. Differences in overall group performance or ability level differences between the groups may cause the interaction effect. This is tied to the assumption that both groups come from the same population which is a tenuous assumption at best (Park & Lautenschlager, 1990).

Modified Delta Plot (Transformed Item Difficulties) Bias is defined by this method when an item is more difficult for one subgroup then for another (Osterlind, 1983, p. 28). The

advantages for this method are that it is relatively easy to compute and that it allows for a graphical representation of bias that can easily be explained to an audience without a statistical background.

This method is concerned with differences in the relative difficulties of items. Item difficulties (p values) are computed for each item for each of two subgroups. These p values are transformed into delta (a standardized score). Delta scores allow the researcher to express the p values of two groups on the same scale with a common mean and standard deviation (Osterlind, 1983, p. 32). After the p values have been converted to z scores, delta is computed by multiplying each z score by 4 and then adding 13. This new distribution has a mean of 13 and a standard deviation of 4. This new distribution has the benefit of having no negative values and having a constant value of the standard error at all levels of item difficulty (Osterlind, 1983, p. 33).

After these delta scores have been obtained, the means and standard deviations for each group (x and y) are obtained. The correlation between delta values for each item are then obtained. This information is used to compute the slope and intercept of the plot of one group's delta values vs. the other group's delta values. Osterlind (1983) reports the necessary calculations for these values:

$$b = \frac{(\sigma_y^2 - \sigma_x^2) \pm \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4r_{xy}^2\sigma_x^2\sigma_y^2}}{2r_{xy}\sigma_x\sigma_y}$$

$$a = M_x - bM_y$$

The distance from this line determines whether or not the item is biased. The distance function is given as:

$$D_i = \frac{bX_i + a - Y_i}{\sqrt{b^2 + 1}}$$

The b and a values are given above and the X and Y values are the two groups' delta values for a given item. The final question is how big the distance function must be for an item to be considered biased. Typically, a computed confidence interval is placed around the major axis line. Distance functions outside this range are classified as biased.

There are several difficulties with the delta plot approach. One is that when a test has items of various levels of difficulty (a common scenario), there may be a groups X items interaction in a test that has no biased

items. However, this approach defines the groups X items interaction as a indication of bias. The second difficulty with this approach is that it assumes that both subgroups come from the same population. In such a case, a common major axis line can be used for both groups. However, this assumption may not be correct. The two sub groups may come from different populations and require two different curves to fit the data. The final difficulty with this approach is that the item difficulties that are obtained --even in their transformed state-- are dependent on the particular sample in which they were obtained. For these reasons, this approach is not recommended for the current study.

Chi Square The chi-square item bias method examines the probability of examinees, at the same ability level, from two or more groups answering an item correctly (Osterlind, 1989, p. 38). The benefit of this method is that it incorporates the idea of ability level of examinee into its computation of item bias. This approach divides the total test score into intervals that correspond to different ability levels. The major question is whether the distribution of obtained scores within each interval is similar to the distribution of expected scores.

The chi square approach treats each item individually, in terms of the different ability levels. Ability level is determined by splitting the total score into intervals. The

number of intervals is determined by examining the frequency of scores at each score level. Typically, three to five intervals are used. It is assumed that the overall test score is a valid indicator of ability. An item characteristic curve (ICC) can be graphed for each item. This is one way of organizing the data (Osterlind, 1983, p. 39). ICCs belonging to different subgroups can then be compared to determine if bias is present. Bias is present when ICCs are not identical. Standard procedure requires the use of a goodness-of-fit test instead of a visual comparison of ICCs. The relative ease of application is one of the benefits of the chi square approach; a secondary benefit is that a large sample size is not necessary.

Scheuneman (1979) proposed the initial formulas for computing the chi square statistic. Scheuneman's statistic was an approximation to the chi squared statistic (as described by Baker, 1981). Marascuilo and Slaughter (1981) proposed six chi square statistics; these methods were tailored to the specific needs of various problems and differed in terms of the ease of computation. The benefits of these methods include ease of computation and ease in explaining results to persons unfamiliar with statistical jargon (Osterlind, 1983). The chi squared method approximates item response theory approaches to item bias research, but does not share in the benefits of the IRT

approach.

Overview of Item Response Theory Bias Methods

The main premise of item response theory (IRT) is that examinee performance on a particular test item can be predicted by traits (Hambleton et al., 1991, p. 7). A item characteristic curve can be developed that is a monotonically increasing function that describes the relationship between each examinee's performance on individual items and the underlying trait.

There are several benefits to using item response theory methods to determine whether items are biased or not. The main benefit is that item and person parameter estimates are invariant across differences between groups in mean trait levels (Hambleton et al., 1991, p. 8). Item parameters are not dependent on any particular sample, and estimates of examinee ability are not dependent on any particular collection of test items.

There are several item response models that are available for use; these models differ in the number of parameters used to describe the relationship between the latent trait and performance on the item. The most common item response models for dichotomous data are the one, two and three parameter logistic model.

The one-parameter model, also know as the Rasch model, explains the relationship between the underlying latent

trait and performance on the item with the b parameter. The b parameter is the item difficulty parameter. This parameter determines the location of the item characteristic curve on the latent trait scale. The b parameter can be defined as the location on the latent trait scale at which the probability of a correct response is 0.5 (Hambleton et al., 1991, p. 13).

The two-parameter logistic model uses the b parameter, as in the one parameter model, as well as the a parameter. The a parameter is the item discrimination parameter. While b determines the location of the ICC on the latent trait scale, a determines the slope of the ICC at that point. Items that are steep, which at the extreme would approach a step function, are better at determining the levels of the latent trait of particular examinees. ICCs that have less steep a parameters do not discriminate as well between examinees of different levels of the latent trait (Hambleton et al., 1991).

The three-parameter logistic model builds upon the two parameter model in that it uses both the b parameter as well as the a parameter. This model adds a third parameter (c) to explain the relationship between the underlying latent trait and performance on the item. In the previous models, it was assumed that the lower asymptote of the monotonically increasing function (the ICC) was at zero. The c parameter

allows for a possible nonzero lower asymptote (Hambleton et al. 1991, p. 17). The c parameter is the "pseudo chance parameter" (Hambleton et al., 1991, p. 17). and is used when there is a possibility that examinees with a low level of the latent trait will get the item correct; in fact, the c parameter represents the probability that examinees with low levels of the latent trait will answer the item correctly (Hambleton, et al., 1991, p. 17).

In the majority of IRT applications, the test in question is measuring some kind of ability. The MBTI is a personality test. There are four latent traits being measured: extraversion-introversion, sensing-intuition, thinking-feeling and judging-perception. Each subject has some true level of each of these latent traits. The major distinction between ability and personality tests is that in ability tests there is always a distinctly correct answer to every item. For example, in a mathematical ability test, there will be one correct answer for each item regardless of its difficulty. In a personality test, it would appear that no one test alternative is distinctly correct. However, in regards to the personality construct being measured, one item will be correct. Endorsement of that alternative will give credit towards the overall score on that personality measure.

The item parameters will function in a similar fashion

for personality tests as they do for ability tests. The b parameter will determine the location of the ICC on the latent trait (personality construct) scale. It can be defined as the point on the latent trait scale at which the probability of a correct response --*in regards to the personality construct in question*-- is 0.5. The a parameter determines how well the item discriminates between subjects at different levels of the latent trait (*the personality construct*). The c parameter acknowledges that even at the lowest level of the personality construct, the correct response --*for that particular personality construct*-- will be endorsed on some items. This may be due to the social desirability of certain responses. While for ability tests, the c parameter is considered a pseudo-guessing parameter, for personality tests it may be considered a "pseudo-social desirability parameter."

No matter which of the three logistic models is chosen, there must be an assessment of the fit between the model and the data. Hambleton and Swaminathan (1985) have suggested three types of evidence that support the fit between the model and the data: a) Correctness of the assumptions of model, b) the invariance of the item and ability parameters, and c) accuracy of the model predictions (Hambleton et al., 1991, p. 55). Regarding the first type of evidence, one of the major assumptions of all three item response theory

models is the unidimensional nature of the data. This can be checked through factor analysis. For the two parameter estimate, lack of guessing can be determined if the performance of examinees with low thetas on the most difficult items is at or approach zero (Hambleton et al. 1991, p. 57). Invariance of ability estimates can be checked by assessing the ability estimates of groups of items of different difficulties. Evidence for invariance occurs when the estimates do not differ greater than the measurement errors of the estimates. The chi squared statistic can also be used to test for the goodness of fit between the model and the data.

Lord's (1980) Chi Squared Method This method of detecting differential item functioning compares the item parameter estimates of two groups. If the parameters numerically are the same, it must also be the case that the item functions are the same, and thus no differential item functioning exists. The null hypothesis for this statistic is:

H0: $b1=b2$; $a1=a2$; $c1= c2$ (Hambleton et al., 1991)

Estimates of a, b, and c and the variance-covariance matrices of the estimates for each of the items are need to compute the chi squared statistic. The metrics of these estimates must first be linked. Several metric linking

techniques will be discussed in momentarily. After scale adjustment, the variance-covariance matrix is obtained for each group. This variance-covariance matrix is the inverse of the information matrix for each item. These two matrices are added to produce the variance covariance matrix of the differences between the estimates. The chi squared statistic is given as (Hambleton et al., 1991, p. 111):

$$\chi^2 = (a_{diff}\ b_{diff}\ c_{diff})'\ \Sigma^{-1}\ (a_{diff}\ b_{diff}\ c_{diff})$$

where

$$a_{diff} = a_2 - a_1 \qquad b_{diff} = b_2 - b_1 \qquad c_{diff} = c_2 - c_1$$

and Σ is the variance–covariance matrix of the differences between the parameter estimates.

In large samples this test statistic approximates a chi square distribution. It has p degrees of freedom; p is defined as the number of estimates that are being estimated in that particular model. In general, and in LOGIST (Wingersky et al., 1982) in particular, the test statistic is often not estimated well; additionally, use of the c parameter in the bias statistics may make the test too conservative and decrease its ability to detect bias (Hambleton et al., 1991, p. 112). One possible solution is to use only the a and b statistics in these analyses: observed differences in a and b would support the conclusion that the item is biased. Differences in c would also prompt the conclusion that the items is biased, but c is so unstable that such conclusions would not be justified.

Raju's (1987) Area Under the Item Characteristic Curve

Instead of comparing the item parameters that define the ICCs, the ICCs themselves can be compared. The premise is that if the ICC's for two groups were placed on common graph (which would be done by converting them to a common scale), a no-bias condition would be indicated by no physical difference between the ICCs. Bias would exist if there was nontrivial area between the ICCs for the two groups. The first attempts to determine this area were based on numerical procedures. This involved "a) dividing ability ranges into k intervals, b) constructing rectangles centered around the midpoint of each interval, c) obtaining the values of the ICC's (the probabilities) at the midpoint of each interval, d) taking the absolute value of the differences between the probabilities, and e) multiplying the difference by the interval width and summing" (Hambleton et al., p. 113, 1991). Raju (1988) developed a formula for computing the area between ICC's for one, two and three parameter models. The formula for the three parameter model is as follows:

$$\text{Area} = (1-c)\left| \frac{2(a_2-a_1)}{Da_1 a_2} \ln\left[1+e^{Da_1a_2(b_2-b_1)/(a_2-a_1)}\right] - (b_2-b_1)\right|$$

In this formula, c is assumed to be identical across the two groups. If c is different for the two groups, the area between the two curves is infinite over the whole range of ability (Hambleton et al., p. 114, 1991). Thus, if the c values are not identical, it is impossible to compute the significance test. Several methods have been proposed to combat this problem. One is to divide the area statistic by its standard error in order to approximate a standard distribution. Another method is to use "bootstrapping" techniques to determine the sampling fluctuations of the parameter estimates. The major problem with this method is that it requires the estimation of parameters in two or more groups. A small sample size will hinder this procedure (Hambleton et al., 1991, p. 115).

Metric Linking Methods

The parameter estimates obtained in item response theory are arbitrary in scale and origin (Divgi, 1985, p. 413). If parameters are estimated separately, there must be some linking technique to place the estimates from one calibration onto the same scale as the other calibration. There are several possible linking techniques (Warm, 1977; Linn, Levine, Hastings & Wardrop, 1981; Stocking & Lord, 1983; Divgi, 1985; Park & Lautenschlager, 1990; Hambleton et al., 1991). They can be categorized into one of two types: a) the non-iterative linking technique (i.e., the linking of

estimates is done in one pass); and b) the iterative linking technique; which has its origins in the purification technique for detecting item bias described by Lord (1980). In this latter technique, the linking is repeated several times, taking account of the items that are determined to be biased and non biased. The iterations continue until the items are that are classified as biased are consistent from one iteration to the next.

Divgi's (1985) Non Iterative Linking Method (DN) In this approach, one calibration is specified as the first calibration (denoted as 1) and the other is specified as the second calibration (denoted by 2). The transformed estimates are denoted as $\underline{a}^*_2$ (the estimate of discrimination) and $\underline{b}^*_2$ (the estimate of difficulty). The formulas for the transformed estimates are as follows:

$$a_2^* = a_2/A \quad ,$$
$$b_2^* = Ab_2 + B \quad ,$$

The initial value of the A transformation coefficient is obtained by dividing the mean value of the a parameter for the second calibration by the mean value of the a parameter for the first calibration. The B transformation coefficient is slightly more difficult to determine. First, the

variance-covariance matrix of sampling errors is obtained for each item in both calibrations. These are the variance-covariance matrices for the two groups. The diagonal elements of Sigma2 must also be transformed. The procedure is as follows:

For a given item, let Σ_1 and Σ_2 be the values of this matrix from the two calibrations. When a_2 and b_2 are transformed, so are the diagonal elements of Σ_2—the *aa* element is divided and the *bb* element multiplied by A^2. Denote this transformed matrix by Σ_2^*.

The transformed matrix is called Sigma*$_2$ (Divgi, 1985, p. 414).

The matrix <u>T</u> is defined as:

$$(\Sigma_1 + \Sigma_2^*)^{-1}$$

Coefficient <u>B</u> can be defined as

$$B = \Sigma[T_{aa}(a_1 - a_2/A) + T_{bb}(b_1 - Ab_2)]/\Sigma T_{bb}$$

The quadratic form (Q) is then determined:

$$(a_1 - a_2^* \quad b_1 - b_2^*)(\Sigma_1 + \Sigma_2^*)^{-1}(a_1 - a_2^* \quad b_1 - b_2^*)'$$

Values for A and B must be selected that minimize Q. One possible method for finding these values is focusing only on A. B is a function of A so this is permissible. After the initial value of A is obtained, a grid of values around A can be selected and a Q value can be obtained for each corresponding value of A. Two values of A can be identified between which the minimum value of Q falls. A finer grid of values of A between these two points can be selected and this process repeated until the desired level of precision has been reached (personal communication, Holtzman and Heise, January, 1992).

Lord's (1980) Purification Procedure Lord (1980) described a "purification of the test" (p. 220) technique that was originated by Marco; Lord suggested that this technique be used when the test consists of many biased items. The procedure is as follows: The total test is analyzed and items are removed that appear to be biased towards the groups of interest; the items that remain are to be considered a unidimensional pool. All groups are combined together and ability estimates are obtained. Fixing these ability estimates at the estimated values,

parameter estimates are computed for each item, separately in each subgroup. All items, even those previously dropped, are used in this step. The obtained item parameters are compared between groups by a method such as Lord's Chi Squared method.

Divgi's Iterative Linking Method (DI) (Park and Lautenschlager, 1990) This procedure is a modification of the iterative linking technique suggested by Drasgow (1987). However, Drasgow used the Stocking and Lord (1983) metric linking method, whereas this technique utilizes the Divgi (1985) method. The procedure is as follows: The Divgi (1985) linking method is used to link the initial parameter estimates, and biased items are classified through an item bias method. The items that are determined to be unbiased are used to link the estimates again. Once items are relinked, item bias methods are used to determine which, if any, items are biased. This procedure continues until the same set of items are identified as biased in two consecutive repetitions.

The Divgi method can be used in place of the Stocking and Lord (1983) method because although it is conceptually similar to the Stocking and Lord method, it is far easier to implement.

Modified Lord Test Purification Method (MLTP) Park and Lautenschlager (p. 165, 1990) describe this method as

follows: All groups are combined and ability estimates are obtained. The groups are separated, but the ability estimates are fixed at their initial estimated values. Then, item parameters are estimated for all individuals, and an appropriate bias statistic is computed. The items that are determined to be biased are removed. Using the unbiased items, the ability estimates are recomputed. The item parameters are then reestimated for each subgroup separately, while holding the ability estimates at the fixed value. Bias statistics are then computed for all items. These steps are repeated until the same set of items are identified as biased in two consecutive repetitions.

Iterative Parameter Linking and Theta Scale Purification (ILAP) (Park & Lautenschlager, 1990)

This procedure starts at the point at which the Divgi Iterative procedure ends. The groups are separated and the unbiased items are used to estimate the latent trait. Theta (the measure of the latent trait) is fixed at this estimated value and item parameters are reestimated. Divgi's linking coefficients (A and B) are estimated using only the unbiased items (as determined by the original Divgi Iterative procedure). Bias statistics are then computed and iterations continue until there is a convergence between iterations.

Pseudo-IRT

The pseudo-IRT method can be seen as a link between the classical test theory methods of determining item bias and the item response theories methods. The primary advantage of this approach is that it may be used when the test has a small number of items and/or subgroup sample size is relatively small. Linn and Harnisch (1981) developed this procedure. Hambleton et al. (1991) noted that this method can be used with small samples because parameter estimates are obtained from the combined majority and minority group. It is also useful when the test has a small number of items because the combined sample is used, prompting very stable estimates of parameters.

The pseudo-IRT differential item functioning technique has several steps, the first being to estimate the a, b and c item parameters as well as theta for each individual using all of the available subjects. The estimated probability that an individual (j) with a given theta score would respond correctly to a specific item (i) is obtained by:

$$P_{ij} = c_i + \frac{1 - c_i}{1 + \exp\left[-1.7a_i(\theta_j - b_i)\right]}$$

The estimated (or expected) probability of a correct response is compared to the observed proportion of correct

responses. If there is a large discrepancy between these two scores, this is an indication that the item exhibits differential item functioning.

This comparison is computed for all target groups and all intervals of the target groups. A target group is defined as the demographic group that is currently being examined. In this study, there were two target groups: males and females. Within each target group, there were several intervals. The intervals were subgroups of the target group as determined by estimates of the latent trait (Linn and Harnisch, 1981, p. 111).

If one interval of the target group in question is designated as g, the following formula gives the expected proportion of the people within interval g who will correctly respond to that item (ng is the number of people in interval g):

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij},$$

The expected proportion of people who will get an item correct in the target group (using all intervals) is given as:

$$P_{i.} = \sum_g n_g P_{ig} / \sum_g n_g.$$

The observed proportion correct on item i for interval g (Oig) is the percent of people in interval g who answer an item correctly (Linn & Harnisch, 1981, p. 111). The following formula gives the observed proportion correct for the target group (using all intervals):

$$O_{i.} = \sum_{g} n_g O_{ig} / \sum_{g} n_g .$$

This procedure allows for the computation of two difference scores. The total difference score for the whole target group is given as:

$$D_{i.} = O_{i.} - P_{i.},$$

The interval difference score for interval g of the target group is given as:

$$D_{ig} = O_{ig} - P_{ig},$$

Consideration of these interval difference scores is vital, because there are plausible cases in which Di. may be small, but upon examination of interval difference scores, there may be a significant difference. In this case, the target group would differ both from the overall (all target groups

combined) group as well as the expected score for that particular interval.

Standardized difference scores are also computed; the standardized interval difference score for interval g is:

$$Z_{ig} = \frac{1}{n_g} \sum_{j \in g} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}(1 - P_{ij})}} \right],$$

Uij is 1 when the item is answered correctly and 0 when there is an incorrect answer. The standardized total difference for this target group (all intervals combined) is:

$$Z_{i.} = \sum_g n_g Z_{ig} / \sum_g n_g .$$

The major disadvantage to this method is that the estimates of the item parameters will be influenced by the relative size of the target group in question. The effect of having a relatively large target group would cause bias indicators to become conservative. If this is a major concern, non-target group members can be used to estimate item parameters, and these item parameters can be fixed so

that estimates of the latent trait could be computed in the target group. The standard procedure would be followed from this point (Linn & Harnisch, 1981).

The second disadvantage for pseudo-IRT concerns the decision regarding how many regions of the latent trait were necessary to determine the intervals. Quintiles have been used, but they are not the necessary choice.

The third disadvantage is that the nonstandardized total and interval difference scores are sample dependent; that is, the difference scores can not be compared across target groups or testing sessions (Linn & Harnisch, p. 116, 1981). However, this purpose could be served by the standardized total and interval difference score in terms of comparing across target groups.

Shepard et al. (1985) reported that the pseudo-IRT approach is the most accurate with respect to estimation to the IRT approach. Hambleton et al. (1991) suggested that pseudo-IRT be used in studies with small samples as an useful alternative to IRT methods. Small samples in this case would be defined as samples that number less than 1,000 subjects in each of their subgroups.

Summary and Hypotheses

Hulin et al. (1983) noted that in research involving the comparison of ICC's, such as item bias research, accurate estimation of ICC's is vital. The number of items required

and the number of examinees will determine the accuracy of the estimates. Scales with a small number of items (defined by Hulin et al. (1983) as at least 30 items), combined with a sample size of at least 1,000 examines are expected to produce stable estimates in the three parameter model. With smaller sample sizes and numbers of items, the root mean squared errors of theta may become unacceptable large for item bias research. The Myers-Briggs Type Indicator subscales consist of from 22 to 26 items which may make them questionable from a true IRT viewpoint. Use of the iterative linking methods described previously would compound the small-number-of-items problem. Once items are determined to be biased and are removed from the analysis, the parameters are reestimated. Without a corresponding increase in sample size, the root mean squared errors of theta will increase, and the item pools for the MBTI subscale may become unacceptably small.

However, these comments concern themselves only with joint maximum likelihood estimation (JMLE) procedures; LOGIST is one estimation program that uses JMLE. The computer program BILOG (Mislevy & Bock, 1986) was used in this study to compute the item parameter and ability estimates. BILOG is the preferred estimation program for several reasons. First, it is recommended for circumstances that entail small samples (under 1,000 subjects per group)

and/or small numbers of items (under 50 items) (Mislevy & Stocking, 1989). Second, it is recommended because it uses either the marginal maximum likelihood estimation (MMLE) or the marginal maximum a posteriori estimation (MMAPE) approach to parameter estimation. Cohen, Kim and Subkoviak (1991) tested the utility of the MMAPE approach using the 2 parameter model; they did not use the 3 parameter model because they acknowledged the problem that small samples can introduce into the estimation of the c parameter. Cohen et al. (1991) defined their large sample as having 1000 subjects in each of the 2 subgroups and their small sample as having 1000 subjects in one subgroup and 200 subjects in another subgroup; their test consisted of 50 items. Parameters were estimated with and without the use of priors using MMLE as well as JMLE. Raju's signed and unsigned area estimates of item bias were then computed. For the two parameter model in a small sample, Cohen et al. (1991) found that the use of BILOG with priors was the most effective estimation procedure.

Hambleton et al. (1991) noted that MMLE (without the use of priors) estimates parameters without incorrect extreme values only with a large number of examinees; a large number of examinees are needed to best approximate the distribution of ability. In turn, this distribution of ability is needed to compute the marginal maximum likelihood

function of the item parameters (Hambleton et al., 1991, p. 43). MMLE was used in this study to estimate parameters for the full sample (capitalizing on the use of MMLE in a large sample). Linn and Harnisch's (1981) Pseudo-IRT method was used to determine which MBTI items are biased for males versus females. Combining the two groups and using MMLE negated concern for the robustness of the estimates; this use of MMLE and pseudo-IRT will incorporate the best features of both procedures.

The disadvantage of using pseudo-IRT in this scenario is that because of the relative size of the male and female group, the bias indicators may become too conservative. On the plus side, the conservative nature of the bias indicator nullifies the possibility of falsely assigning items that are unbiased the label of being biased. The male and female subgroups are very similar is size, negating the concern for the relative influence of each subgroup on the estimates. Using both groups independently as target groups also gave some measure of convergence on DIF items. If DIF was found for both males and females on a particular item, it would be a strong indicator that item has psychometric problems.

Hypothesis I Items on the Thinking-Feeling scale of the Myers-Briggs type indicator that are differential scored for males and females under the published scoring system will be determined to function differently for males and females based on Pseudo-IRT bias methods.

Hypothesis II Items on the remaining three scales will not be determined to be biased on the basis of sex.

Method

Instrument: The Myers-Briggs Type Indicator

Form F of the MBTI was used. Only the 94 scored items were examined. Table 1 reports the scoring system developed by the test constructors. For the purpose of this bias research, the indicator was treated as four independent scales. The T and F scales have been modified to nullify the differential scoring used by the test authors. Items on the TF scale were given a value of 1 if answered in the F direction and 0 otherwise (i.e., the differential weighting on the basis of gender was removed). It was hypothesized that it is these differentially scored (by the test authors) items that would be determined to be biased. This was not the case, and seriously calls into question the scoring system recommended by the test authors. These modified

scales are shown in Table 2. For the remaining three scales, items were given a value of 1 if the response was in the I, N, or P direction and 0 otherwise. It is important to note in Table 3 that some items are only valid for one of the two poles (as denoted by e, i, s, n, t, f, j, or p). For the items that are classified as valid (by the test authors) for one pole, but the item is scored for the opposite pole, reverse coding of responses is used. In essence, the item is treated as though it was valid for both poles. Theoretically, the traits being measured are bipolar and to treat them otherwise violates the assumption of bipolarity. In a practical sense, if an item is defined as valid only towards one pole, lack of endorsement towards that pole in essence is support for the opposite pole, Finally, and most importantly, it is important to examine if these items indeed function for only one pole.

Data Set

The data consists of an archival data base of Myers-Briggs Type Indicator item responses obtained at a large southeastern university, as well as item responses collected in two data gathering waves. The data consists of both graduate students and undergraduates. There are 883 males and 1004 females in the sample.

Parameter Estimation technique

BILOG was used to estimate the item and ability

parameters; the three parameter model was used. There were four estimation runs to determine the estimates (a, b, c, and theta; the items in the EI, SN, TF, and JP scales were analysed separately using all 1883 subjects to obtain these estimates.

It is appropriate to combine the items for the dichotomous poles into one item pool for two reasons. One is that the original MBTI researchers as well as recent factor analytic studies (Harvey, Becker, Murry, Lawless, Stamoulis & Brill, 1991), support the premise that the four scales are each unidimensional. The second reason is the property of invariance assumed in the use of a IRT model. This property states that one of the benefits of the use of IRT methods is that the item parameters computed for a particular examinee do not rely on the particular set of test items (Hambleton et al., 1991, p. 18), as long as the total pool is effectively unidimensional.

Pseudo-IRT methods

The pseudo-IRT approach was used on each of the four subscale calibrations. It was hypothesized that only a subset of items in the TF scale analysis would be determined to be biased; These items are labeled in table 3 by a "M" if they were hypothesized to be biased against males, and by a "F" if they were hypothesized to be biased against females. It was also hypothesized that if different items were

determined to be biased, this would call into question the scoring system recommended by the test authors.

Figure 2 describes the sequence of events in the Pseudo-IRT analysis. The first step was the division of the data into the four sub sets for the four sub scales. The next steps were be carried out on each of the data sets. The data for males and females were combined and a, b, c and theta were be computed. Males were assigned as the target group. The appropriate levels of theta were determined by an examination of the frequency distribution of theta. The observed and expected scores were computed for each interval of theta for each item, as well as observed and expected scores for the overall data set. Difference scores were computed using these observed and expected scores. Females were then assigned as the target group and the process repeated. Standardized difference scores were also computed.

Results

Parameter Estimation

Each subset of the data (N=883 males, N=1004 females) was dichotomously scored based on the scoring system in Table 3. Four PC-Bilog programs were run to estimate item, ability, and test parameters for the EI, SN, TF, and JP data sets. Figure 3 shows the information function for the EI test. Table 4 presents the item parameters and standard

errors of these estimates for the EI data set. The item characteristic curve for each of the EI items is given in Appendix A. Table 5 presents the item parameters and standard errors of these estimates for the SN data set. The item characteristic curve for each of the SN items is given in Appendix B. Figure 4 shows the information function for the SN test. Table 6 presents the item parameters and standard errors of these estimates for the TF data set. The item characteristic curve for each of the TF items is given in Appendix C. Figure 5 shows the information function for the SN test. Table 7 presents the item parameters and standard errors of these estimates for the JP data set. The item characteristic curve for each of the JP items is given in Appendix D. Figure 6 shows the information function for the JP data set.

There are two main conclusions to be drawn from these data. The first is that by examination of the standard errors of the item parameter estimates, it is apparent that the MMLE estimation technique compensated for the relatively small number of items with the large sample size; these standard errors are acceptably low, indicating that the parameter estimates should be stable. The stability of the estimates is especially important in differential item functioning research, because differences between subgroups when there is a high degree of error in measurement may be

due solely to measurement error. Differences between groups when parameters are stable may be accounted for by actual differences between the groups.

The second conclusion, based on the test information functions, is that the four subtests are not equally precise at estimating the latent trait at all levels of theta. The standard error of theta is relatively small at the range of -.5 to +.5 theta, but much larger outside that range. On all four subtests the level of the standard error of the estimated theta changes dramatically over the full range of theta values. In general, for the non extreme theta values, these standard errors are probably acceptable.

Difference Score Computation

Item and ability estimates produced by PC-Bilog and dichotomously scored raw data were used to produce interval and total difference scores, as well as standardized difference scores. Before these values were generated, intervals on theta for each of the four data sets were examined in order to determine the appropriate interval difference score parameters. There was some congruence across the four combined samples, but when the data were split up into target groups, there was no match between the distributions of theta. In classical test theory, finding this differential ability in subgroups would indicate that items are biased. In IRT analysis (and Pseudo-IRT

analysis), differences in mean ability levels will not necessarily indicate that the item is biased. In this type of analysis, bias is indicated if, when ability level is held constant, items differ based solely on subgroup membership of respondents. For this analysis, levels of theta must be defined to produce interval scores. If theta levels are not consistent across target groups, it is then advised to set up intervals on theta with roughly equivalent sample sizes.

Consequently, for the thinking-feeling subscale, quintiles with near equivalent cell sizes were examined. The division points on theta for the male target group were -1.163733, -.634374, -.189976 and .377199. The division points for the female target group were -.341306, .209974, .648837, and 1.302337.

This procedure was followed for the other three subscales. For the judging-perception subscale, the division points on theta for the male target group were -1.052086, -.263965, .253577 and .879064. The division points for the female target group were -1.038055, -.305514, .184735, and .74962. For the extraversion-introversion subscale, the division points on theta for the male target group were -.756272, -.155076, .427447 and 1.050193. The division points for the female target group were -1.221407, -.514122, .077521, and .663311. For the sensing-intuition subscale, the division points on theta for the male target group were

-.904382, -.286328 .206856 and .943907. The division points for the female target group were -.791249, -.186519, .275615, and .867026.

These quintiles did not converge across the four subtests because of mean differences in distribution of theta for males and females. Table 8 shows the total, male, and female mean levels of theta for each subtest. As indicated in Table 8, mean levels of theta differ across gender groups on each of the subtests. In classical test theory methods of detecting item bias, this alone would promote a conclusion that items were biased. In IRT techniques, mean differences in level of theta do not necessarily cause items to be classified as having DIF.

Because quintiles differ across subgroups. it would not be useful to compare these interval difference score across subscales because each interval is based on a different level of theta dependent on the subtest. However, these were still computed in order to check on the accuracy of the computation of the total difference scores.

Standardized total difference scores for each item in both target groups for each sub scale were then computed. For the thinking/feeling scale, the mean difference score was -.0336884 (SD = .0544623). The minimum value was -.1554881 and the maximum was .0636397 (range = .2191278). For the extroversion/intraversion scale, the mean total

difference score was -.0299287 (SD = .1358581). The minimum value was -.6601512 and the maximum was .1433621 (range = .8035133). For the sensing/intuition scale, the mean total difference score was -.0110811 (SD = .0786800). The minimum value was -.1519357 and the maximum was .1595772 (range = .3115129). For the judging/perception scale, the mean total difference score was .000359341 (SD = .1155137). The minimum value was -.4235477 and the maximum was .2000619 (range = .6236096).

Total difference scores were also examined across sub scales in terms of male and female target groups. For all male target groups, the mean total difference score was -.0064209 (SD = .1064168). The minimum value was -.6601512 and the maximum value was .2000619 (range = .8602131). For all female target groups, the mean total difference score was -.0300282 (SD = .0927194). The minimum value was -.4235477 and the maximum value was .1915416 (range = .6150893).

Standardized total difference scores were also examined across all target groups and subscales. The mean of this distribution was -.0182243 (SD = .1002373). The minimum was -.6601512 and the maximum was .2000619 (range = .8602131). This distribution approximates a normal distribution (see Figure 7) except for three outliers. The total difference score for males on EI item 50 is -.6601512, while the total

difference score for females on EI item 50 is -.3911648. This congruence gives strong evidence that there is a problem with item 50. Item 50 reads as follows:

" 50. Are you usually

A) a "good mixer", or

B) quiet and reserved? "

Alternative B is the introverted response. This item uses slang that was appropriate in the 1940's when the test was first developed, which may be ambiguous to the young collegiate sample used in this study. Hulin et al. (1983) noted that DIF may occur if an item has different meanings for different subgroups. They cite an item on a children's intelligence test that asks what the child would do if he or she lost one of their friend's balls. Some rural children were penalized when they responded that they would get a doctor. If this item is interpreted in an anatomical fashion, this might be a correct response. While this is an extreme example, it illustrates how differences in word meaning can prompt DIF.

Congruence among total difference scores across target groups is not necessary for an item to be considered to have DIF. Another equally plausible result is for an item to have a large positive total difference score for one target

group and a large negative total difference score for the other target group. An item may also show a large positive or negative total difference for a particular target score. This result would show that the item only has DIF in one group. The third extreme score is an example of this last case. The total difference score for females on SN item 85 is -.4235477. Item 85 is a word pair that reads as follows:

> *"Which word in each pair appeals to you more? Think what the words mean, not how they look or how they sound.*
>
> *85. A) scheduled B) unplanned"*

B is the perceiving alternative. The reason for DIF on this item is not as clear as with item 50. It warrants further study with a more diverse sample. However, it is important to note that even for these DIF items, the total difference scores are not especially large.

An alternate explanation for the large total difference scores for item 50 and item 85 is that because they have the most extreme a parameters of all 94 items, the target groups cannot accurately estimate the proportion of the respondents who answered the item correctly. The a parameter for item 50 is 2.213 and the a parameter for item 85 is 1.978. The

full sample may be necessary to reproduce the extreme ICC's produced by these values.

Discussion

There were two possible outcomes to this study. The first possible outcome was that the items on the thinking/feeling scale that are differentially scored and/or differentially weighted for males and females would show evidence of DIF. This result would support the use of the differential scoring system devised by the test authors. The second possible outcome was that the DIF findings would not match the scoring system pattern, and thus support the view that the differential weighting and scoring of thinking/feeling items based on gender is inappropriate. This evidence would occur at the item level, as simple mean differences in the latent trait would not necessarily indicate differential item functioning. The results of this study support the latter conclusion.

Myers and McCaulley (1985) give the following rational for the differential scoring on the TF scale. At first, male and female keys were used for all the subscales. After form E, separate keys were only retained for the TF subscales because item popularity and prediction ratios were the same for items on the other scales (p. 148). The TF subscale retained differential scoring because females who showed that they were thinking types (through inferences

from behavior and attitudes) sometimes preferred feeling responses. This incorrect response was thought to be due to a social desirability effect or cultural norms (p. 148). Weights, by gender, were applied to the appropriate TF items, to match endorsement by the original criterion subjects.

Studies were undertaken using the same methodology that not surprisingly supported these findings. At first the split between thinking/feeling types in males was 60% T and 40 % F. The split among females was 33% T and 66 % F. In the 1970's, it became apparent that there was a shift in the distribution of TF feeling types. Evidence was gathered that showed 44% T males and 30% T females (p. 148). When the differential scoring system is applied to the current sample, there were 66.9% Thinking males and 31.1% Thinking females. However, when an unweighted scoring system was used, there were 41.7% Thinking males and 13.6% Thinking females.

There are two explanations for these findings. One is that there are simply less Thinking types in the population, which is unacceptable in type theory. When Jung described his typology, he was describing "basic psychological processes (Hall & Nordby, 1973)." He was describing what he considered to be the building blocks of personality. Type theory would also indicate that the proportion of types in

the population should remain constant across time. Jung considered males to primarily thinking types and females to be primarily feeling types (Hall & Nordby, 1973).

The position that the test authors took was that certain feeling responses were more socially acceptable now than when the test was initially constructed. Evidence of socially desirability would be discovered by pseudo-IRT analysis. This technique looks at both the level of the latent trait and the response so responses inconsistent with the latent trait would become apparent.

Item total correlations and prediction ratios were then computed by the MBTI authors on a new standardization sample. Feeling choices were shown to be more popular and reweighted and Thinking responses were less popular and were also reweighted. The new weightings produced 61 % T males and 30% thinking females.

The question then becomes, should items be reweighted to prompt some change in the mean test level? If there is some external criterion, such as the test constructors' belief in a fixed proportion of thinking/feeling types in the population, the answer may be yes. However, that "yes" is based on many inferences. If item level analysis, based on quantitative information is preferred, such in this case, the evidence that items should be weighted differently for males and females should be based on item level information

without regard to some external criterion.

The pseudo-IRT technique allows for mean differences to exist in subgroups. If items are shown to function differently for males and females, this would be evidence to justify differentially scoring those items. According to the test authors, initial differential scoring for males and females was done because the latent trait was not prompting the correct item response. Of course, the best course of action would have been to remove --or rewrite-- the offending items if indeed they were biased. However, the results obtained in this study show that there is no differential item functioning on the TF scale. The next course of action is to tackle the test authors' and Carl Jung's conception that the proportion of types in the population is fixed. If there truly is this set proportion of thinking/feeling types in the population, it should not be necessary to tamper with the item responses by weighting. If the items were bad (if they had DIF) then some weeding out or perhaps reweighting of items could be justified. However, at this point there is no evidence for this item level manipulation. Future studies should examine a non weighted, non differential scored TF scale to examine what the true population thinking/feeling proportion may be.

References

Baker, F. B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement. 18, 59-62.

Carskadon, T. G. (1977). Test-reliabilities of continuous scores on the Myers-Briggs Type Indicator. Psychological Reports. 41, 1011-1012.

Carskadon, T. G. (1982). Sex differences in test-retest reliabilities of continuous scores on form G of the Myers-Briggs Type Indicator. Research in Psychological Type. 5, 78-79.

Cohen, A. S., Kim, S. H., & Subkoviak, M. J. (1991). Influence of prior distributions on detection of DIF. Journal of Educational Measurement. (28)1, 49-59.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. Applied Psychological Measurement. (9)4, 413-415.

Drasgow, F. (1987). Study of the measurement of bias of two standardized psychological tests. Journal of Applied Psychology. 72, 19-29.

Guzie, T. & Guzie, N. M. (1984). Masculine and feminine archetypes: a complement to the psychological types. Journal of Psychological Type. 7, 3-11.

Hall, C. S. & Nordby, C. S. (1973) A Primer of Jungian Psychology. New York, NY: Penguin Books.

Harris. A.H., & Carskadon, T. G. (1988). Comparative validity of the old and new scoring weights on the MBTI Thinking-Feeling scale. Journal Of Psychological Type. 15, 54-62.

Harvey, R. J., Becker, R. L., Murry, W., Lawless, W., Stamoulis, D. & Brill, R. (1991). Dimensionality of the MBTI. Paper presented at the annual convention of the American Psychological Association.

Hambleton, R. K. & Swaminathan, H. (1985). Item Response Theory: Principles and applications. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H. & Rogers H. J (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, Inc.

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). Item Response Theory. Homewood, Il: Dow-Jones Irwin.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. Journal of Applied Psychology, 71(2), 327-333.

Linn, R. L. & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18(2), 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension (Tech Rep. No. 163). Urbana IL: University of Illinois , Center for the Study of Reading.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Marascuilo. L. A. & Slaughter. R. E. (1981) Statistical procedures for identifying possible sources of item bias based on chi square statistics. Journal of Educational Measurement. 18, 229-248.

McCaulley, M. H. (1990). The Myers-Briggs Type Indicator and Leadership. In K. E. Clark & M. B. Clark (Eds.), Measures of Leadership (pp. 381-418). Greensboro, NC: Center for Creative Leadership.

Mislevy, R. J., & Bock, R. D. (1986). PC-BILOG: item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13(1), 57-75.

Myers, I. B, & McCaulley, M. H. (1985). Manual: A Guide to the Development of the Myers-Briggs Type Indicator. Palo Alto, CA: Consulting Psychologists Press, Inc.

Osterlind, S. J. (1983). Test Item Bias. Sage University Paper series on Quantitative Application in the Social Sciences, 07-030. Beverly Hills and London: Sage Pubns.

Padgett, V. R. Cook, D. D., Nunley, M. E., & Carskadon. T.G. (1982). Psychological type, androgyny, and sex typed roles. Research in Psychological Type. 5, 69-77.

Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. Applied Psychological Measurement. (14)2, 162-173.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika. 53(4), 495-502.

Scheimeiser, C. B. (1982). Use of experimental methods in item bias studies. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias (pp. 64-95). Baltimore: Johns Hopkins University Press.

Scheuneman, J. D. (1979). A method of assessing bias in test items. Journal of Educational Measurement. 16, 143-152.

Shepard, L., Camilli, G. & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22(2), 77-105.

Stocking, M. L. & Lord, F. M. (1983). Developing a common f metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stokes, J. (1987a). Exploring the relationship of type and gender - part 1: anecdotal experiences of MBTI users. Journal of Psychological Type. 13, 34-43.

Stokes, J. (1987b). Exploring the relationship of type and gender - part 1: a review and critique of empirical research and other data. Journal of Psychological Type. 13, 34-43.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias (pp. 31-63). Baltimore: Johns Hopkins University Press.

Warm, T. A. (1977). A primer of item response theory. (Technical Report No. 941078). Washington, D.C.: U.S. Coast Guard Institute.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1: Scoring System for MBTI Form F

E	I	S	N	T(m)	T(f)	F(m)	F(f)	J	P
6a	6b*	2b	2a*	4b*	4b	26a*	26a*	1a*	1b*
15a	19a	11a	11b	26b*	26b*	29a	29a	13b	9b,
19b*	25b	17b	37b*	29b*	29b*	79b*	79b*	20b	c
25a	33b*	37a	53b	72a*	72a*	81b	81b	27a*	13a
33a	41a	53a	70a*	79a*	79a	84a	84a	35a*	20a
41b*	47c*	64a*	76a*	81a	81a	86b*	86b	42b	27b
47b	50b*	70b	78b	84b	84b	100b	100b*	55a*	35b
50a*	58a	73b*	88b	86a*	86a*	103a	103a	60a	42a
58b*	66b	76b	102b	89b*	89b*	105b*	105b*	74a*	49b*
66a	87a	78a*	104b*	91b*	91b*	111a	114b	5A*	55b*
77a	92b*	88a*	112b*	93b	93b	114b*	122a	94b	60b,
87b*	95b	90a	115a*	100a	100a	147a	147a*	97a	c
92a	106a	98a*	119b	103b*	103b*	154a	154a*	99b	74b*
106b	116b	102a*	128b*	105a	105a	158a	158a*	109a*	85b*
116a	126b*	104a	145b*	108a*	108a*			118a*	94a*
126a	129b*	107a*	149a	111b*	111b*			124a	97b
134a	134b*	117a		114a*	114a*			132a	109b
138a	138b	119a		120b*	120b*			151A	113a
148a*	148b*	121a		133a	133a			153b	118b*
160a	160b	128a*		154b	154b*				132b*
		140b							142b
		145a*							151b
		149b*							153a
		165b							

* given a weighting of 2

Table 2: Scoring System for MBTI Form F (combined)

E	I	S	N	T	F	J	P
6a	6b*	2b	2a*	4b*f	26a*	1a*	1b*
15a	19a	11a	11b	26b*	29a	13b	9b,c
19b	25b	17b	37b*	29b*	79b*	20b	13a
25a	33b*	37a	53b	72a*	81b	27a*	20a
33a	41a	53a	70a*	81a	84a	35a*	27b
41b*	47c*	64a*	76a*	84b	86b*f	42b	35b
47b	50b*	70b	78b	86a*	100b*m	55a*	42a
50a*	58a	73b*	88b	89b*	103b	60a	49b*
58b*	66b	76b	102b	91b*	105b*	74a*	55b*
66a	87a	78a*	104b*	93b	111a*f	85a*	60b,c
77a	92b*	88a*	112b*	100a	114b*f	94b	74b*
87b*	95b	90a	115a*	103a*	122a*m	97a	85b*
92A	106a	98a*	119b	105a	147a*m	99a	94a*
106b	116b	102a*	128b*	108a*	154a*m	109a*	97b
116a	126b*	104a	145b*	111b*	158a*m	118a*	109b
126a	129b*	107a*	149a	114a*		124a	113a
134a	134b*	117a		120b*	n=15	132a	118b*
138a	138b	119a	n=16	133a		151a	132b*
148a*	148b*	121a		154bm		153b	142b
160a	160b	128a*					151b
		140b		n=19		n=19	153a
n=20	n=20	145a*					
		149b*					n=21
		165b					
		n=24					

* given a weighting of 2

m = assumed to be biased against males
f = assumed to be biased against females

Table 3: Item Pools for Pseudo IRT Analysis

EI	SN	TF	JP
scored towards I	scored towards N	scored towards F	scored towards P
6b	2a	4at	1b
15be	11b	26a	9b,cp
19a	17as	29a	13a
25b	37b	72bt	20a
33b	53b	79bM	27b
41a	64bs	81b	35b
47c	70a	84a	42a
50b	73as	86bF	49b
58a	76a	89at	55b
66b	78b	91at	60b,c
77be	88b	93at	74b
87a	90bs	100bM	85b
92b	98bs	103a	94a
95bi	102b	105b	97b
106a	104b	108bt	99aj
116b	107bs	111aF	109b
126b	112bn	114bF	113ap
129bi	115an	120at	118b
134b	117bs	122aM	124bj
138b	119b	133bt	132b
148b	121b	147aM	142bp
160b	128b	154aM	151b
	140as	158aM	153a
n=22	145b		
	149a	n=23	n=23
	165as		
	n=26		

e = scored for E scale only
i = scored for I scale only
s = scored for S scale only
n = scored for N scale only
t = scored for T scale only
f = scored for F scale only
j = scored for J scale only
p = scored for P scale only
M = assumed to be biased against males
F = assumed to be biased against females

Table 4: Item Parameter and Standard Errors for EI subscale

SUBTEST EI : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT S.E.	A S.E.	B S.E.	DISPERSN S.E.	C S.E.	CHISQ (PROB)	DF
M6	.075	.810	-.092	1.234	.018	23.5	9.0
	.053*	.060*	.068*	.091*	.027*	(.0053)	
M15	.086	.782	-.110	1.279	.094	27.1	8.0
	.107*	.095*	.148*	.156*	.064*	(.0008)	
M19	.527	.729	-.723	1.371	.015	39.4	8.0
	.043*	.051*	.072*	.095*	.024*	(.0000)	
M25	-.724	.994	.729	1.006	.003	18.4	9.0
	.049*	.063*	.043*	.064*	.005*	(.0312)	
M33	-.657	.927	.708	1.078	.004	14.4	9.0
	.048*	.061*	.045*	.071*	.006*	(.1096)	
M41	-.230	.438	.526	2.285	.009	22.8	9.0
	.038*	.038*	.087*	.196*	.014*	(.0068)	
M47	-.458	.383	1.194	2.609	.015	13.5	9.0
	.060*	.041*	.143*	.277*	.024*	(.1406)	
M50	-.524	2.213	.237	.452	.002	27.2	5.0
	.069*	.155*	.024*	.032*	.003*	(.0001)	
M58	.278	.416	-.667	2.404	.022	18.4	9.0
	.048*	.037*	.137*	.215*	.035*	(.0311)	
M66	-.537	.667	.806	1.500	.006	12.6	9.0
	.044*	.048*	.063*	.107*	.009*	(.1816)	
M77	-.544	.436	1.248	2.295	.016	24.3	9.0
	.069*	.047*	.129*	.246*	.025*	(.0039)	
M87	-.449	.930	.483	1.076	.006	11.5	9.0
	.046*	.061*	.043*	.070*	.009*	(.2437)	
M92	-.454	.675	.672	1.482	.013	17.1	9.0
	.061*	.056*	.070*	.122*	.020*	(.0471)	
M95	-.481	.514	.937	1.945	.028	11.9	9.0
	.095*	.062*	.122*	.234*	.037*	(.2182)	
M106	-.525	.696	.753	1.436	.008	17.7	9.0
	.051*	.053*	.062*	.110*	.013*	(.0387)	
M116	-1.704	.995	1.712	1.005	.007	10.5	8.0
	.136*	.111*	.091*	.112*	.007*	(.2324)	
M126	.147	.975	-.151	1.025	.006	39.0	8.0
	.038*	.059*	.040*	.062*	.010*	(.0000)	
M129	.514	.353	-1.458	2.835	.020	13.3	9.0
	.043*	.036*	.184*	.286*	.032*	(.1500)	
M134	.046	.804	-.057	1.244	.022	16.3	9.0
	.058*	.063*	.074*	.098*	.030*	(.0601)	
M138	.231	.718	-.321	1.393	.040	10.4	8.0
	.074*	.065*	.122*	.127*	.052*	(.2383)	
M148	-.075	.976	.077	1.024	.015	12.9	9.0
	.054*	.070*	.053*	.073*	.021*	(.1673)	
M160	-.499	.545	.916	1.833	.005	37.5	9.0
	.039*	.043*	.080*	.146*	.009*	(.0000)	

Table 5: Item Parameter and Standard Errors for SN subscale

SUBTEST SN : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT S.E.	A S.E.	B S.E.	DISPERSN S.E.	C S.E.	CHISQ (PROB)	DF
M2	-.327	.963	.340	1.038	.023	17.0	9.0
	.055*	.070*	.048*	.076*	.016*	(.0491)	
M11	-.010	.609	.017	1.641	.102	13.4	9.0
	.095*	.069*	.155*	.185*	.055*	(.1463)	
M17	-.313	.875	.358	1.143	.146	16.8	9.0
	.107*	.103*	.093*	.135*	.038*	(.0516)	
M37	.158	.754	-.209	1.326	.285	15.4	8.0
	.129*	.106*	.193*	.186*	.068*	(.0515)	
M53	-.638	.541	1.180	1.849	.233	12.5	9.0
	.194*	.108*	.188*	.370*	.053*	(.1882)	
M64	.017	.497	-.035	2.014	.232	11.4	9.0
	.156*	.081*	.318*	.329*	.088*	(.2460)	
M70	-.112	.648	.172	1.543	.064	3.8	9.0
	.076*	.061*	.107*	.146*	.039*	(.9249)	
M73	1.089	1.114	-.977	.897	.060	4.7	6.0
	.065*	.088*	.080*	.071*	.042*	(.5854)	
M76	-.859	.885	.971	1.129	.049	16.5	9.0
	.110*	.096*	.066*	.123*	.021*	(.0569)	
M78	.248	.821	-.302	1.218	.020	34.6	9.0
	.039*	.052*	.052*	.077*	.015*	(.0001)	
M88	.454	.810	-.561	1.235	.055	7.3	8.0
	.056*	.063*	.089*	.097*	.038*	(.5084)	
M90	.516	.703	-.735	1.423	.038	22.9	9.0
	.045*	.048*	.082*	.097*	.028*	(.0066)	
M98	.580	.772	-.751	1.295	.057	13.7	8.0
	.055*	.057*	.096*	.096*	.040*	(.0880)	
M102	.513	1.014	-.506	.986	.025	19.2	8.0
	.043*	.064*	.050*	.063*	.018*	(.0141)	
M104	-.218	1.241	.176	.806	.016	16.9	9.0
	.048*	.080*	.036*	.052*	.011*	(.0497)	
M107	.951	1.073	-.887	.932	.055	11.9	6.0
	.060*	.083*	.076*	.072*	.039*	(.0646)	
M112	-.797	.605	1.317	1.652	.030	10.7	9.0
	.084*	.065*	.099*	.177*	.020*	(.2945)	
M115	-1.952	.671	2.910	1.490	.060	12.4	9.0
	.318*	.172*	.349*	.382*	.016*	(.1886)	
M117	.273	.667	-.410	1.499	.083	9.6	9.0
	.074*	.062*	.134*	.138*	.053*	(.3818)	
M119	.001	.864	-.001	1.157	.033	15.9	9.0
	.051*	.064*	.059*	.085*	.023*	(.0680)	
M121	-.085	.460	.184	2.173	.087	12.8	9.0
	.095*	.055*	.192*	.259*	.057*	(.1708)	
M128	-.433	1.055	.411	.948	.050	25.9	8.0
	.081*	.100*	.054*	.090*	.024*	(.0012)	
M140	.861	.399	-2.160	2.508	.079	12.2	9.0
	.062*	.041*	.263*	.256*	.058*	(.2010)	
M145	.014	1.018	-.014	.982	.031	31.5	8.0
	.051*	.076*	.050*	.073*	.021*	(.0001)	
M149	.066	.827	-.080	1.210	.090	4.4	9.0
	.079*	.078*	.100*	.114*	.043*	(.8817)	
M165	.126	.812	-.155	1.232	.111	12.4	9.0
	.084*	.082*	.115*	.124*	.049*	(.1905)	

Table 6: Item Parameter and Standard Errors for TF subscale

SUBTEST TF : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT S.E.	A S.E.	B S.E.	DISPERSN S.E.	C S.E.	CHISQ (PROB)	DF
M4	1.127	.726	-1.553	1.378	.032	5.9	7.0
	.058*	.056*	.112*	.106*	.038*	(.5479)	
M26	.074	1.141	-.065	.876	.036	21.1	7.0
	.056*	.094*	.052*	.072*	.024*	(.0037)	
M29	.565	.796	-.711	1.257	.019	7.7	8.0
	.043*	.050*	.067*	.080*	.023*	(.4678)	
M72	1.651	1.311	-1.259	.763	.025	12.9	6.0
	.081*	.086*	.061*	.050*	.029*	(.0436)	
M79	-.240	.973	.246	1.028	.006	21.6	8.0
	.040*	.062*	.039*	.065*	.008*	(.0058)	
M81	-.105	.580	.181	1.723	.012	15.9	9.0
	.038*	.043*	.064*	.129*	.014*	(.0685)	
M84	.073	.542	-.135	1.844	.069	24.2	9.0
	.095*	.063*	.186*	.214*	.064*	(.0041)	
M86	.387	1.231	-.314	.812	.014	23.0	7.0
	.045*	.083*	.040*	.055*	.016*	(.0018)	
M89	.952	.922	-1.033	1.085	.034	5.1	6.0
	.057*	.068*	.086*	.080*	.039*	(.5331)	
M91	.923	.645	-1.430	1.550	.027	17.3	8.0
	.050*	.050*	.112*	.120*	.033*	(.0272)	
M93	.542	.461	-1.177	2.171	.030	15.6	9.0
	.047*	.039*	.143*	.185*	.036*	(.0758)	
M100	.234	.647	-.362	1.545	.055	13.6	8.0
	.073*	.062*	.136*	.147*	.053*	(.0913)	
M103	.609	1.129	-.540	.886	.033	13.5	7.0
	.053*	.086*	.063*	.068*	.031*	(.0607)	
M105	-.526	.650	.809	1.537	.008	44.2	9.0
	.044*	.051*	.068*	.121*	.010*	(.0000)	
M108	1.174	.503	-2.332	1.986	.032	5.2	8.0
	.055*	.047*	.202*	.184*	.039*	(.7383)	
M111	.978	1.127	-.868	.887	.020	8.9	7.0
	.055*	.075*	.056*	.059*	.023*	(.2581)	
M114	.520	1.500	-.347	.667	.025	26.3	6.0
	.051*	.114*	.040*	.051*	.019*	(.0002)	
M120	.949	.674	-1.409	1.484	.027	5.6	7.0
	.051*	.051*	.107*	.111*	.033*	(.5936)	
M122	-.667	.271	2.458	3.687	.064	19.0	9.0
	.164*	.059*	.385*	.801*	.058*	(.0251)	
M133	1.020	.560	-1.822	1.787	.026	20.8	8.0
	.049*	.046*	.144*	.146*	.032*	(.0078)	
M147	-1.055	.526	2.005	1.901	.028	11.8	9.0
	.133*	.085*	.171*	.307*	.024*	(.2257)	
M154	.069	.740	-.093	1.351	.016	21.4	9.0
	.042*	.051*	.059*	.094*	.019*	(.0112)	
M158	-.567	.709	.800	1.410	.052	23.9	9.0
	.106*	.089*	.086*	.177*	.032*	(.0045)	

Table 7: Item Parameter and Standard Errors for JP subscale

SUBTEST JP : ITEM PARAMETERS AFTER CYCLE 12

ITEM	INTERCEPT S.E.	SLOPE S.E.	THRESHOLD S.E.	DISPERSN S.E.	ASYMPTOTE S.E.	CHISQ (PROB)	DF
M1	-1.257	1.135	1.108	.881	.021	18.6	9.0
	.109*	.104*	.052*	.080*	.011*	(.0289)	
M9	-.197	.616	.320	1.624	.041	16.5	9.0
	.070*	.057*	.098*	.151*	.034*	(.0571)	
M13	-.269	.899	.299	1.113	.094	4.3	9.0
	.096*	.094*	.084*	.117*	.036*	(.8919)	
M20	-.802	.728	1.102	1.373	.115	5.6	9.0
	.150*	.108*	.098*	.205*	.034*	(.7794)	
M27	.280	.903	-.311	1.108	.040	14.5	8.0
	.057*	.068*	.074*	.083*	.033*	(.0701)	
M35	-.179	.636	.281	1.574	.066	11.0	9.0
	.090*	.067*	.122*	.165*	.045*	(.2734)	
M42	-.690	1.231	.561	.813	.194	17.1	9.0
	.139*	.156*	.062*	.103*	.026*	(.0468)	
M49	-1.749	1.216	1.438	.822	.034	22.6	9.0
	.171*	.137*	.064*	.093*	.010*	(.0074)	
M55	-.065	.999	.065	1.001	.032	19.7	9.0
	.059*	.077*	.057*	.077*	.024*	(.0198)	
M60	-.770	.884	.871	1.131	.042	10.4	9.0
	.100*	.091*	.064*	.116*	.022*	(.3186)	
M74	.739	1.212	-.610	.825	.024	34.0	7.0
	.052*	.080*	.049*	.054*	.022*	(.0000)	
M85	-.695	1.978	.351	.505	.008	31.0	6.0
	.079*	.160*	.025*	.041*	.006*	(.0000)	
M94	.043	.860	-.050	1.163	.014	27.1	9.0
	.039*	.053*	.046*	.071*	.013*	(.0014)	
M97	-.090	.736	.122	1.360	.027	12.1	9.0
	.052*	.055*	.067*	.101*	.024*	(.2061)	
M99	.440	.522	-.842	1.914	.074	12.1	9.0
	.075*	.048*	.186*	.177*	.064*	(.2052)	
M109	.926	.874	-1.060	1.144	.031	47.2	8.0
	.052*	.060*	.076*	.078*	.029*	(.0000)	
M113	-.535	.446	1.199	2.242	.029	12.5	9.0
	.071*	.048*	.127*	.239*	.025*	(.1843)	
M118	.512	.999	-.512	1.001	.016	45.2	8.0
	.042*	.062*	.047*	.062*	.015*	(.0000)	
M124	-.149	.454	.329	2.202	.076	17.9	9.0
	.104*	.057*	.201*	.277*	.059*	(.0367)	
M132	-.676	1.022	.661	.979	.042	13.3	9.0
	.092*	.099*	.054*	.095*	.021*	(.1495)	
M142	-.242	.704	.345	1.421	.107	6.1	9.0
	.110*	.085*	.125*	.172*	.048*	(.7293)	
M151	-1.141	1.191	.958	.840	.064	8.8	9.0
	.126*	.123*	.051*	.087*	.015*	(.4574)	
M153	-.801	.524	1.529	1.908	.047	3.9	9.0
	.122*	.074*	.131*	.269*	.032*	(.9149)	

Table 8: Distributions of theta for the Four Subscales

Scale	Sample	Mean	Std
EI	Total	-.1165683	1.2663275
	Males	.1153885	1.2181329
	Females	-.3205702	1.2733624
SN	Total	-.0051203	1.2619974
	Males	-.0261548	1.3422675
	Females	.0133791	1.1872942
TF	Total	.0683945	1.2531808
	Males	-.4207417	1.1231509
	Females	.4985810	1.2032752
JP	Total	-.1059643	1.2608171
	Males	-.0688969	1.3181030
	Females	-.1385644	1.2079147

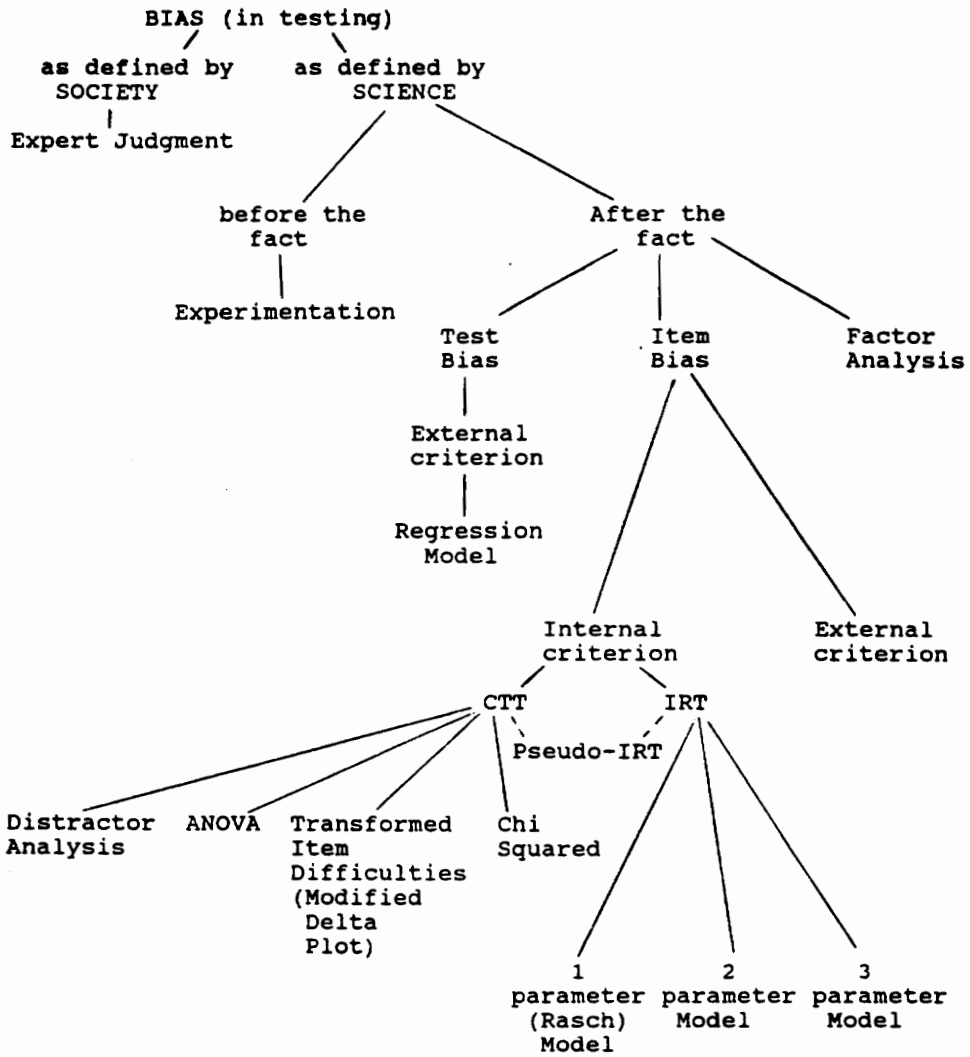


Figure 1) General Conceptualization of Bias in Testing

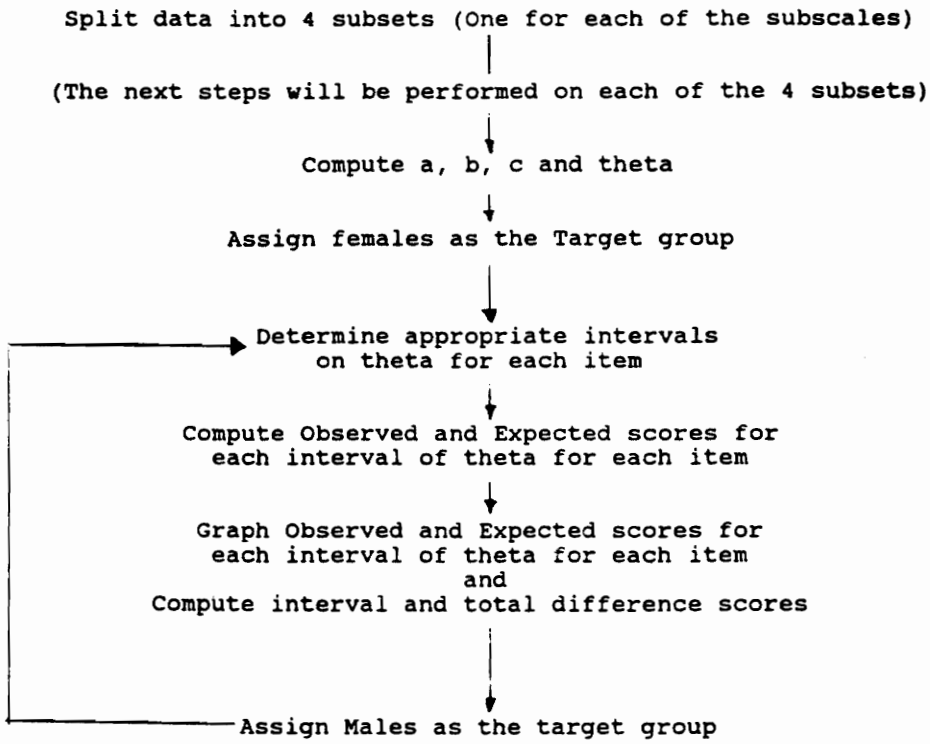


Figure 2) Plan for Pseudo-IRT item analysis of Myers-Briggs Type Indicator

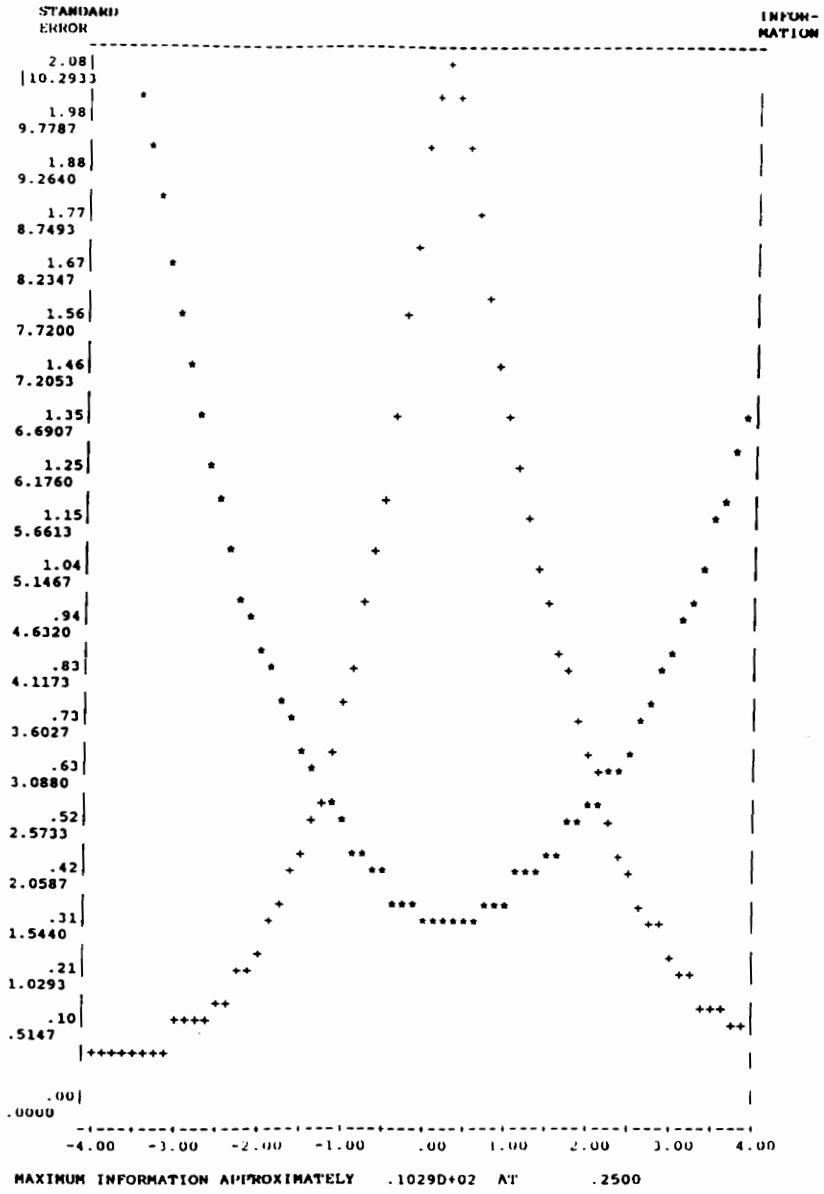


Figure 3: Test Information Function for EI subscale

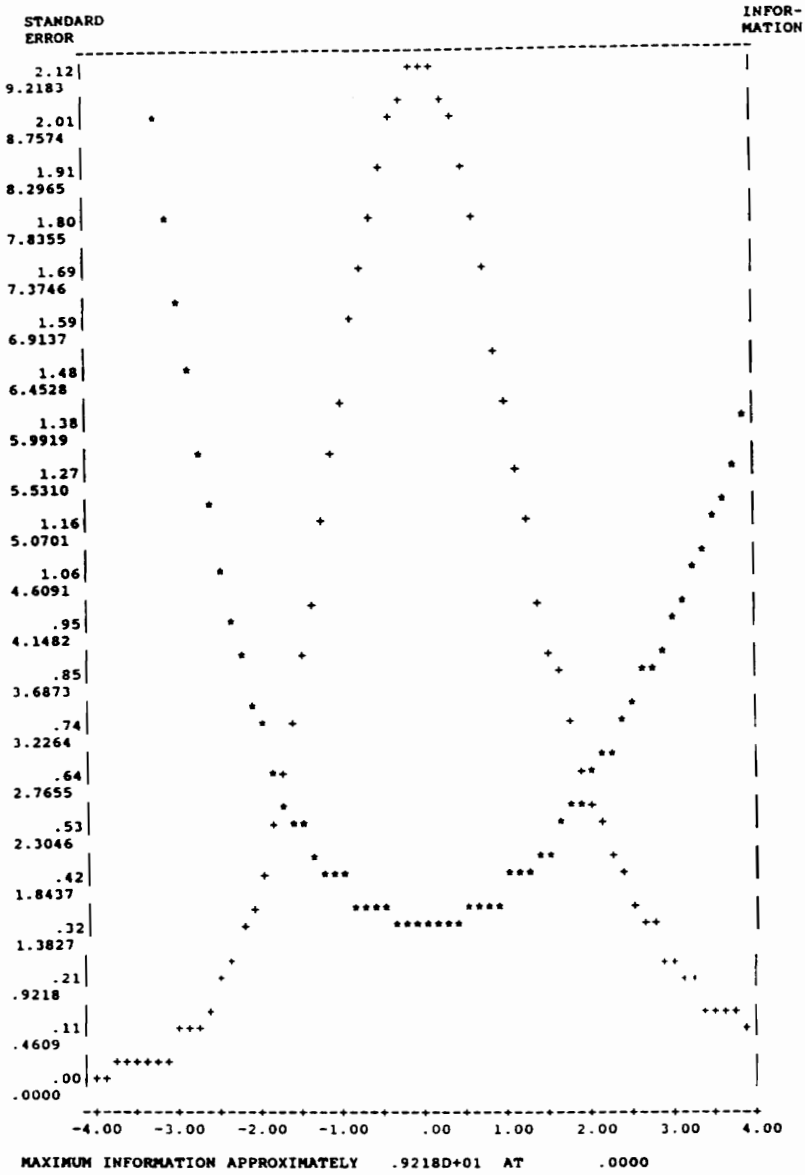


Figure 4: Test Information Function for SN subscale

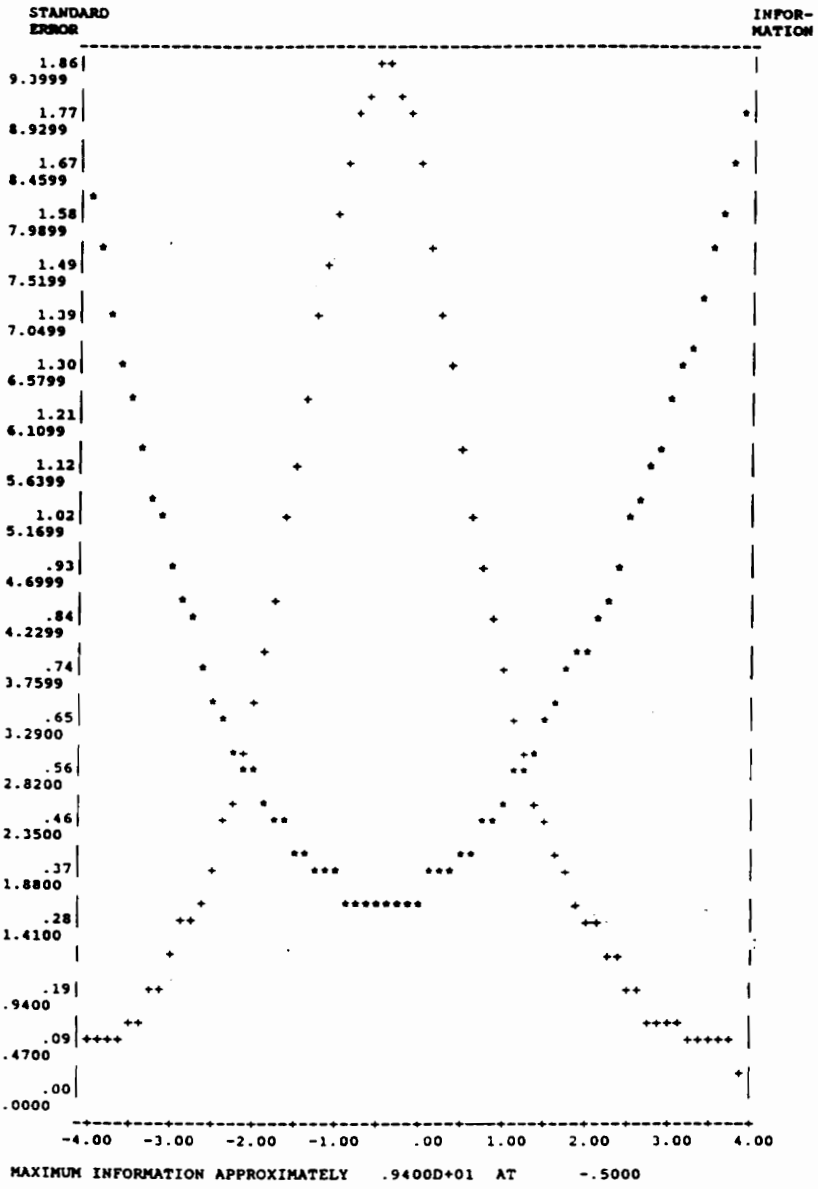


Figure 5: Test Information Function for TF subscale

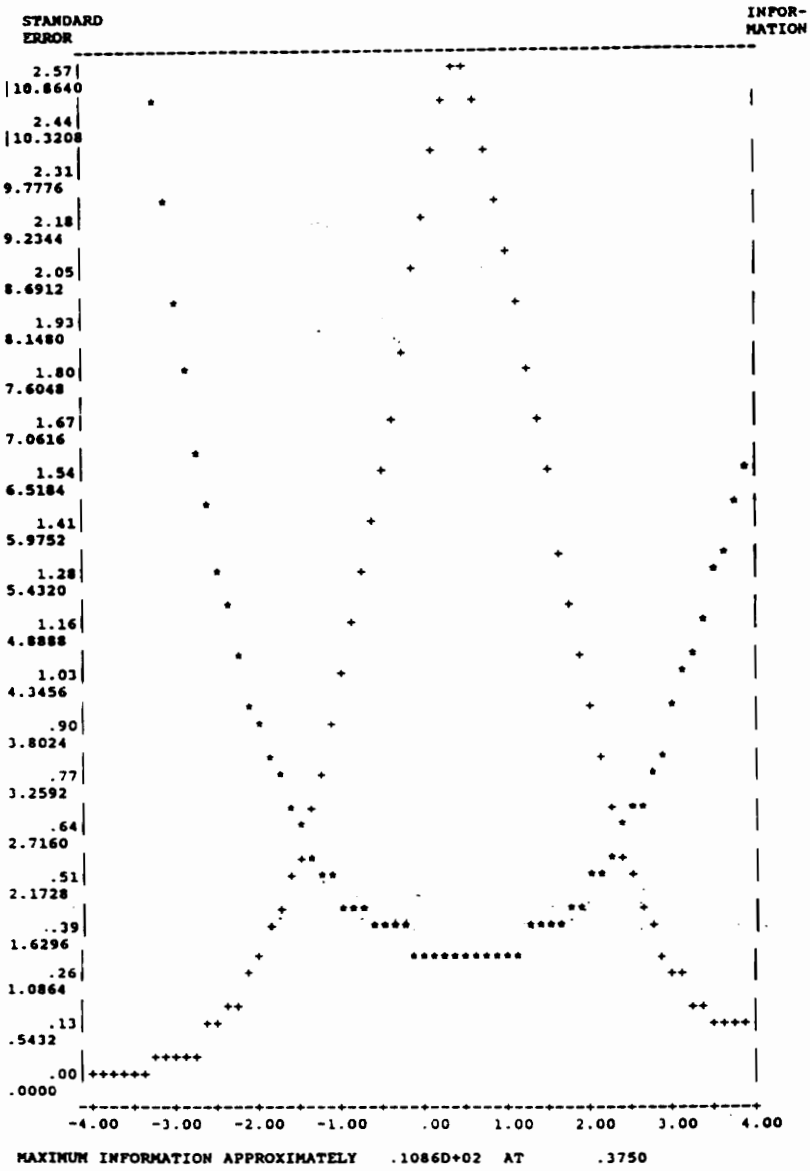


Figure 6: Test Information Function for JP subscale

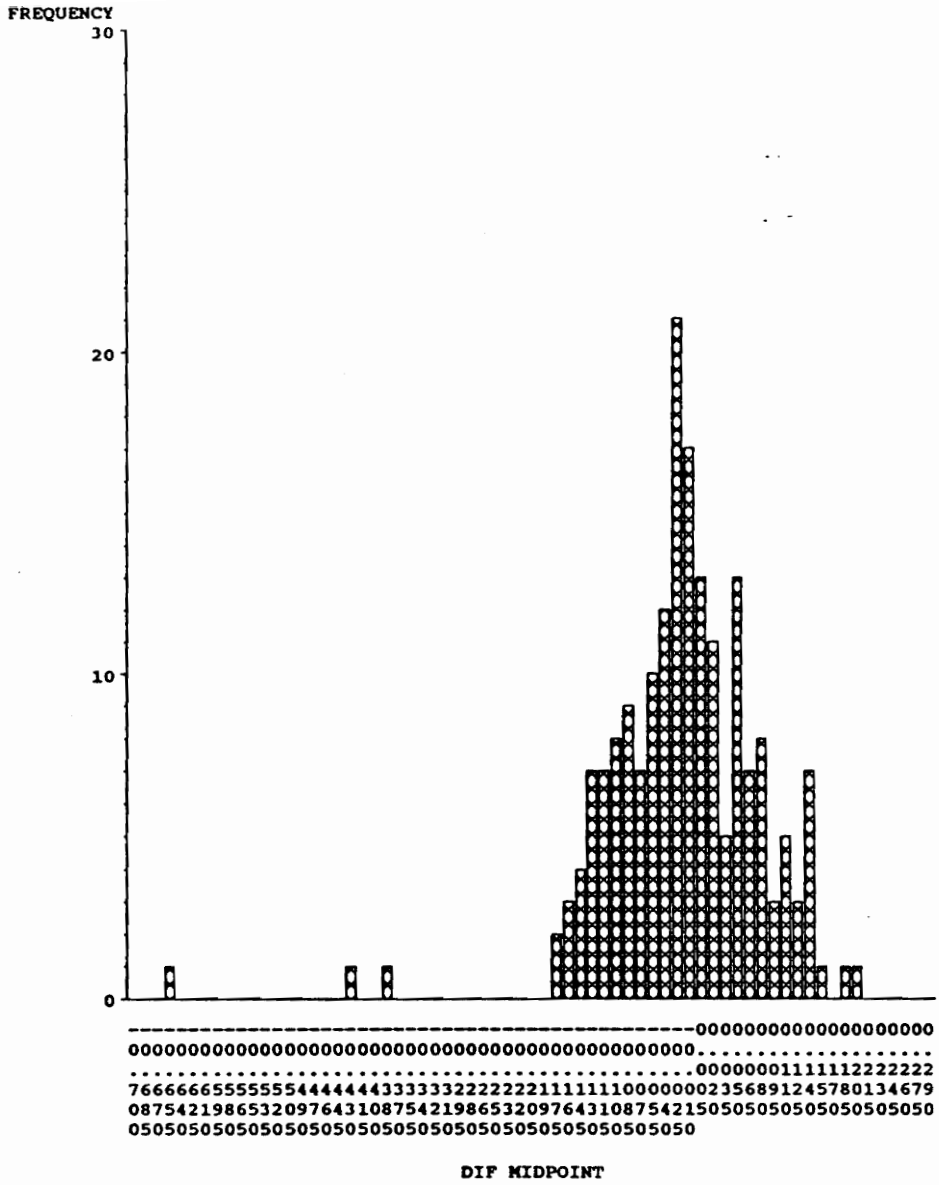
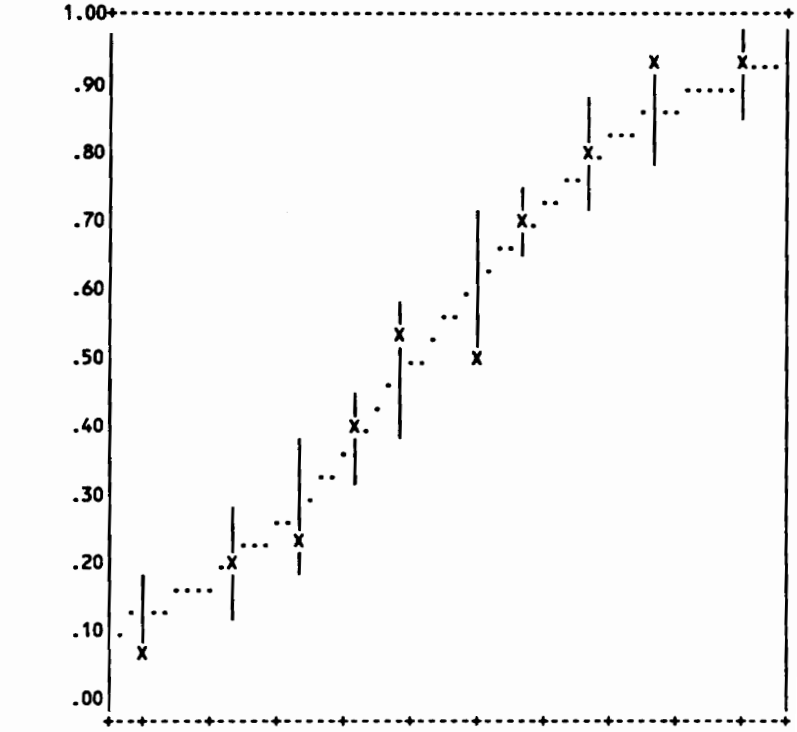


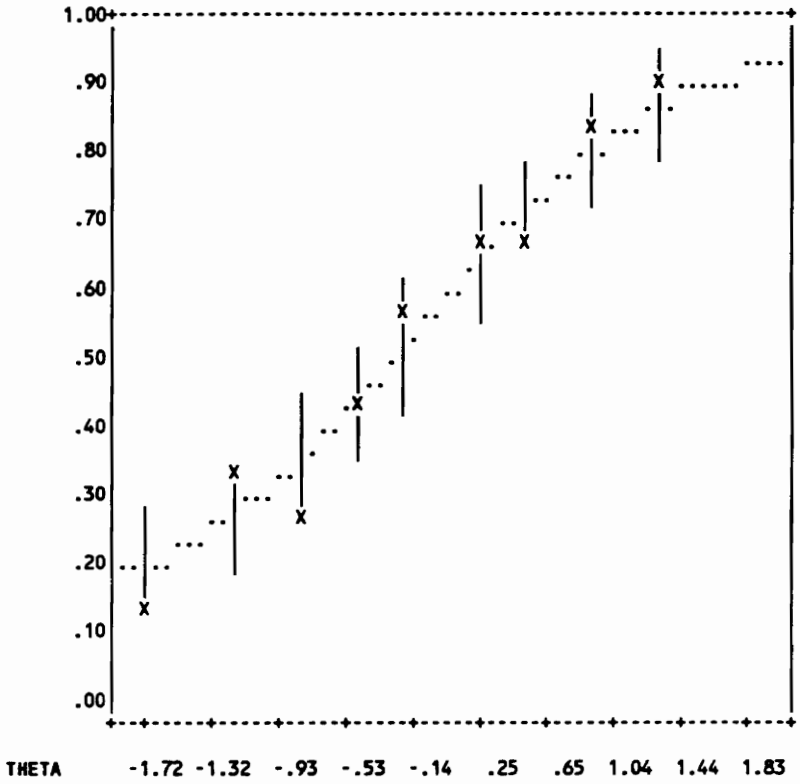
Figure 7: Frequency Distribution for Total Difference Scores

Appendix A

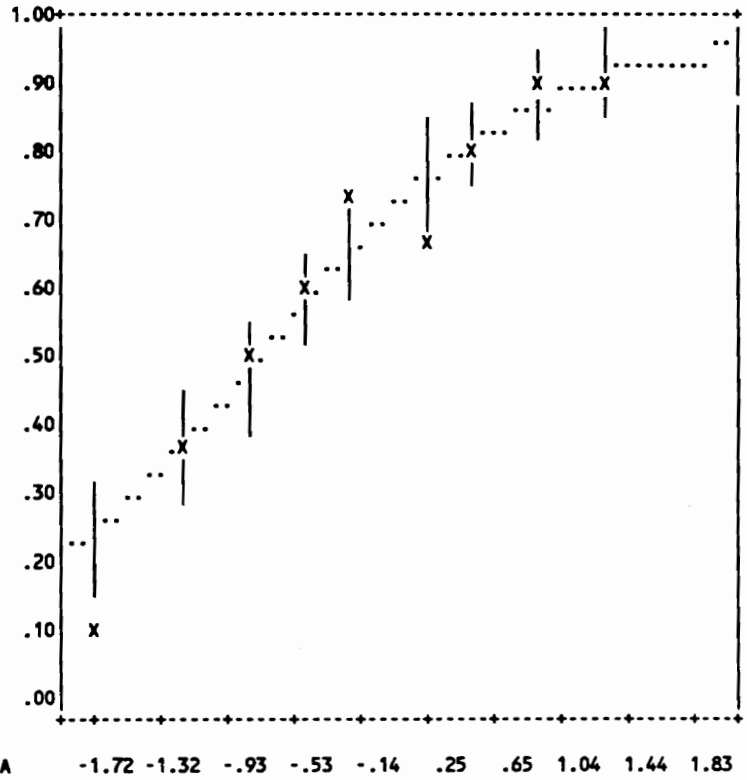
SUBTEST E1
ITEM M6 PROB< .0053



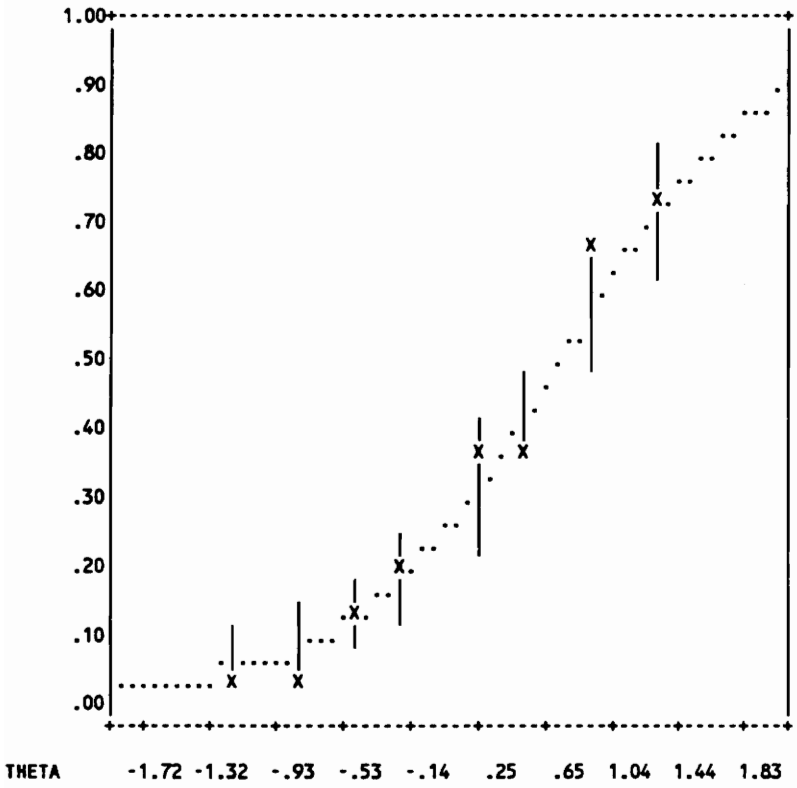
SUBTEST EI
ITEM M15 PROB< .0008



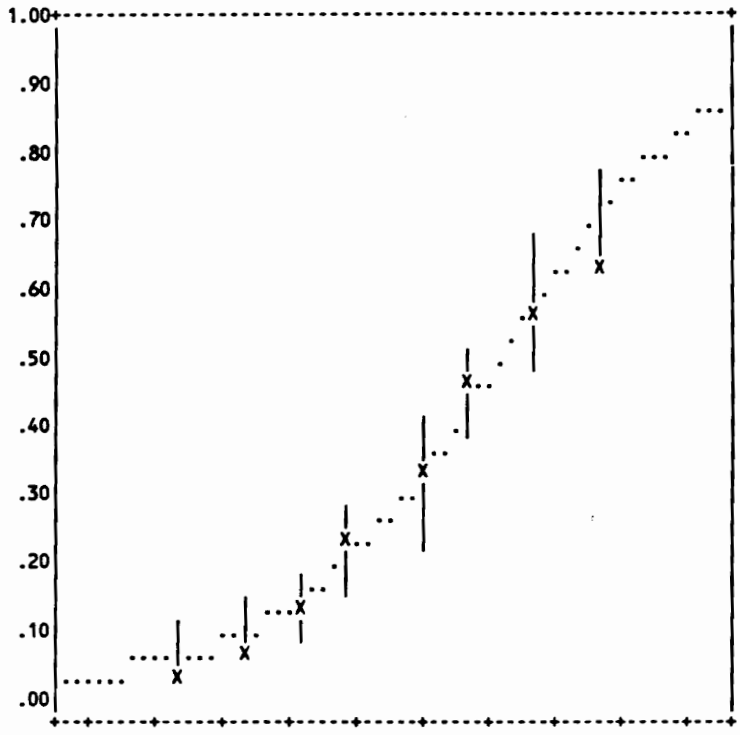
SUBTEST EI
ITEM M19 PROB< .0000



SUBTEST EI
ITEM M25 PROB< .0312

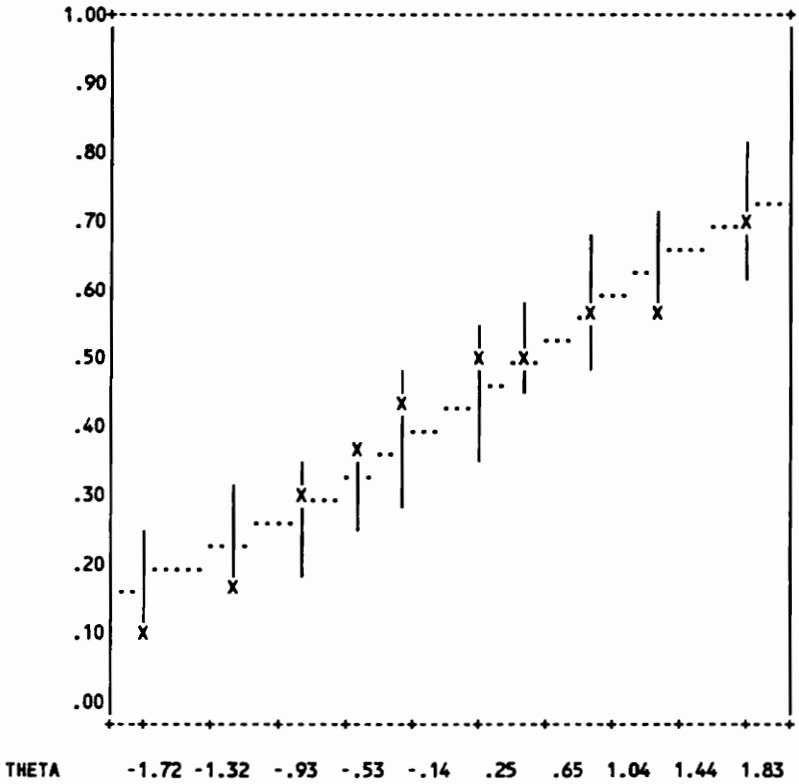


SUBTEST EI
ITEM M33 PROB< .1096

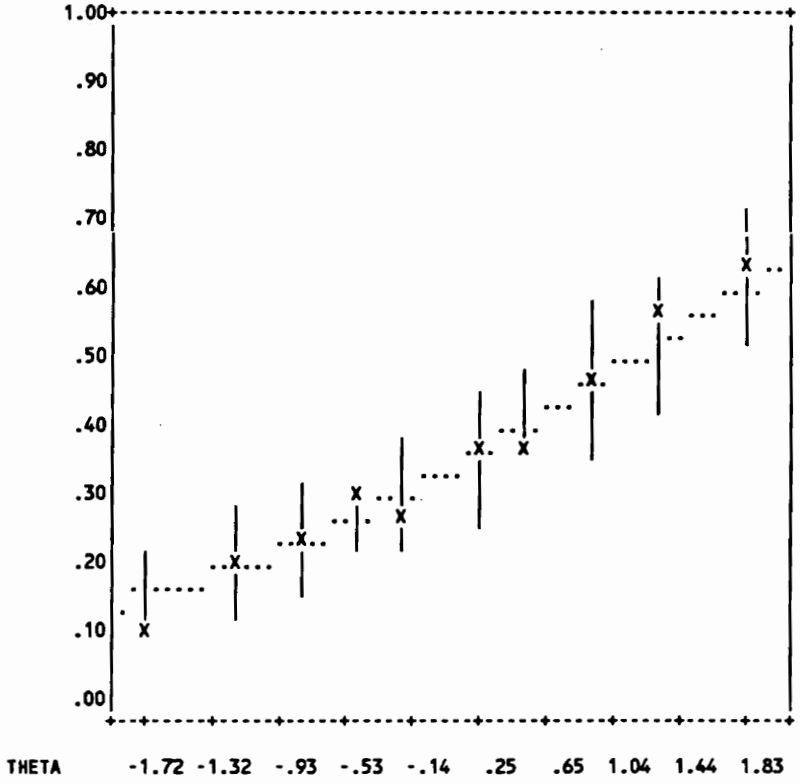


THETA -1.72 -1.32 -.93 -.53 -.14 .25 .65 1.04 1.44 1.83

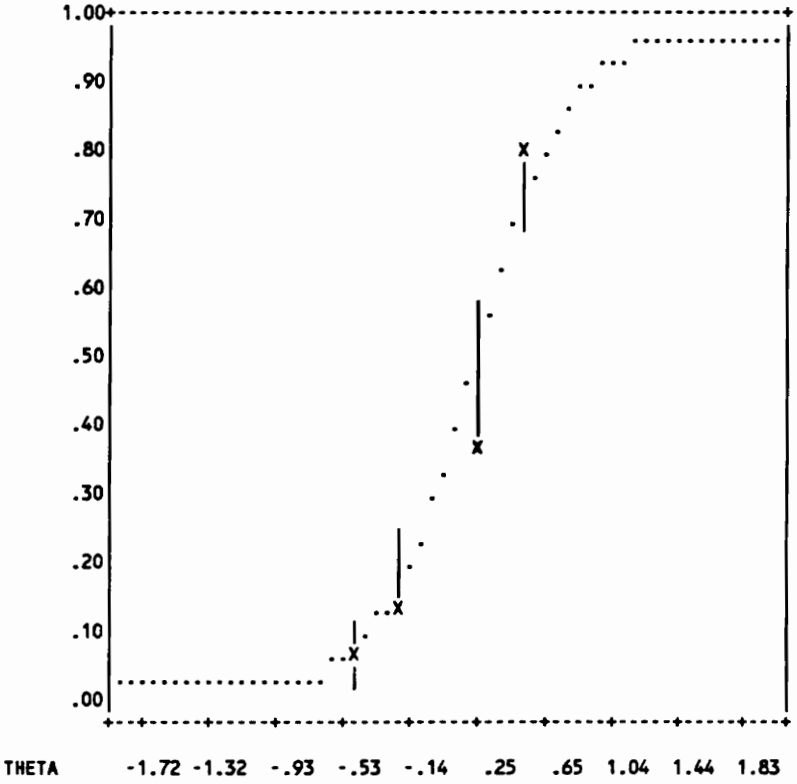
SUBTEST EI
ITEM M41 PROB< .0068



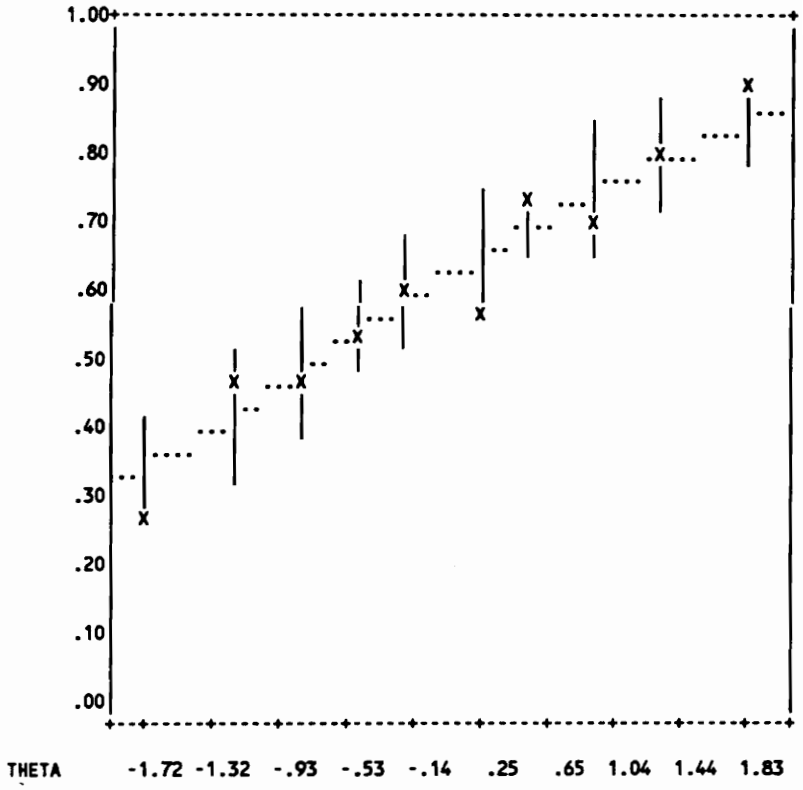
SUBTEST EI
ITEM M47 PROB< .1406



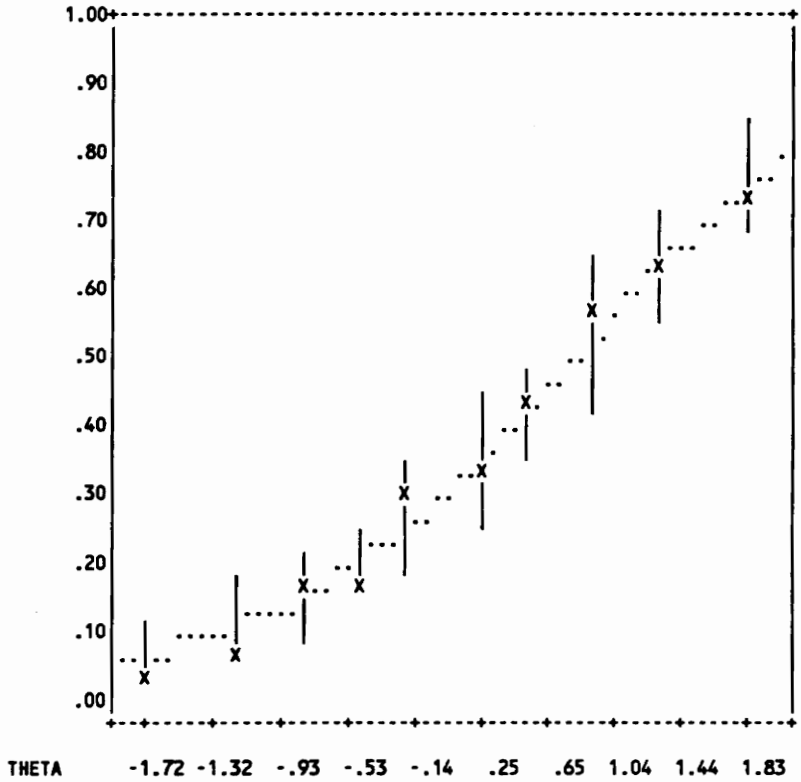
SUBTEST EI
ITEM M50 PROB< .0001



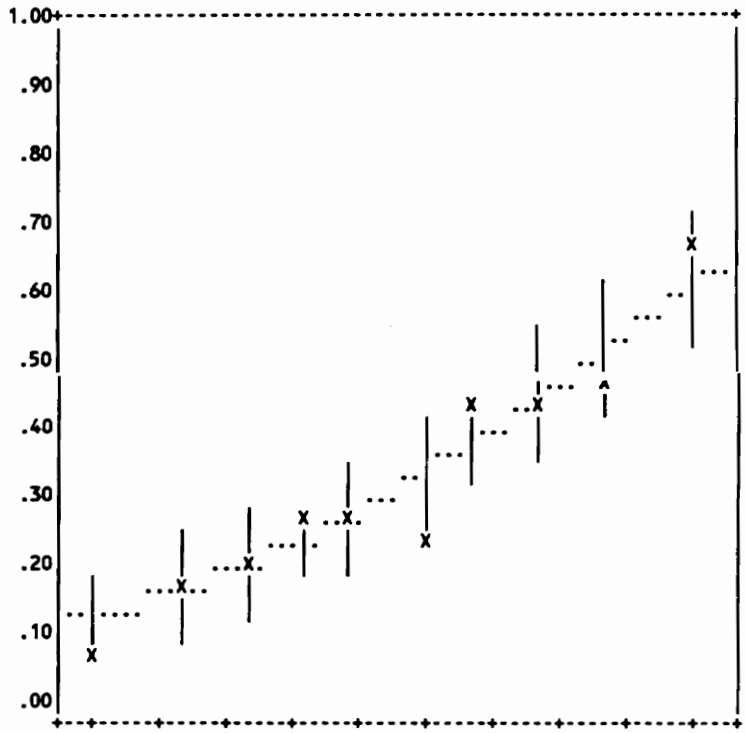
SUBTEST EI
ITEM M58 PROB< .0311



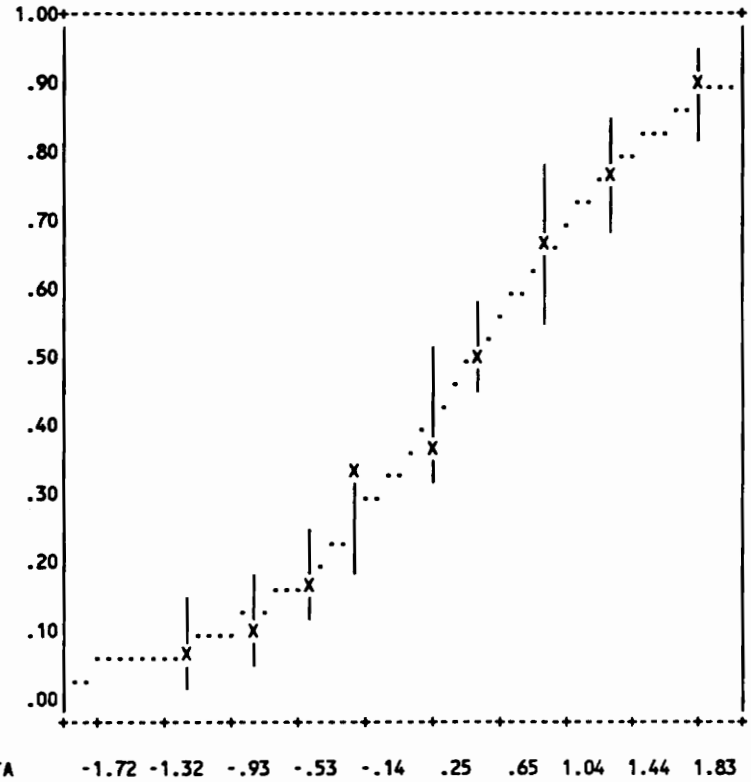
SUBTEST EI
ITEM M66 PROB< .1816



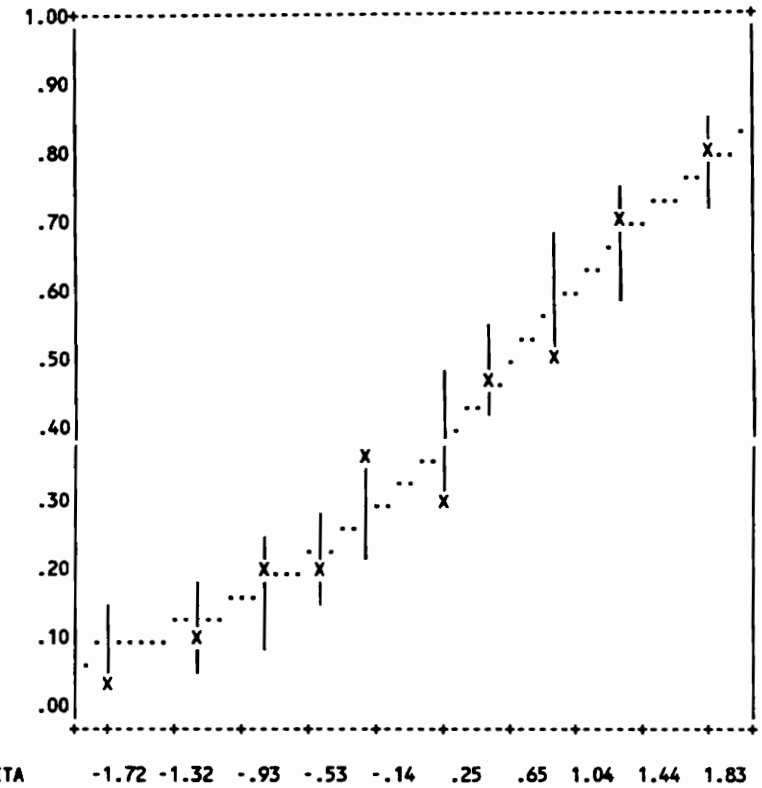
SUBTEST EI
ITEM M77 PROB< .0039



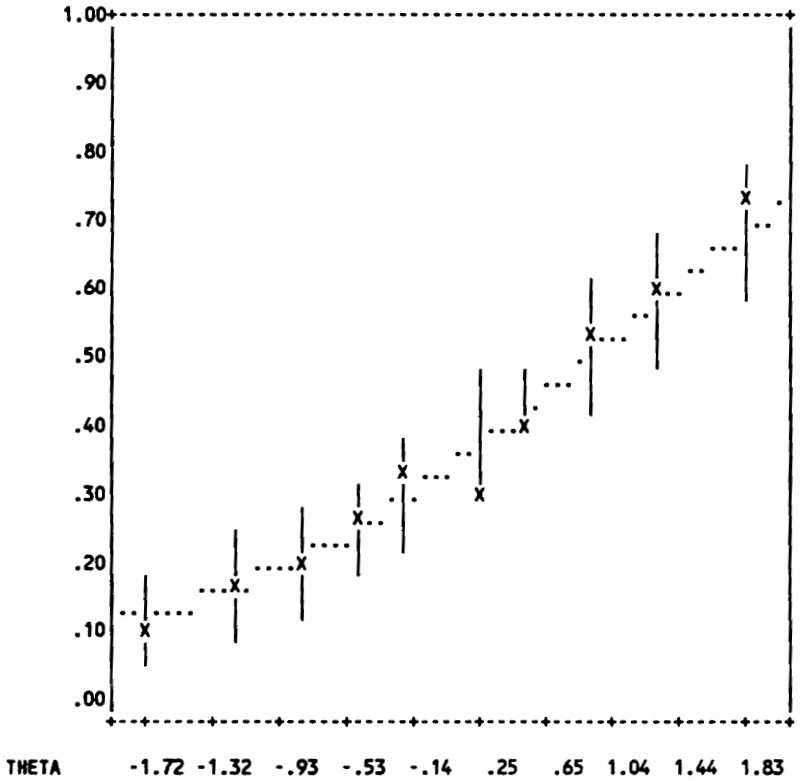
SUBTEST EI
ITEM M87 PROB< .2437



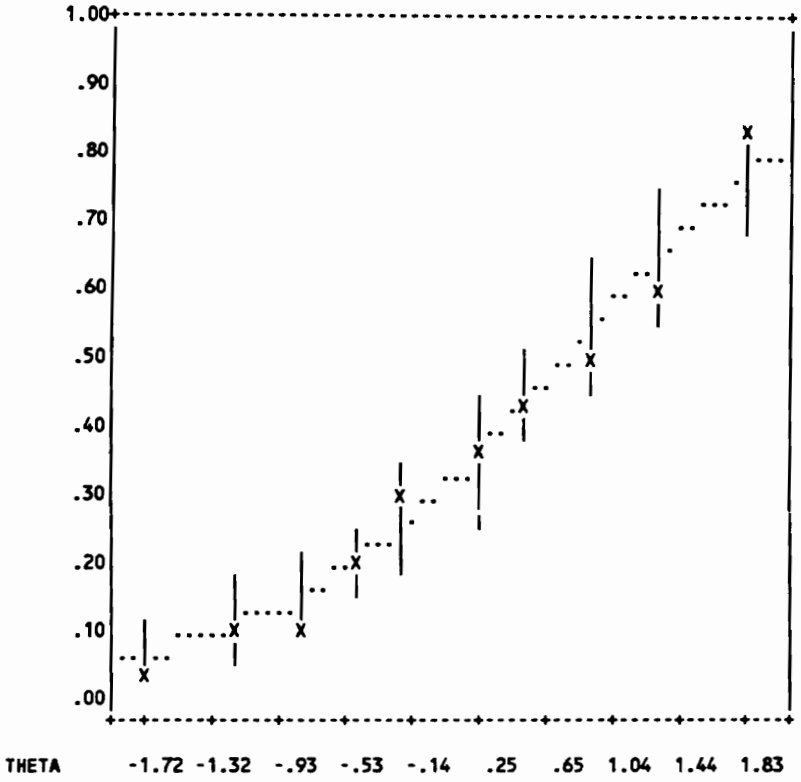
SUBTEST EI
ITEM M92 PROB< .0471



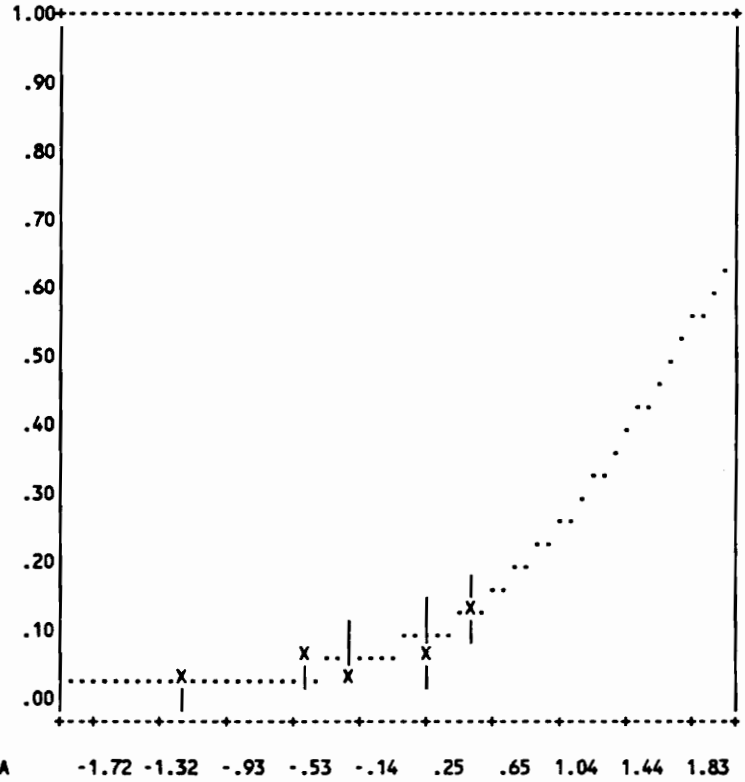
SUBTEST EI
ITEM M95 PROB< .2182



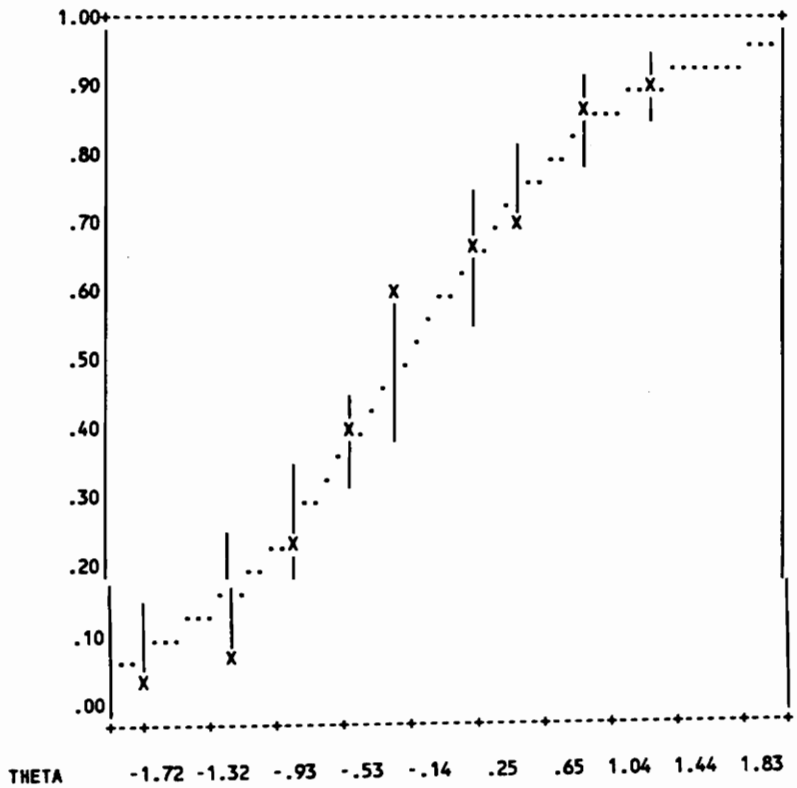
SUBTEST EI
ITEM M106 PROB< .0387



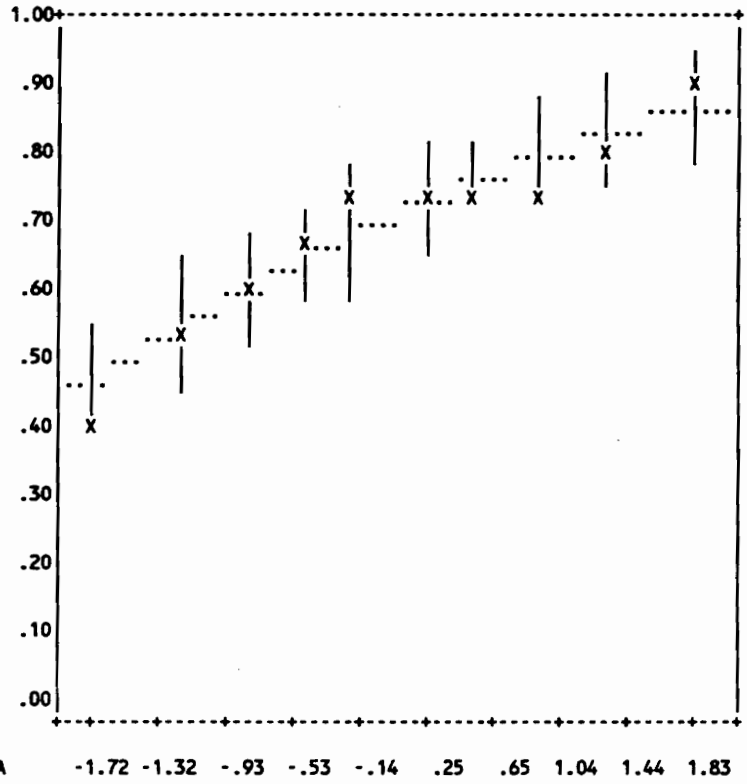
SUBTEST EI
ITEM M116 PROB< .2324



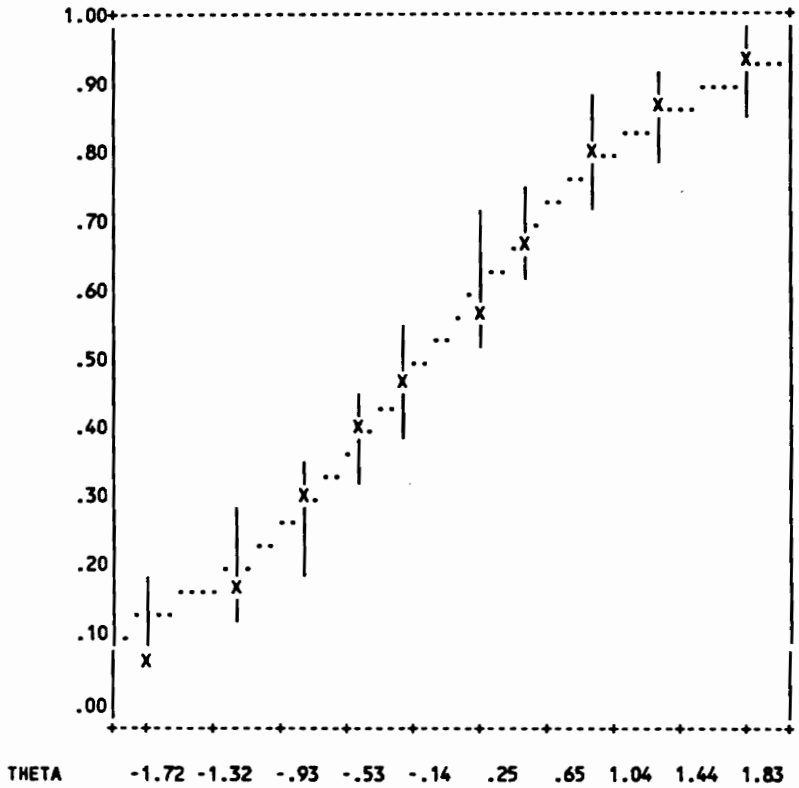
SUBTEST EI
ITEM M126 PROB< .0000



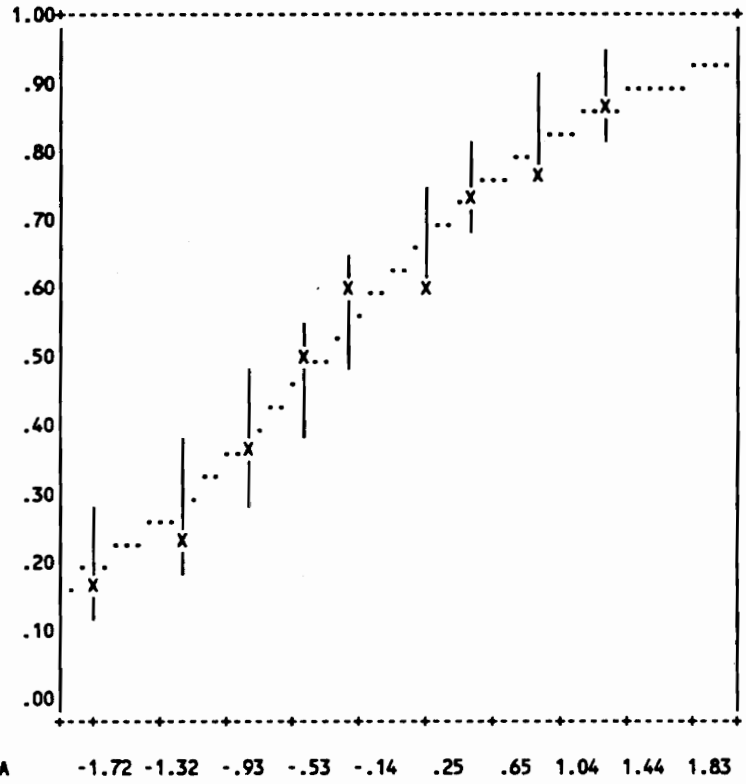
SUBTEST EI
ITEM M129 PROB< .1500



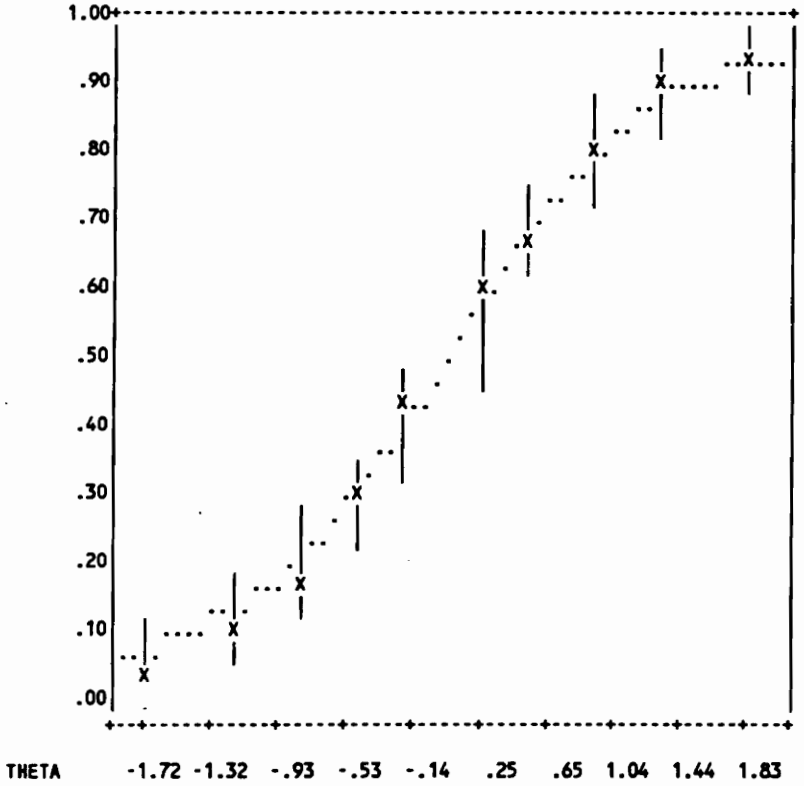
SUBTEST EI
ITEM M134 PROB< .0601



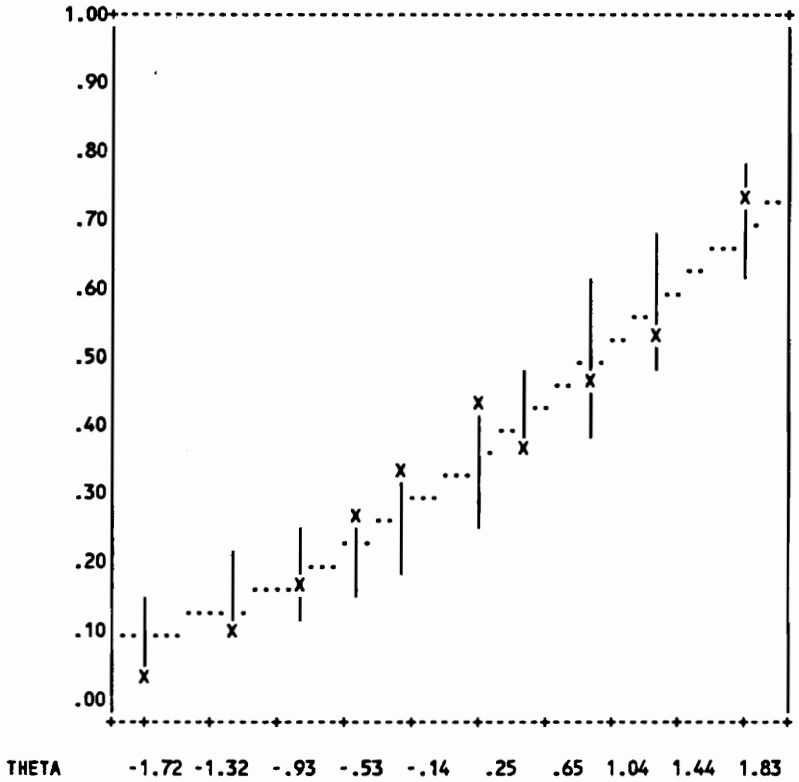
SUBTEST EI
ITEM M138 PROB< .2383



SUBTEST EI
ITEM M148 PROB< .1673

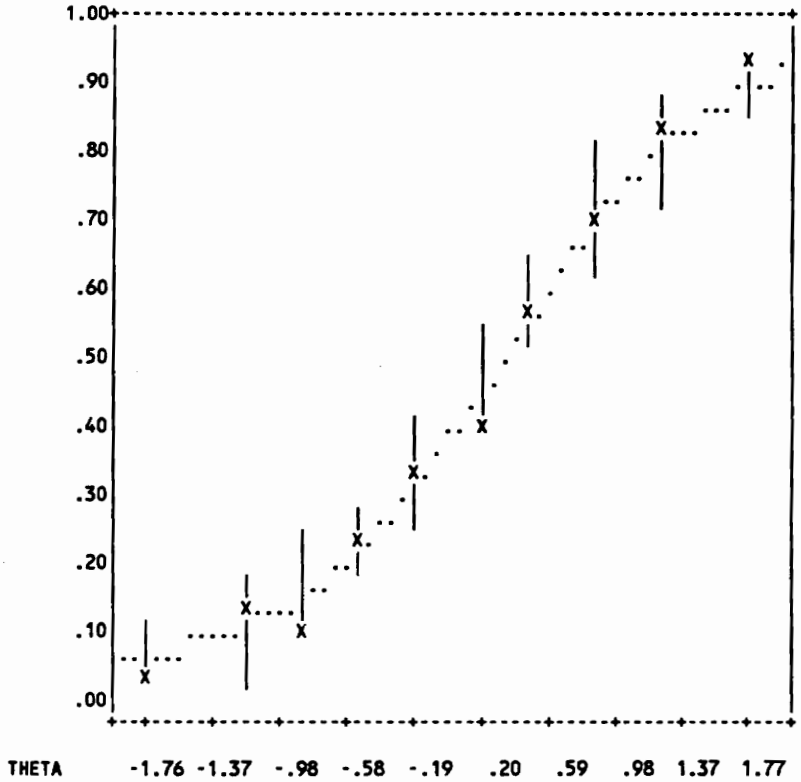


SUBTEST EI
ITEM M160 PROB< .0000

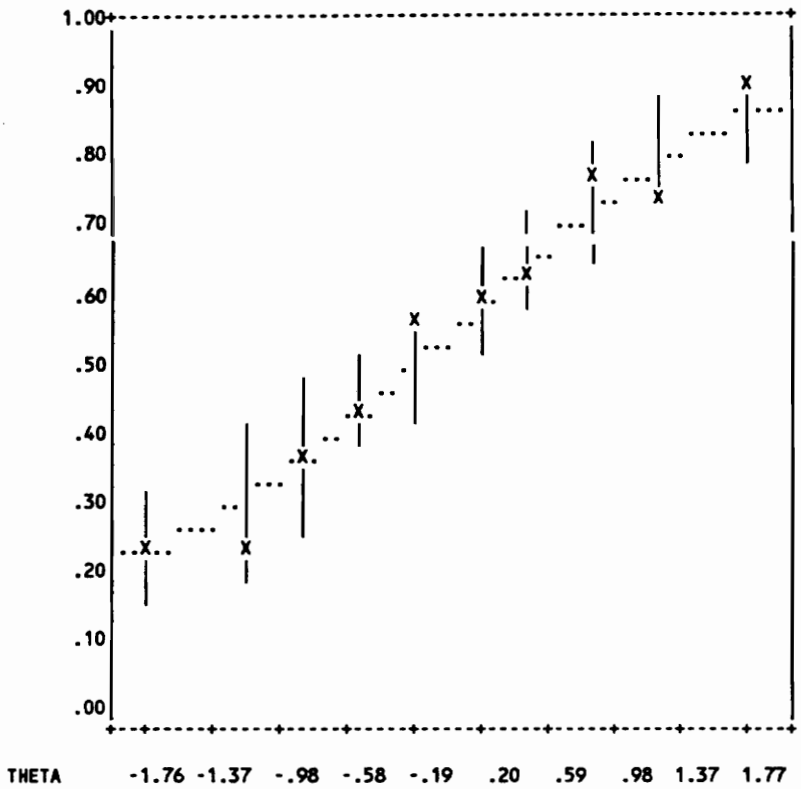


Appendix B

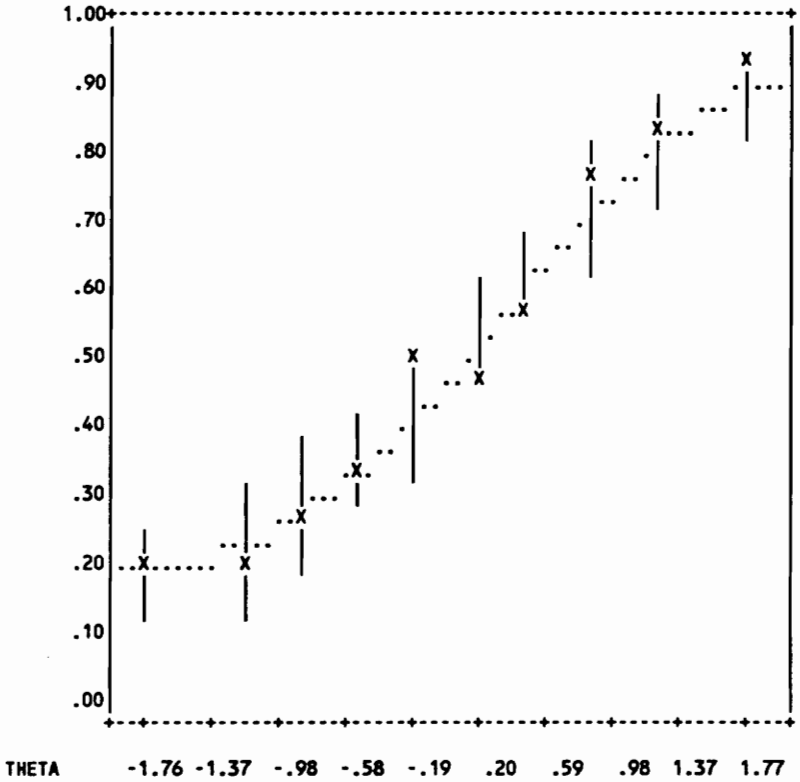
SUBTEST SN
ITEM M2 PROB< .0491



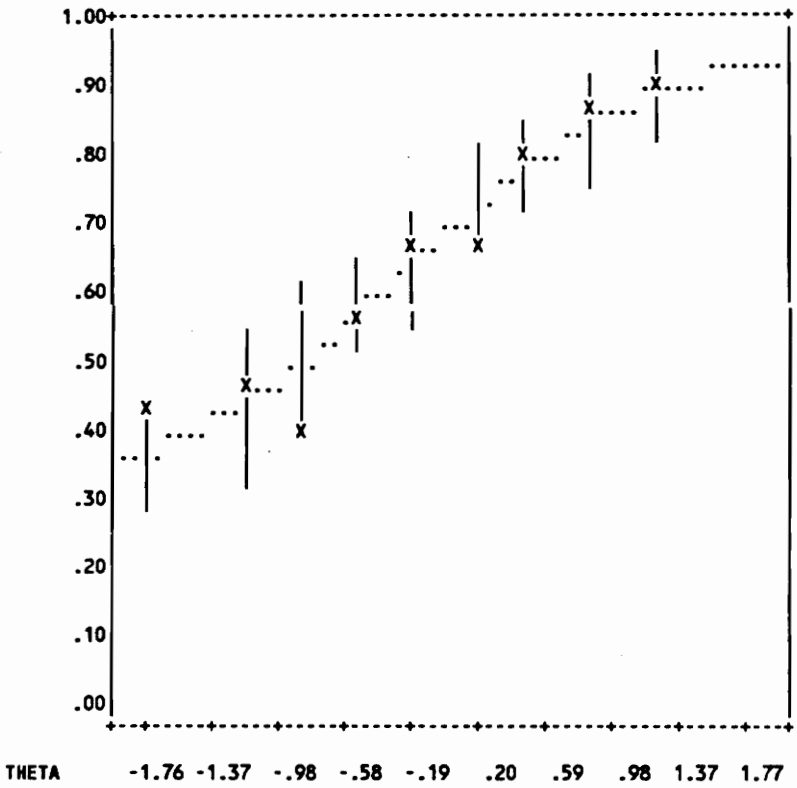
SUBTEST SN
ITEM M11 PROB< .1463



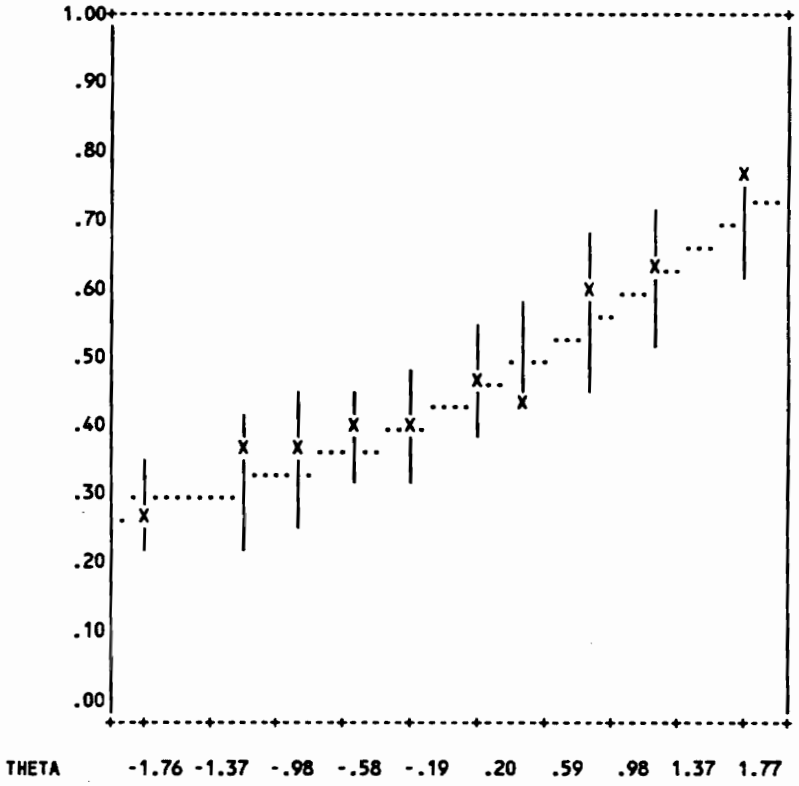
SUBTEST SW
ITEM M17 PROB< .0516



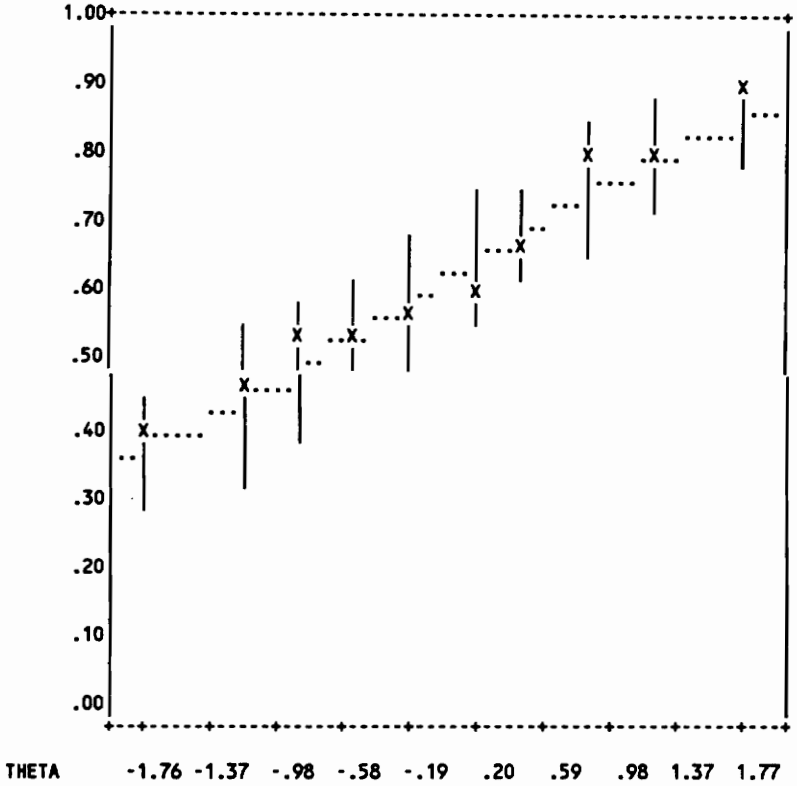
SUBTEST SN
ITEM M37 PROB< .0515



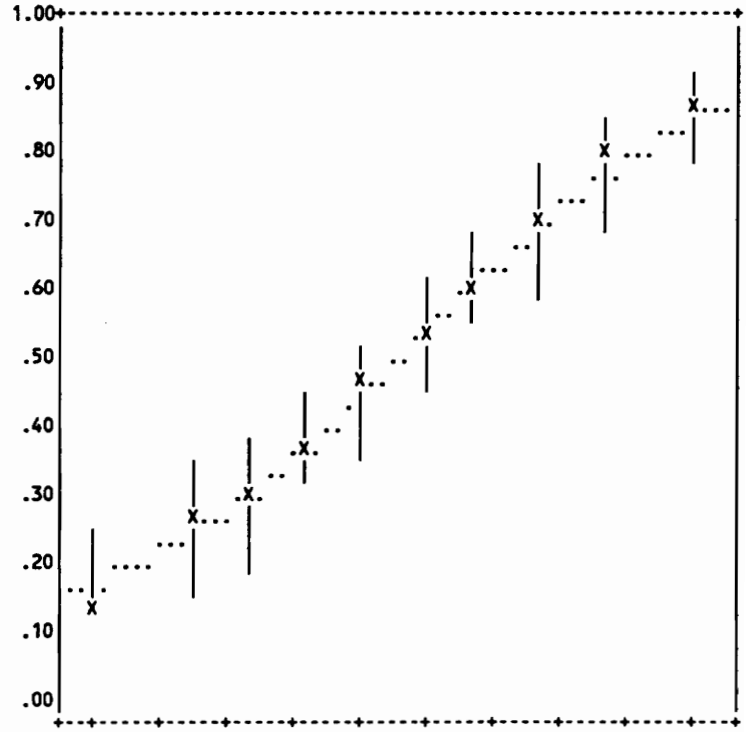
SUBTEST SN
ITEM M53 PROB< .1882



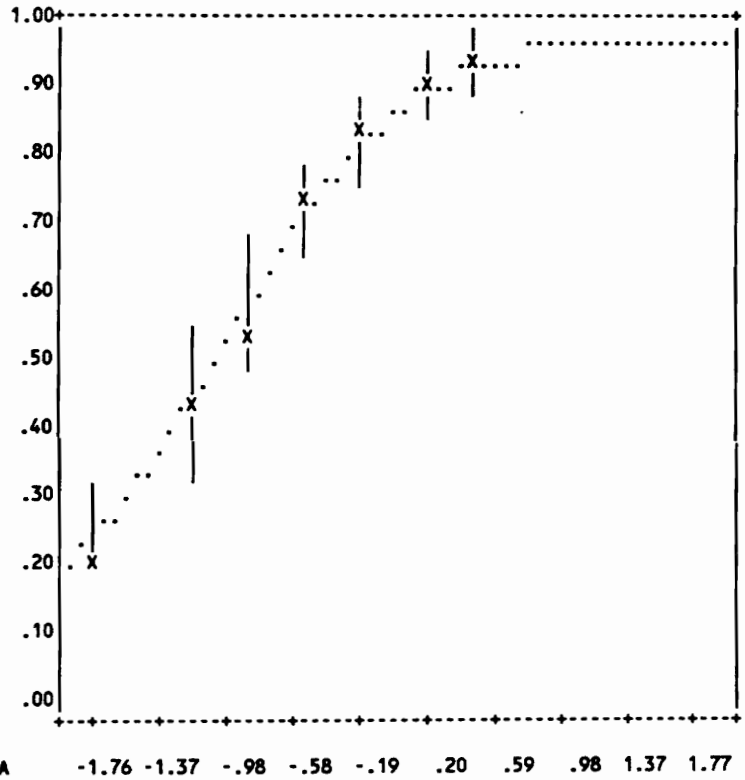
SUBTEST SN
ITEM M64 PROB< .2460



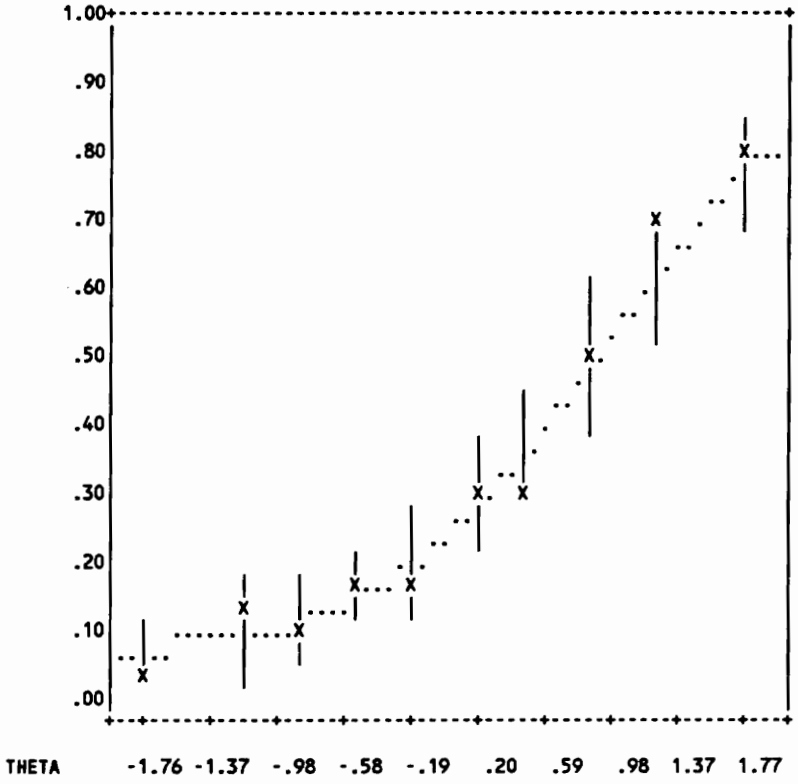
SUBTEST SN
ITEM M70 PROB< .9249



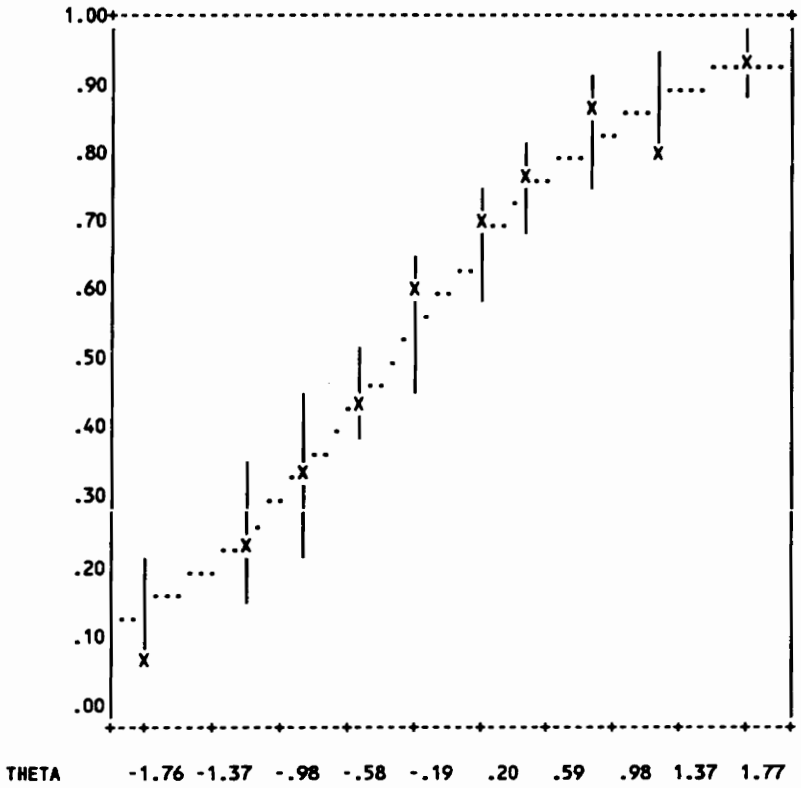
SUBTEST SN
ITEM M73 PROB< .5854



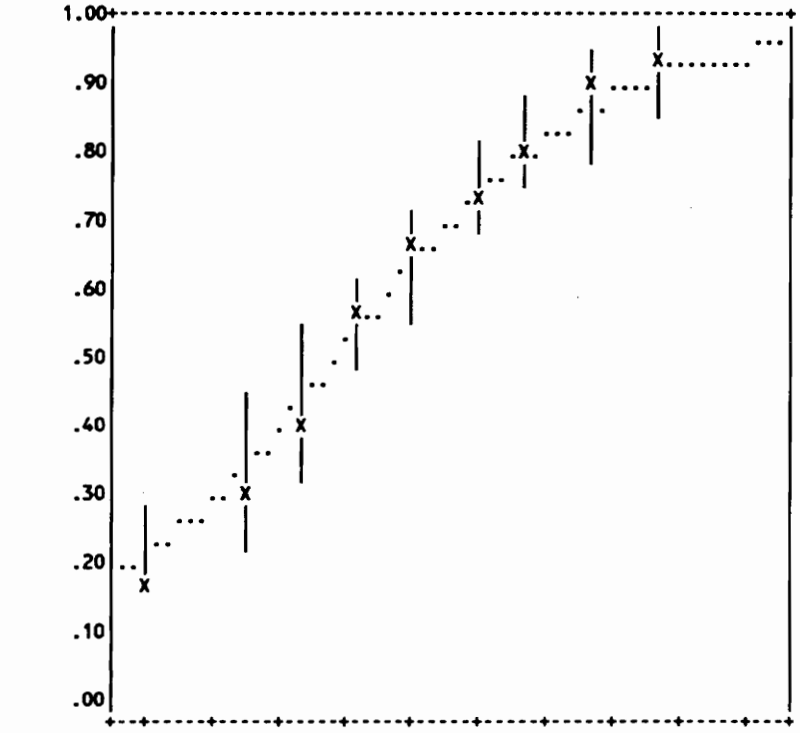
SUBTEST SN
ITEM M76 PROB< .0569



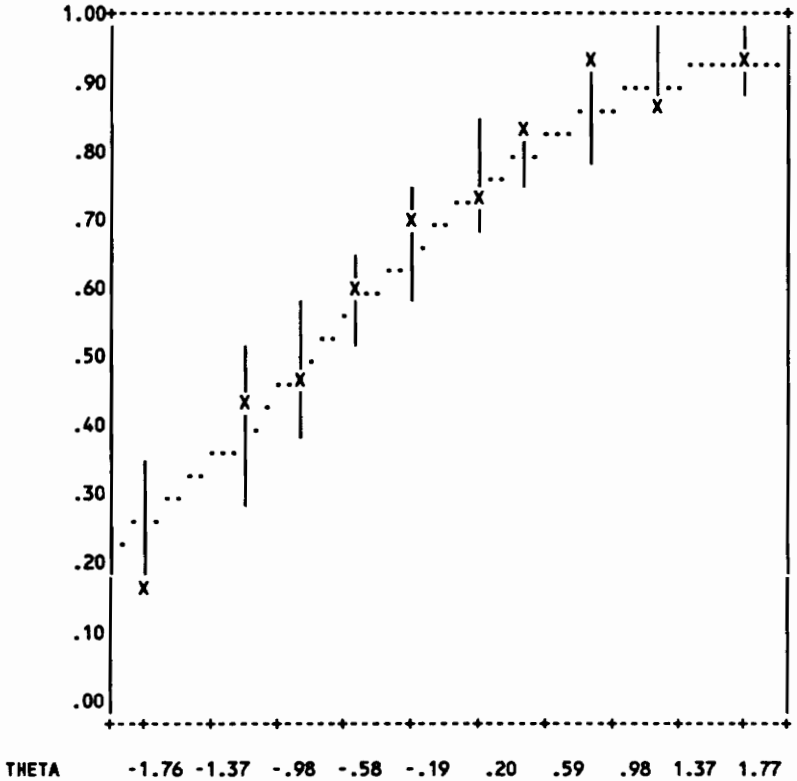
SUBTEST SN
ITEM M78 PROB< .0001



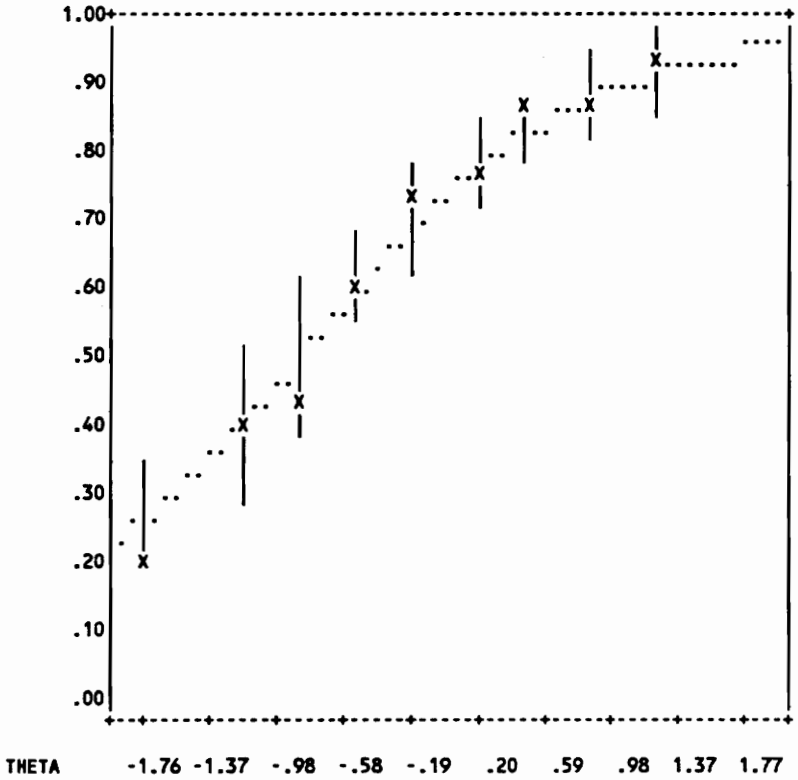
SUBTEST SN
ITEM M88 PROB< .5084



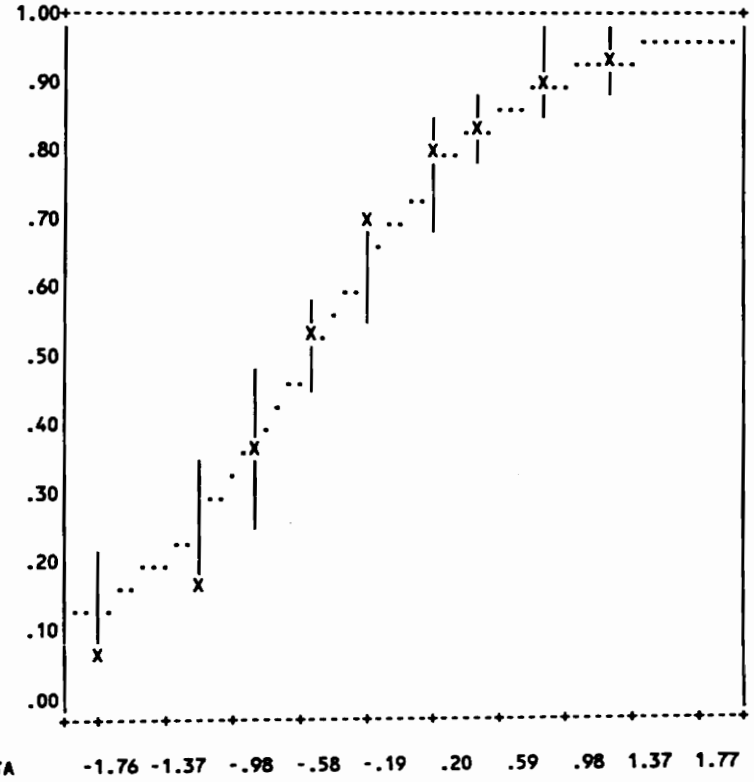
SUBTEST SN
ITEM M90 PROB< .0066



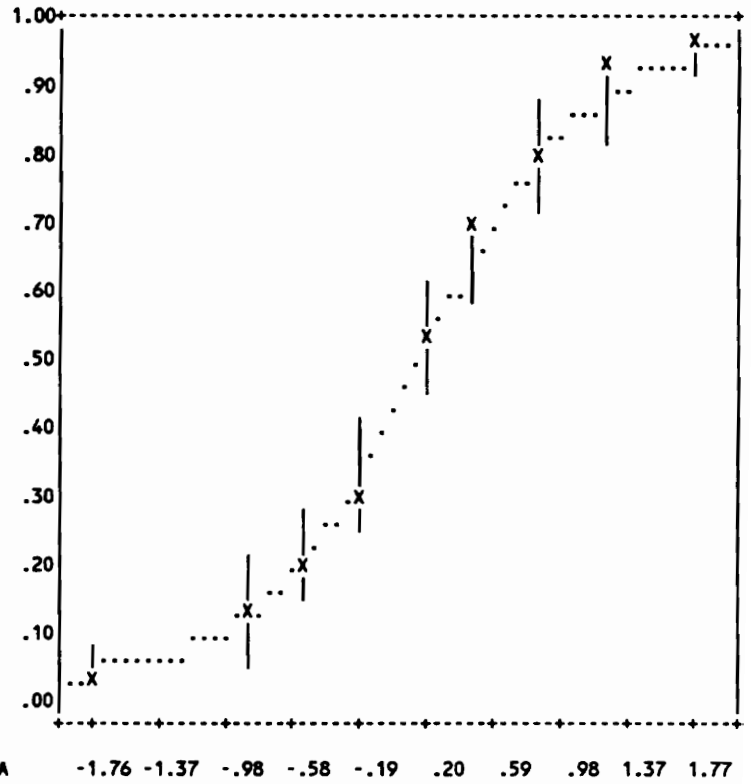
SUBTEST SN
ITEM M98 PROB< .0880



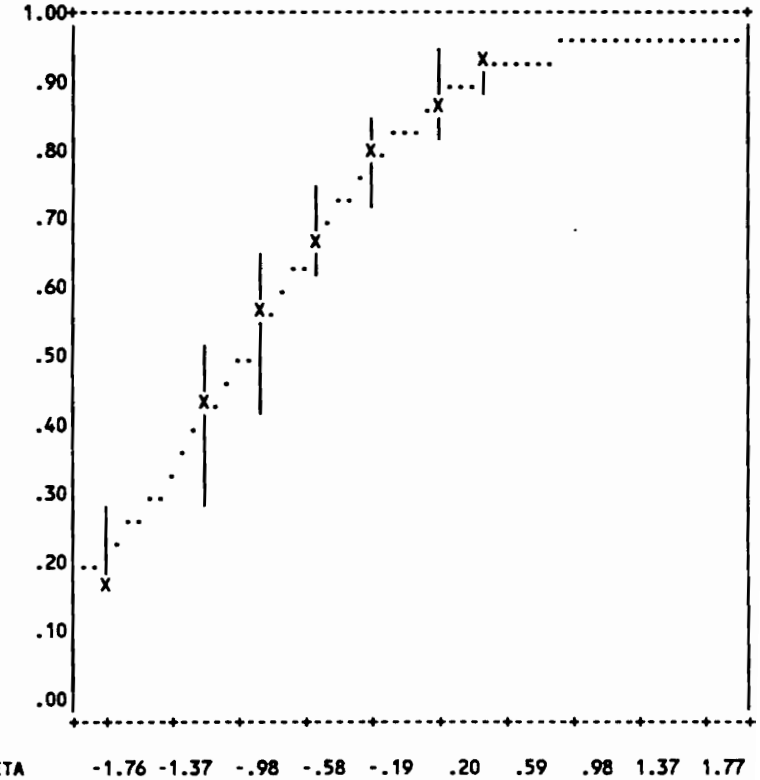
SUBTEST SW
ITEM M102 PROB< .0141



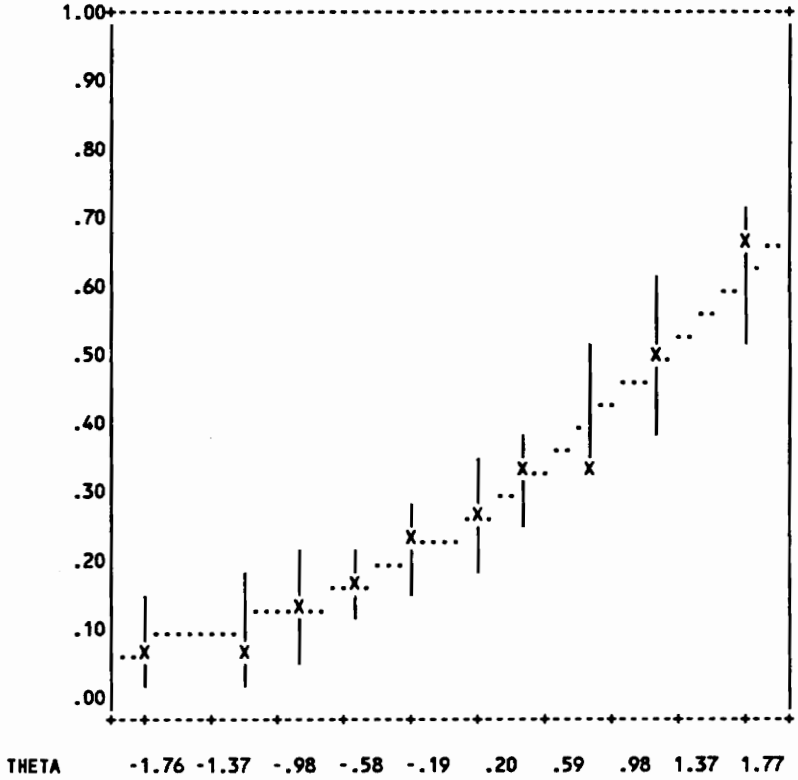
SUBTEST SN
ITEM M104 PROB< .0497



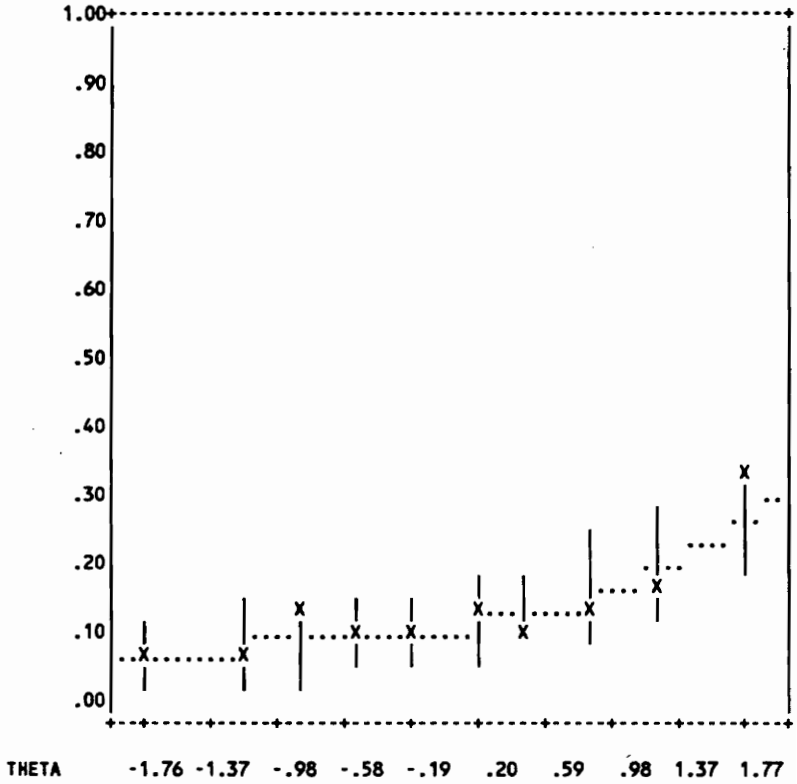
SUBTEST SN
ITEM M107 PROB< .0646



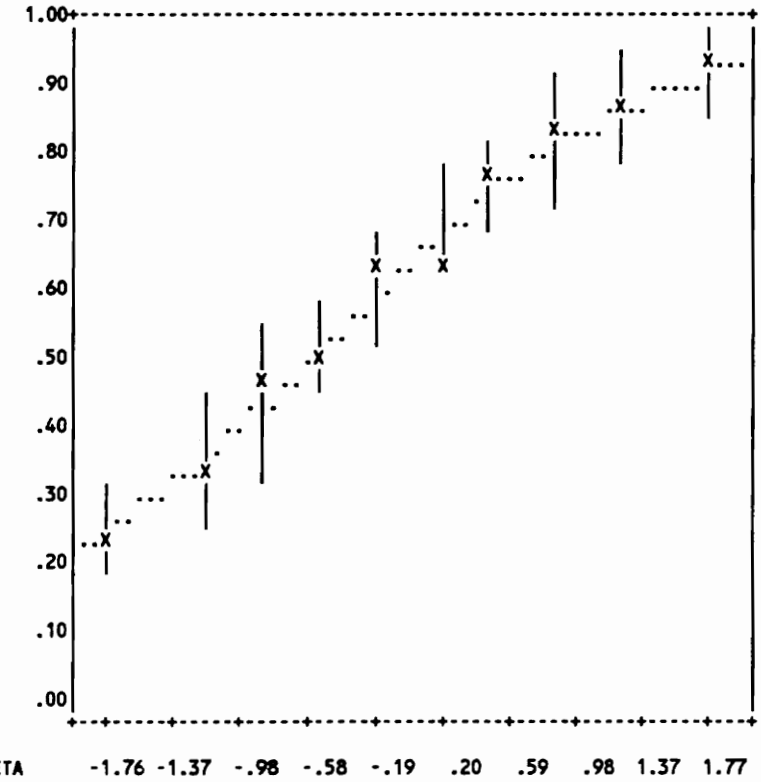
SUBTEST SN
ITEM M112 PROB< .2945



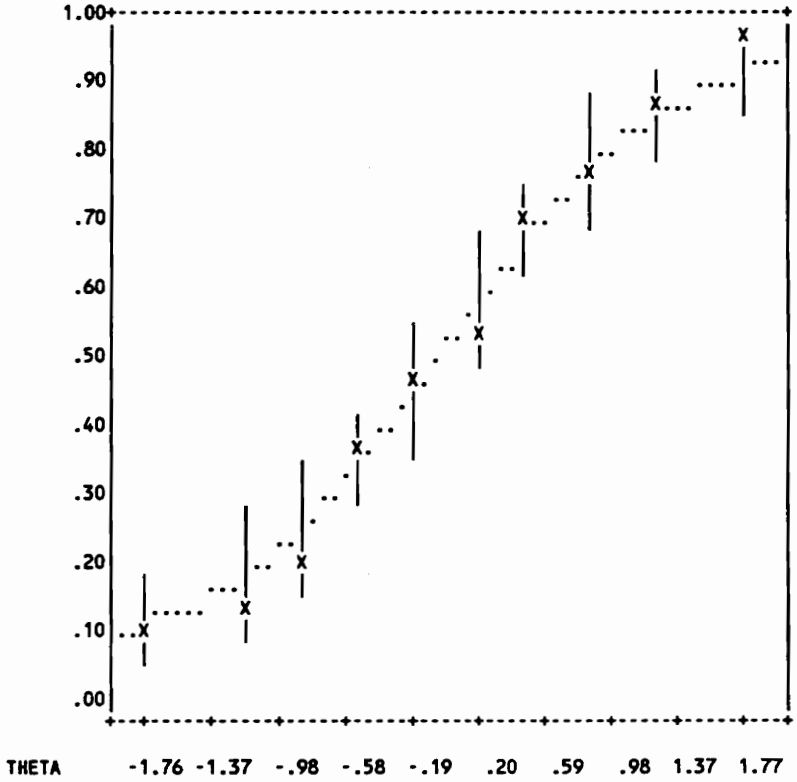
SUBTEST SN
ITEM M115 PROB< .1886



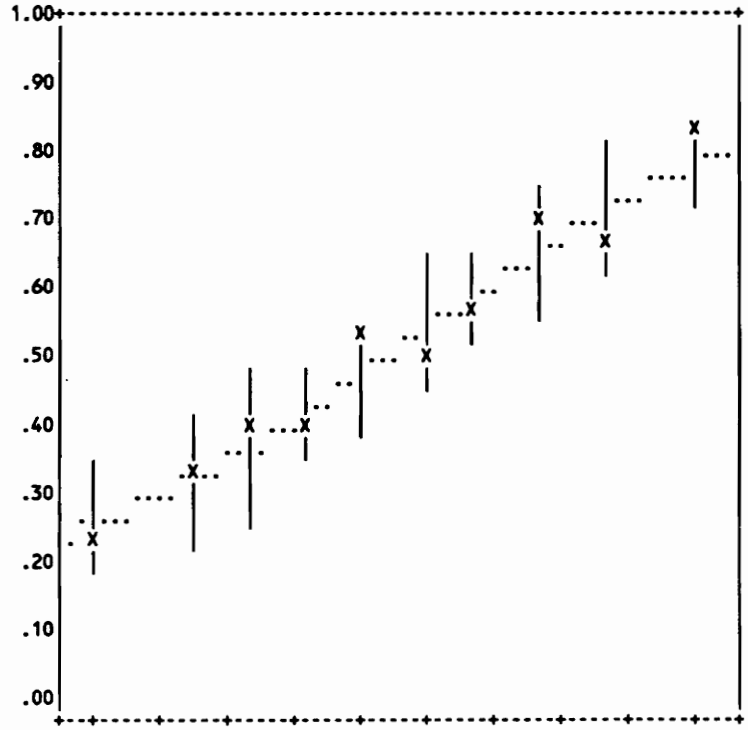
SUBTEST SW
ITEM M117 PROB< .3818



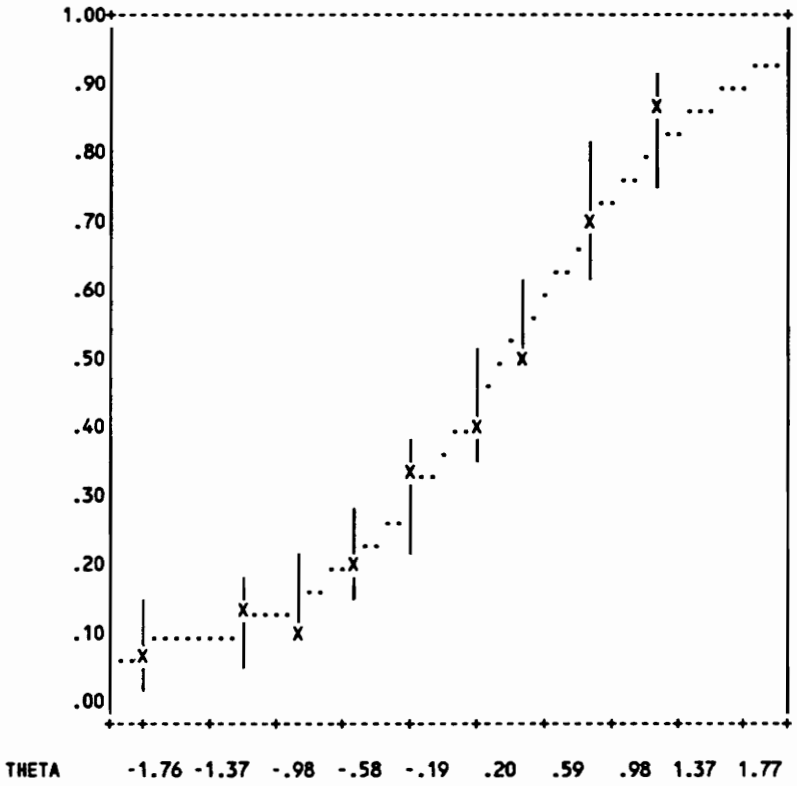
SUBTEST SN
ITEM M119 PROB< .0680



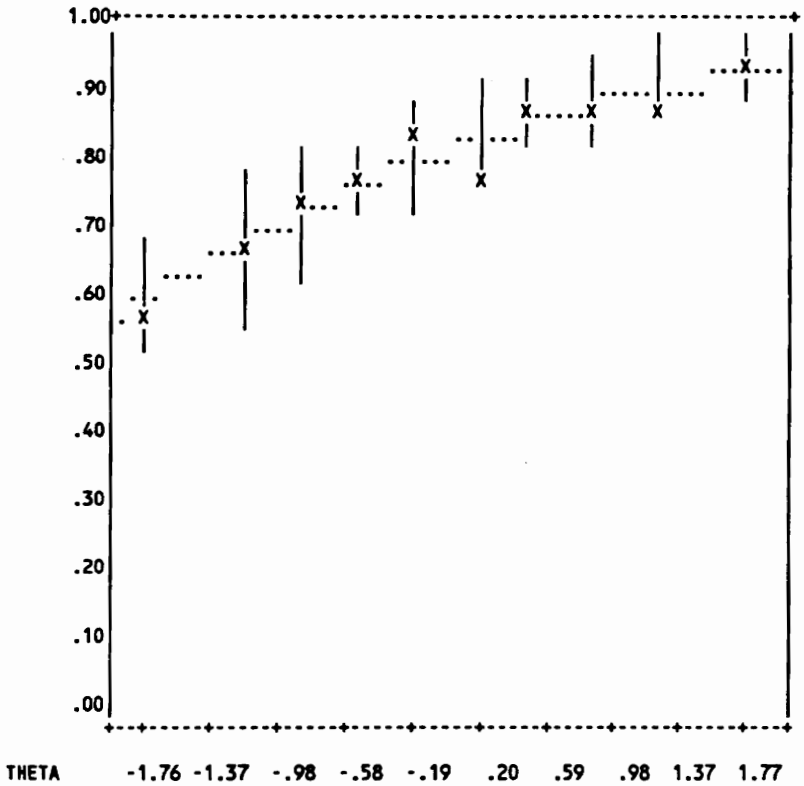
SUBTEST SN
ITEM M121 PROB< .1708



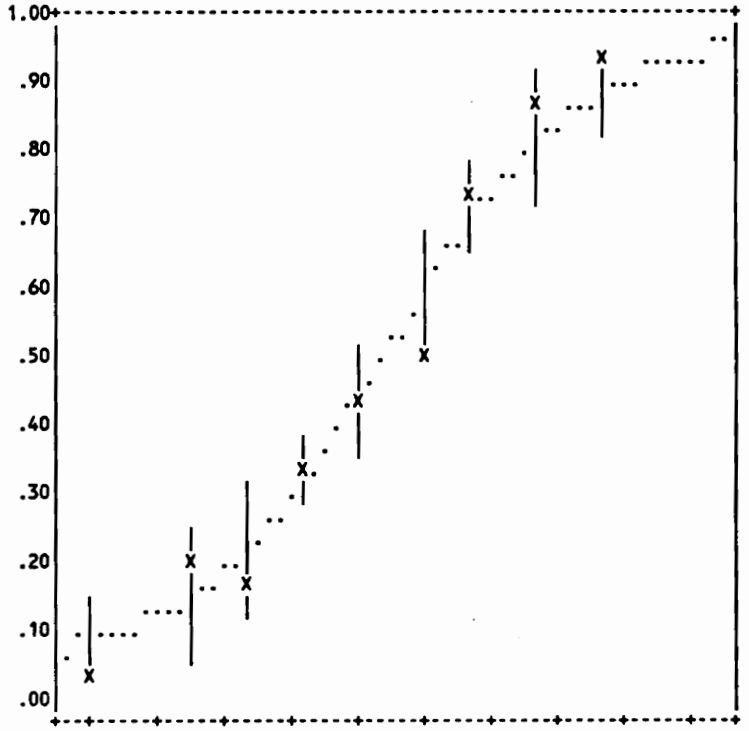
SUBTEST SN
ITEM M128 PROB< .0012



SUBTEST SN
ITEM M140 PROB< .2010

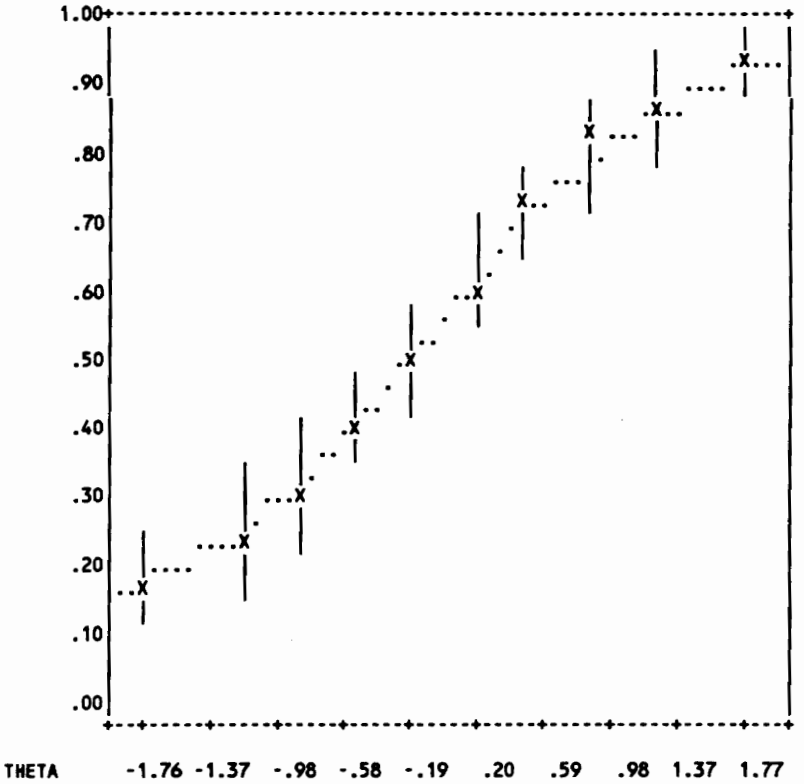


SUBTEST SN
ITEM M145 PROB< .0001

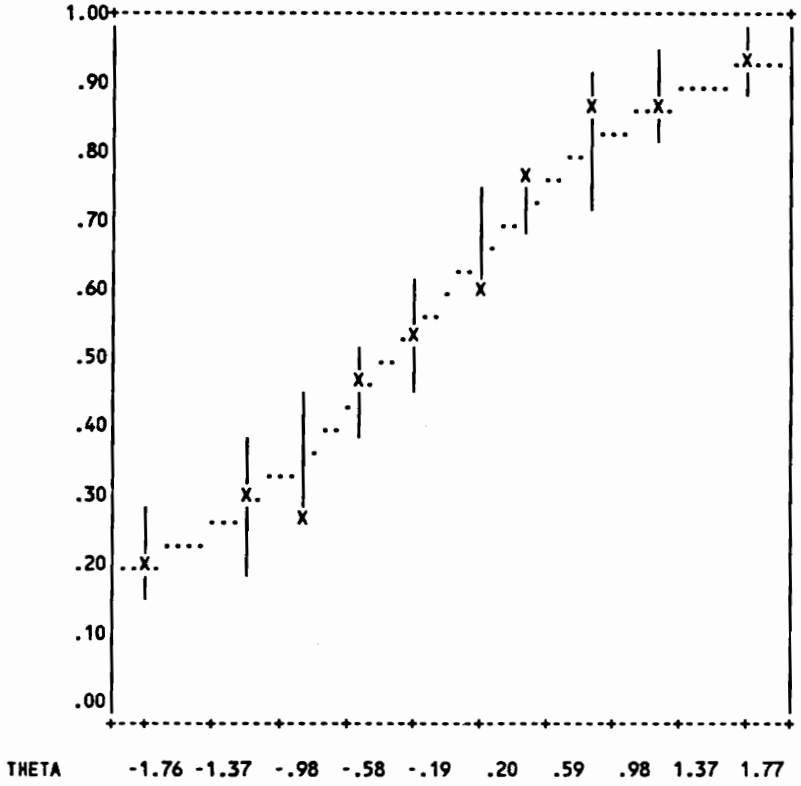


THETA -1.76 -1.37 -.98 -.58 -.19 .20 .59 .98 1.37 1.77

SUBTEST SN
ITEM M149 PROB< .8817

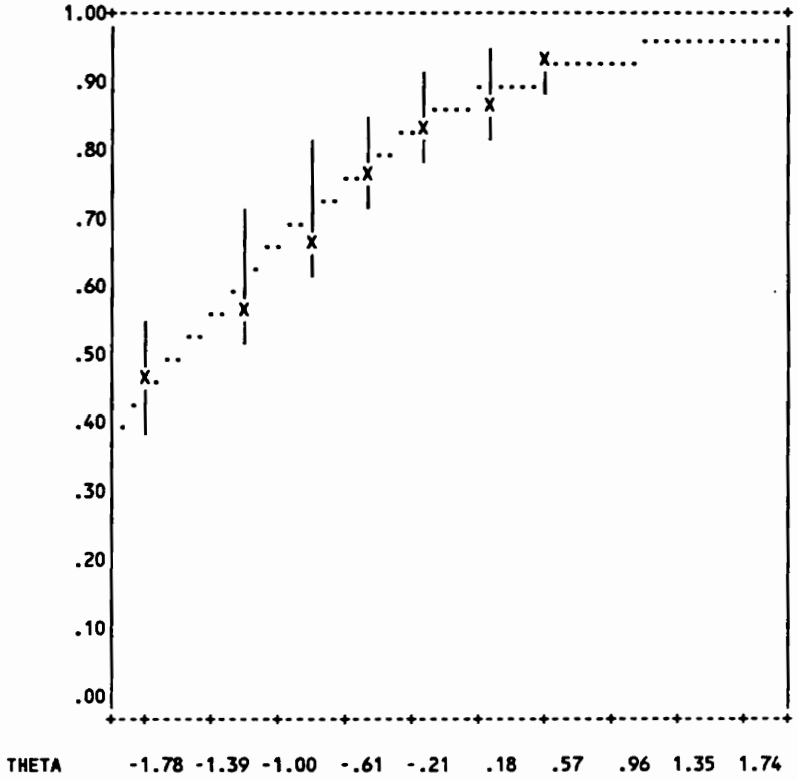


SUBTEST SN
ITEM M165 PROB< .1905

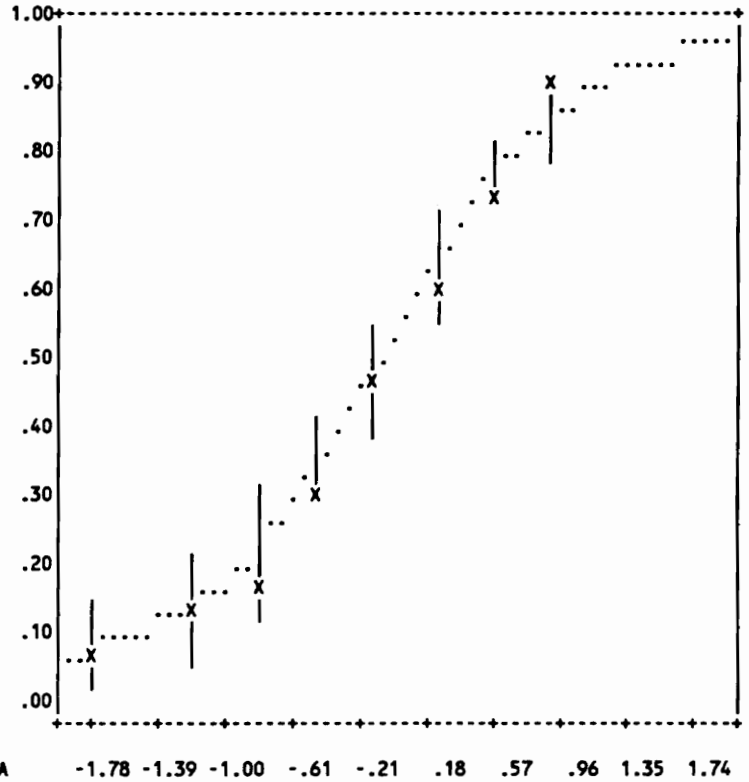


Appendix C

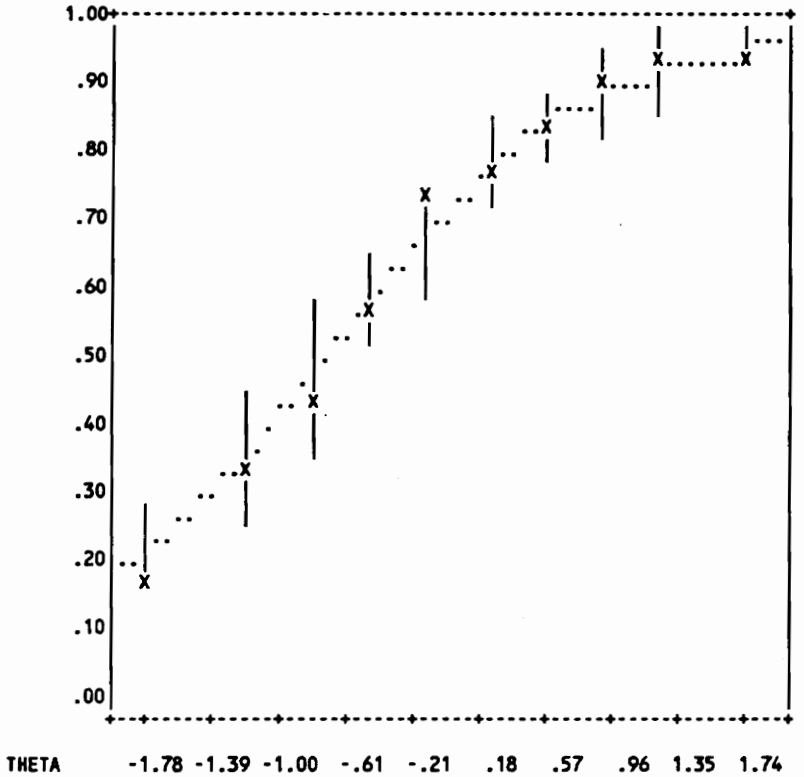
SUBTEST TF
ITEM M4 PROB< .5479



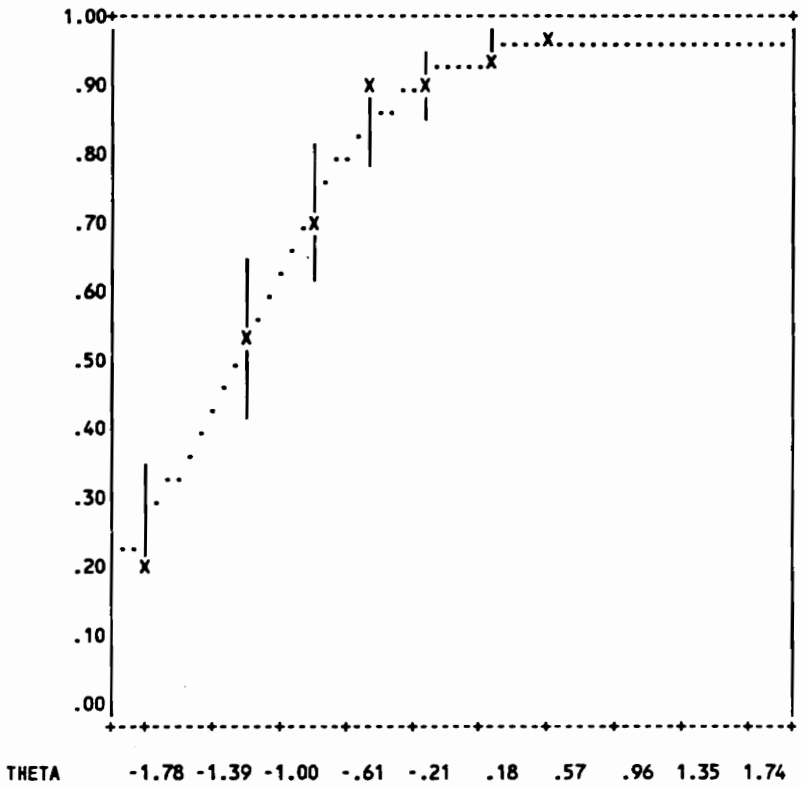
SUBTEST TF
ITEM M26 PROB< .0037



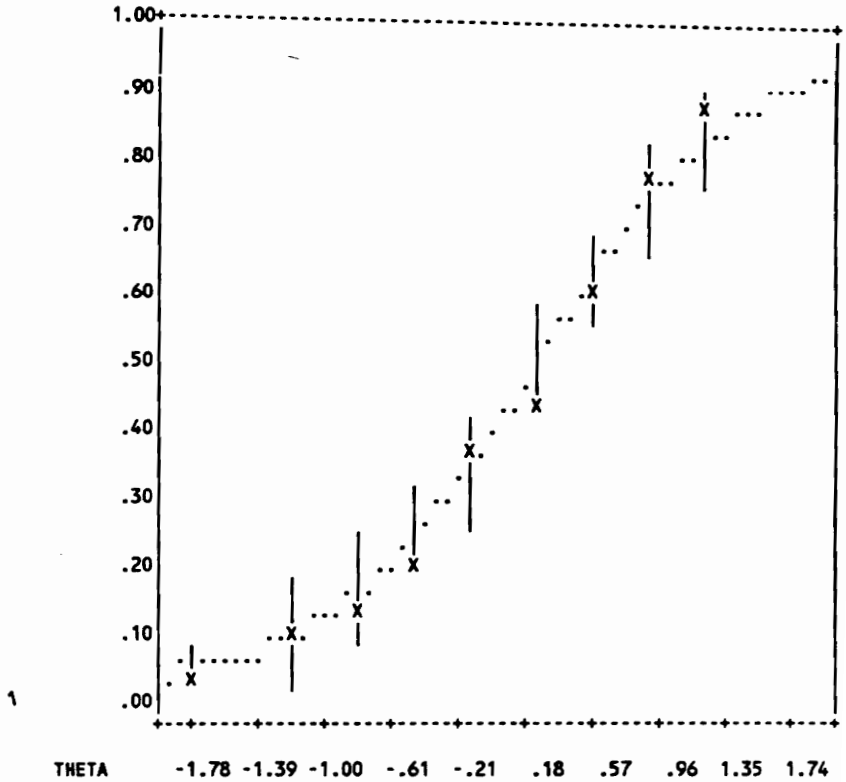
SUBTEST TF
ITEM M29 PROB< .4678



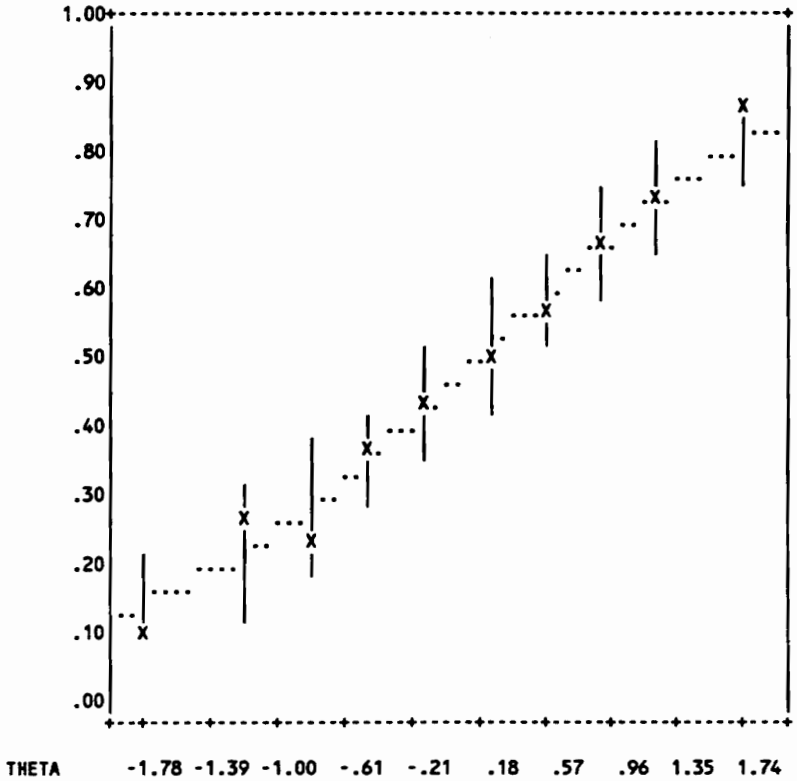
SUBTEST TF
ITEM M72 PROB< .0436



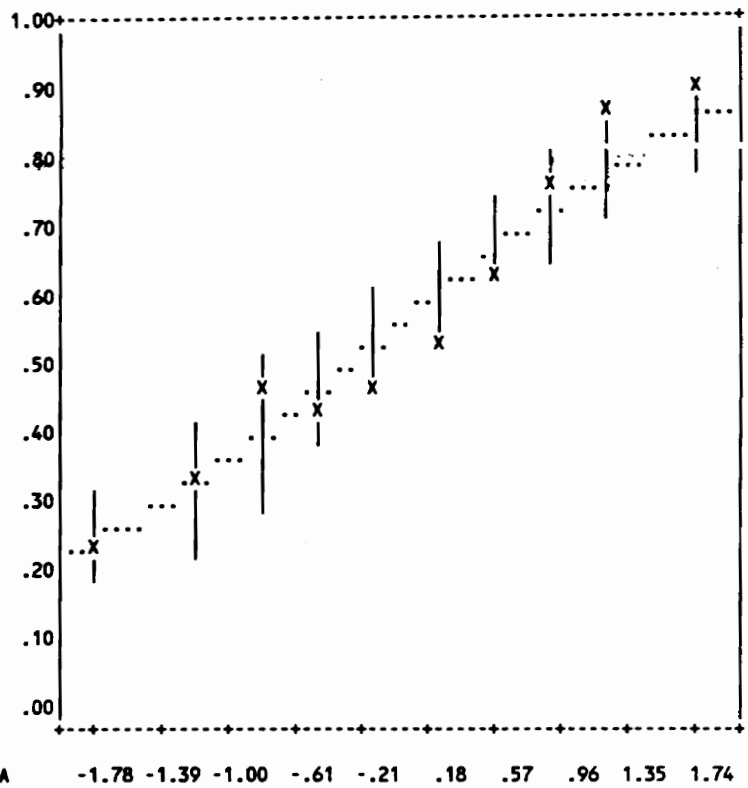
SUBTEST TF
ITEM M79 PROB< .0058



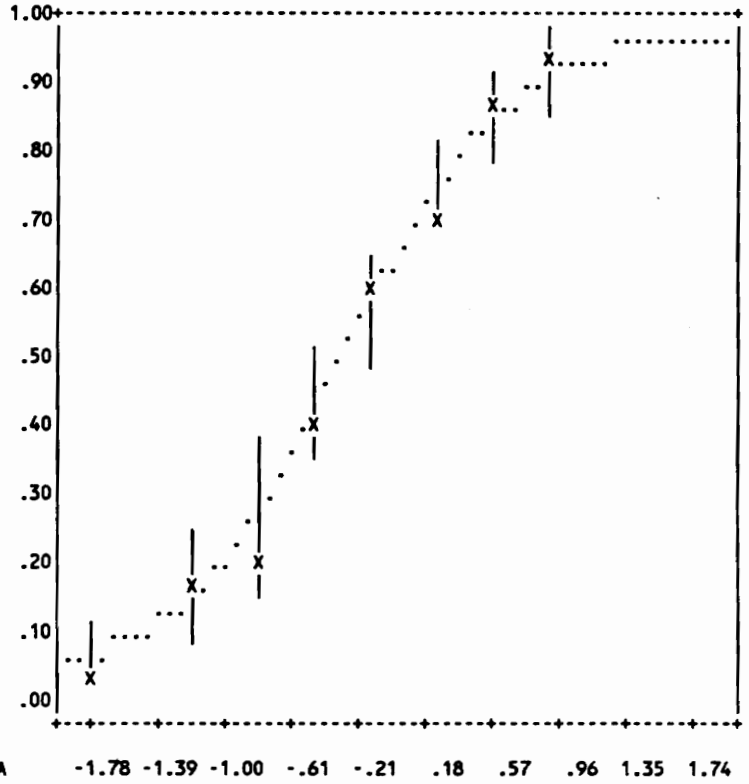
SUBTEST TF
ITEM M81 PROB< .0685



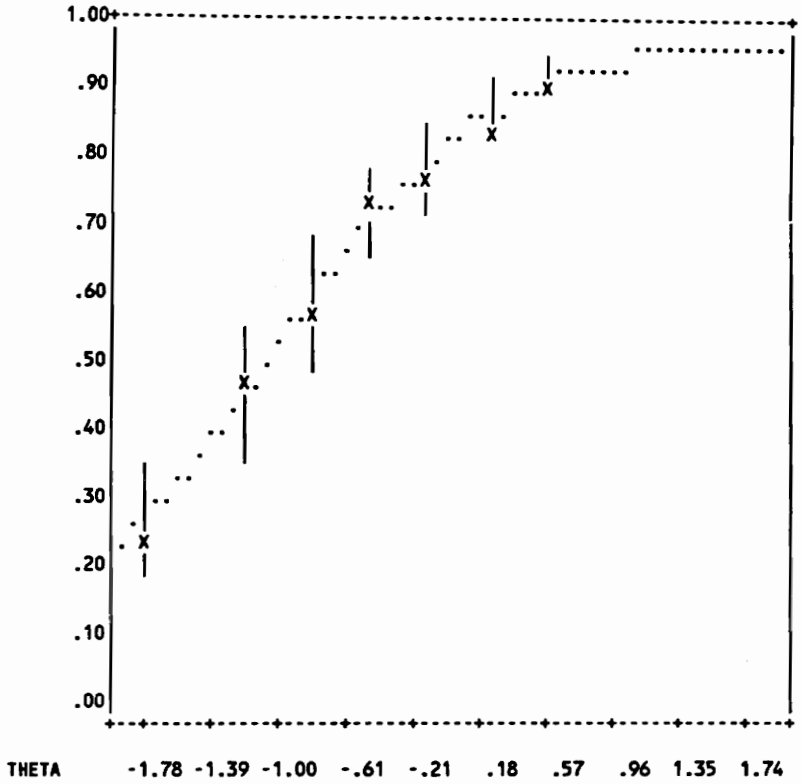
SUBTEST TF
ITEM M84 PROB< .0041



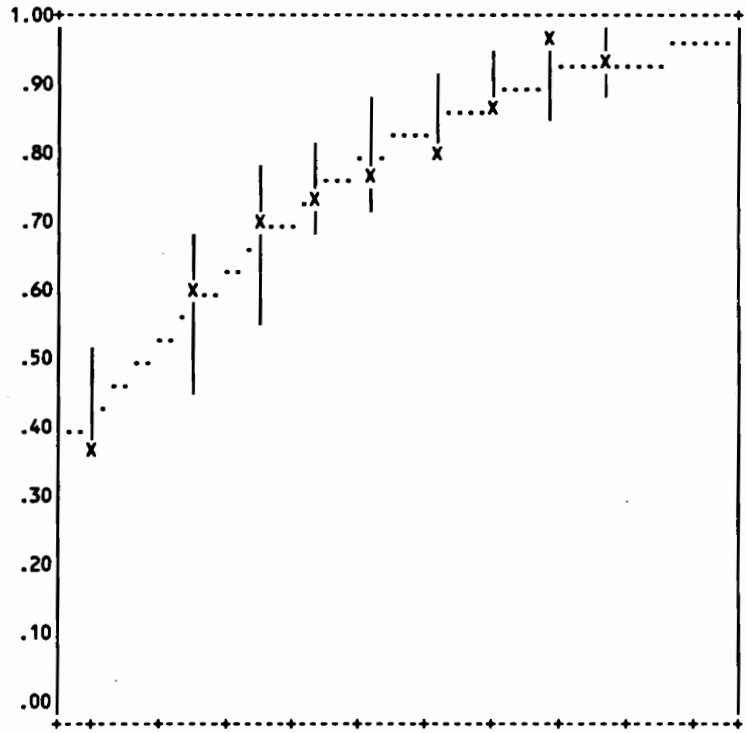
SUBTEST TF
ITEM M86 PROB< .0018



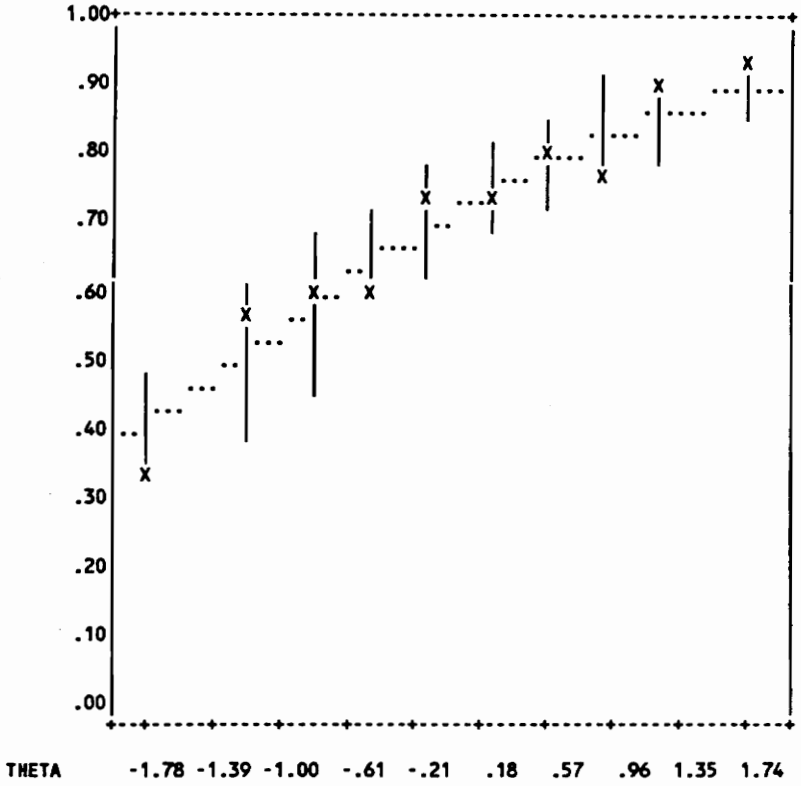
SUBTEST TF
ITEM M89 PROB< .5331



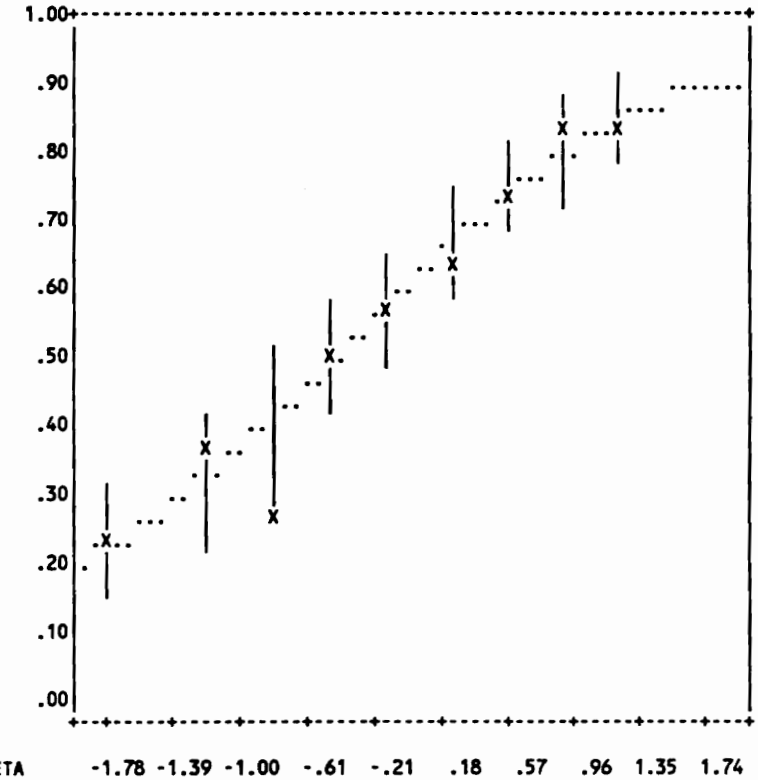
SUBTEST TF
ITEM M91 PROB< .0272



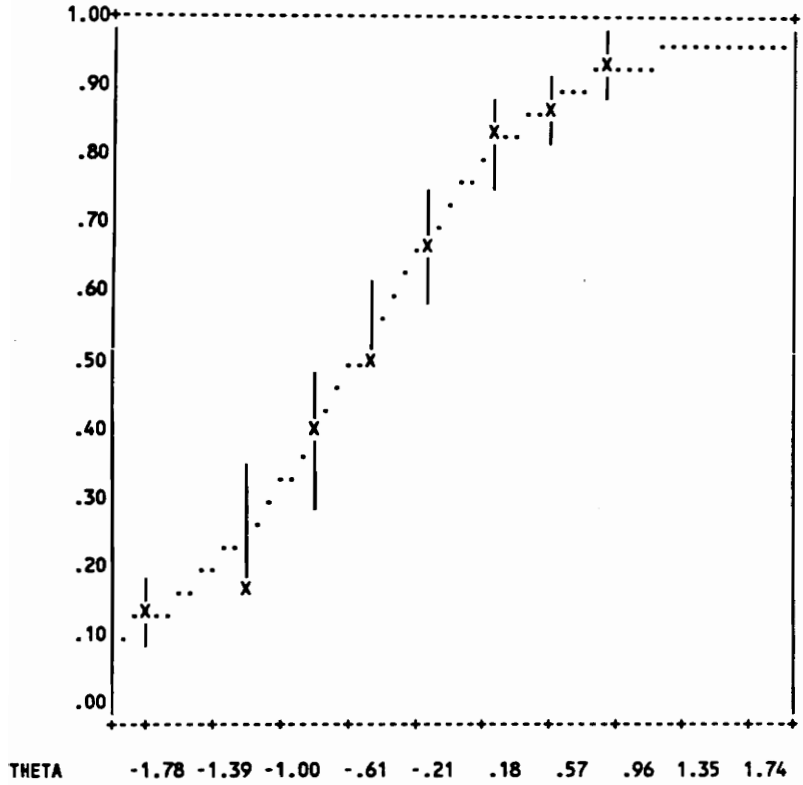
SUBTEST TF
ITEM M93 PROB< .0758



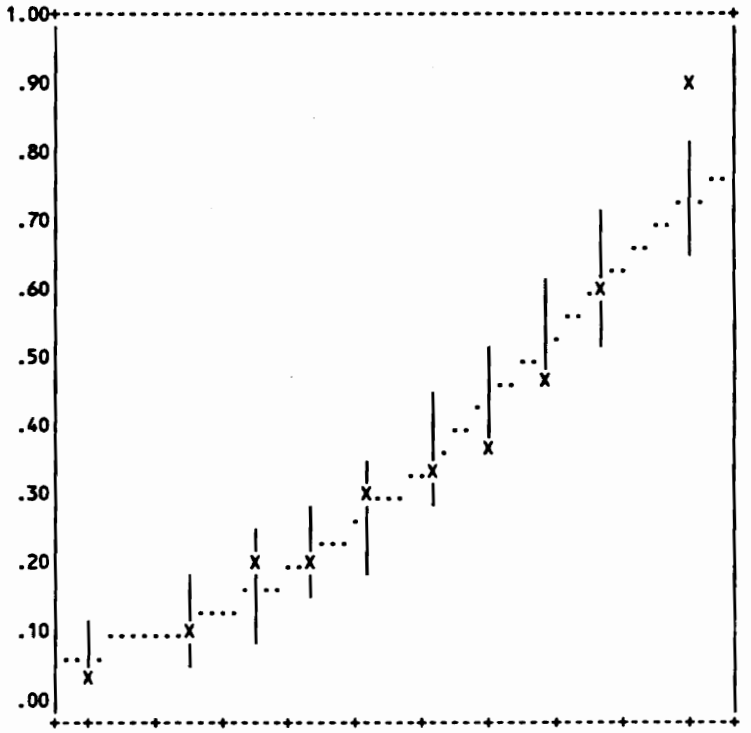
SUBTEST TF
ITEM M100 PROB< .0913



SUBTEST TF
ITEM M103 PROB< .0607

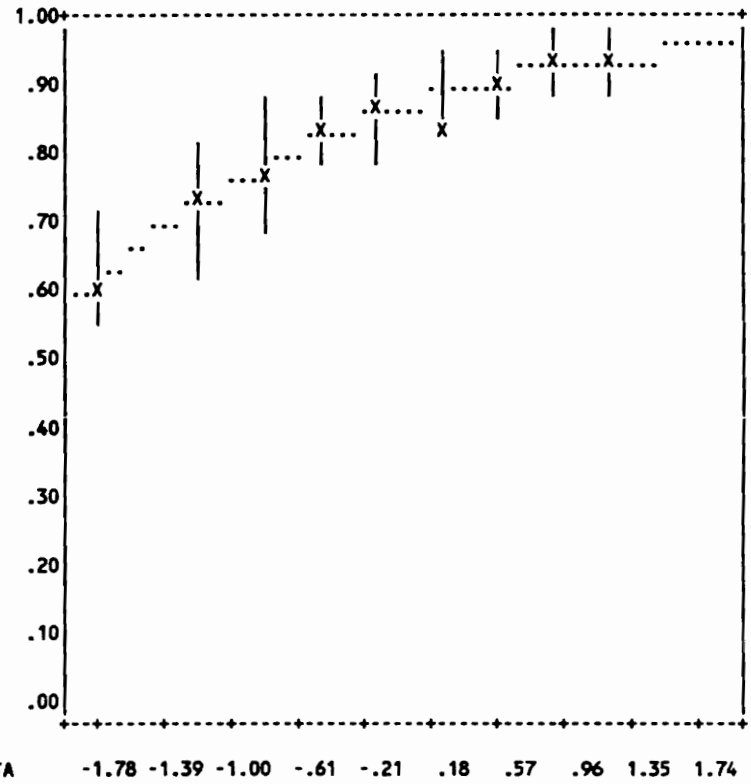


SUBTEST TF
ITEM M105 PROB< .0000

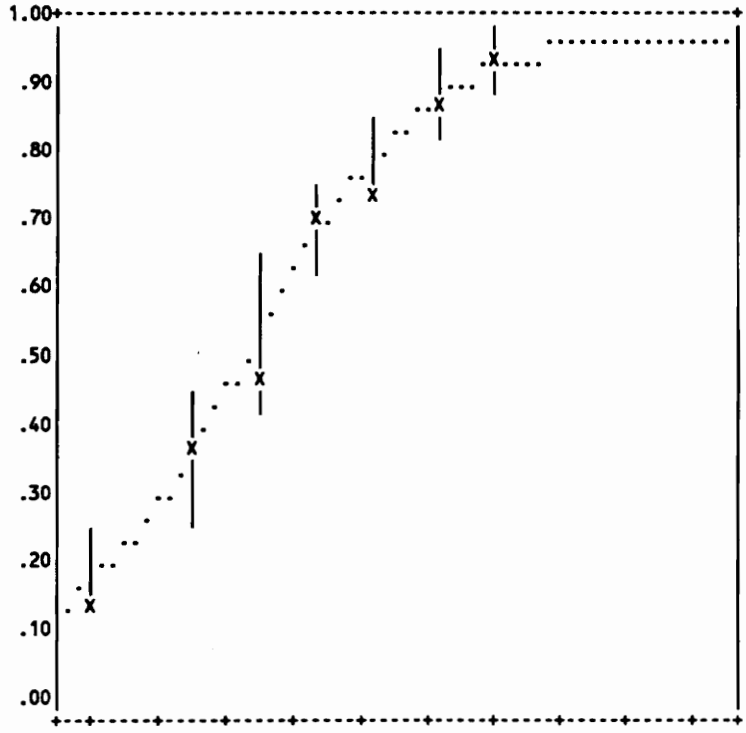


THETA -1.78 -1.39 -1.00 -.61 -.21 .18 .57 .96 1.35 1.74

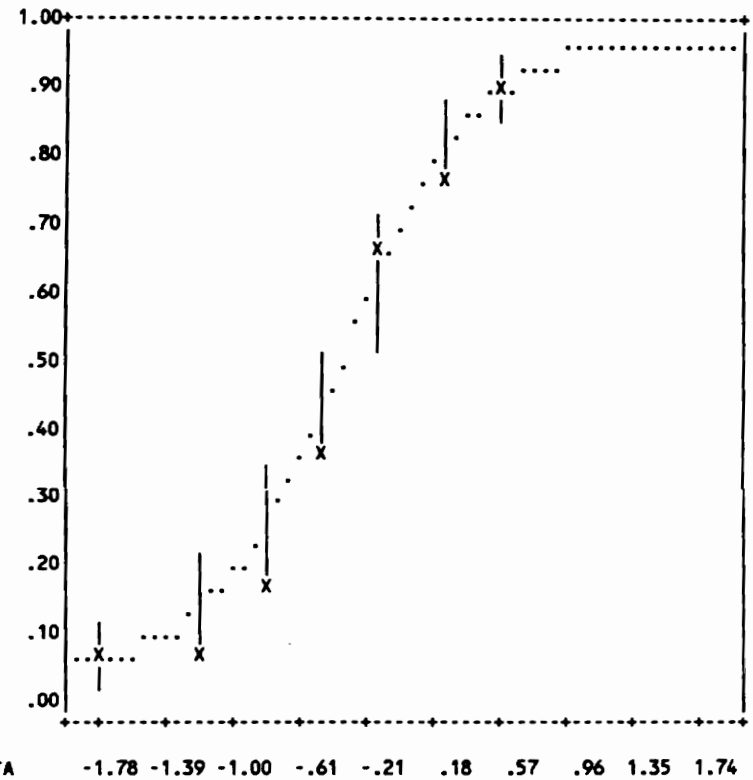
SUBTEST TF
ITEM M108 PROB< .7383



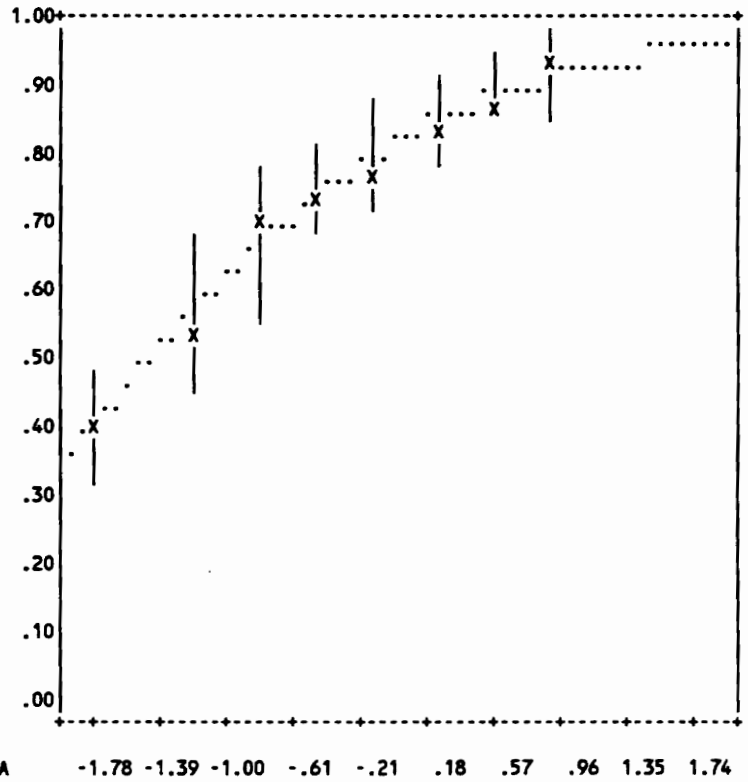
SUBTEST TF
ITEM M111 PROB< .2581



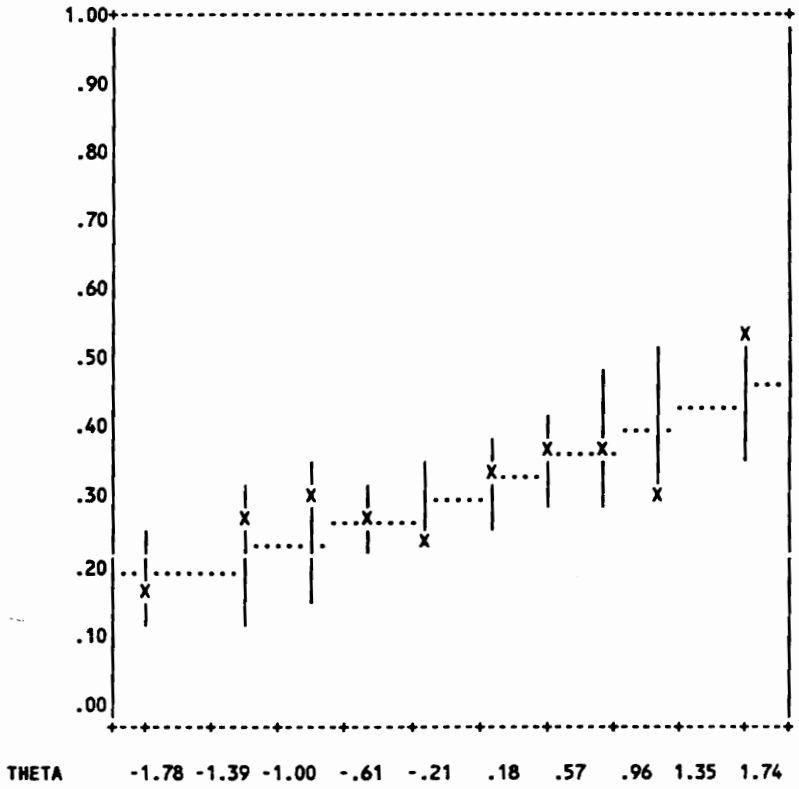
SUBTEST TF
ITEM M114 PROB< .0002



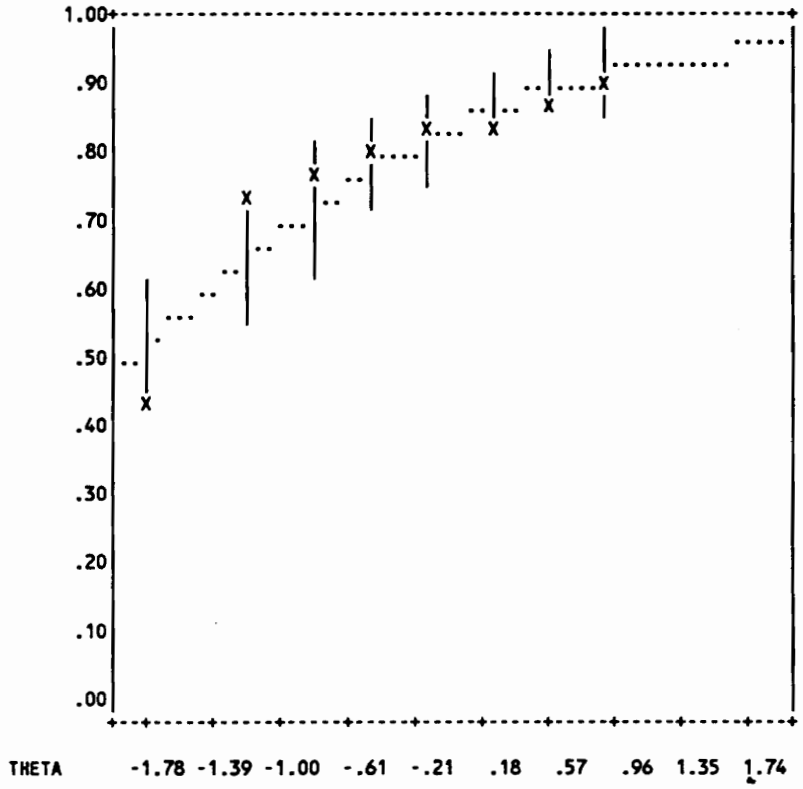
SUBTEST TF
ITEM M120 PROB< .5936



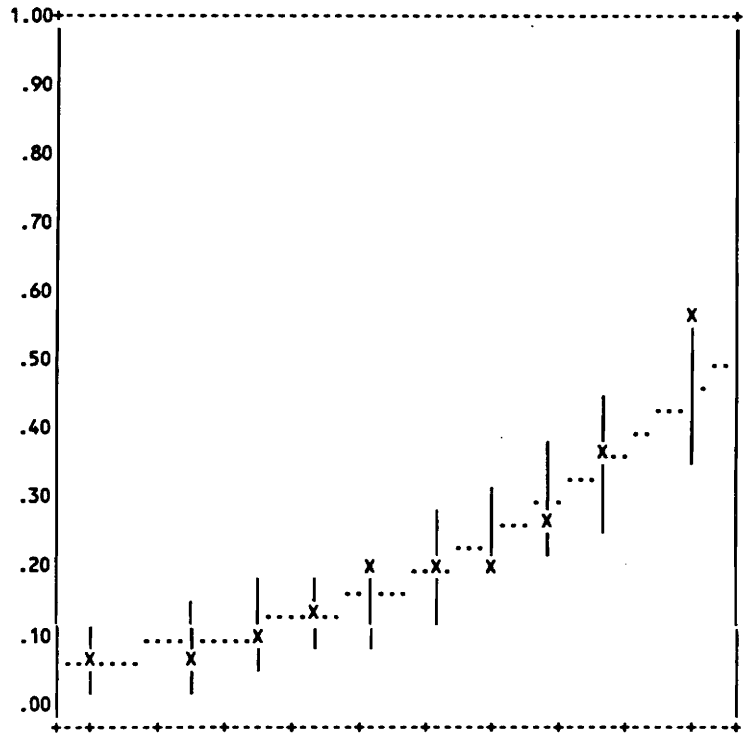
SUBTEST TF
ITEM M122 PROB< .0251



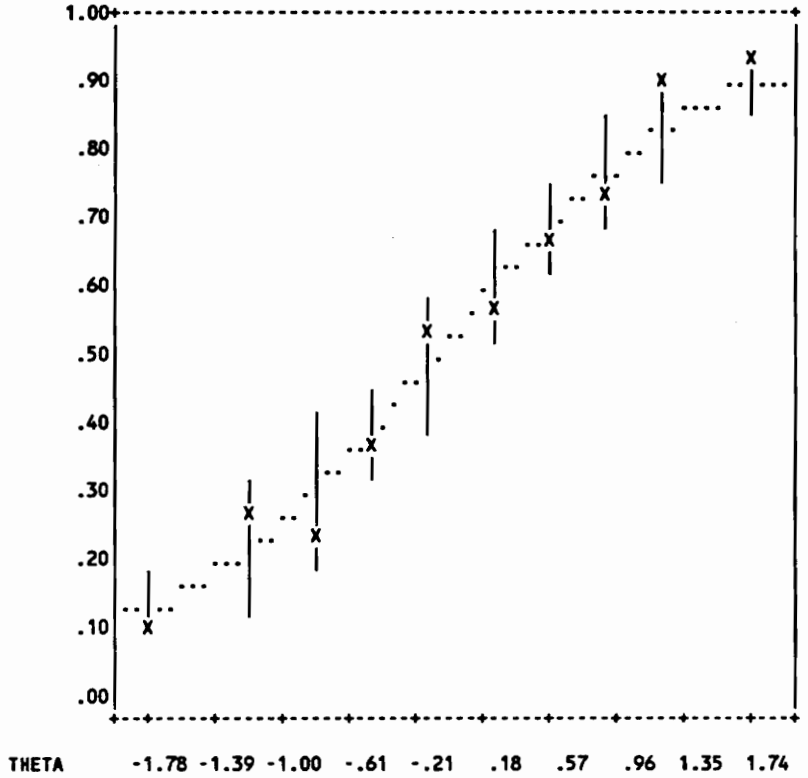
SUBTEST TF
ITEM M133 PROB< .0078



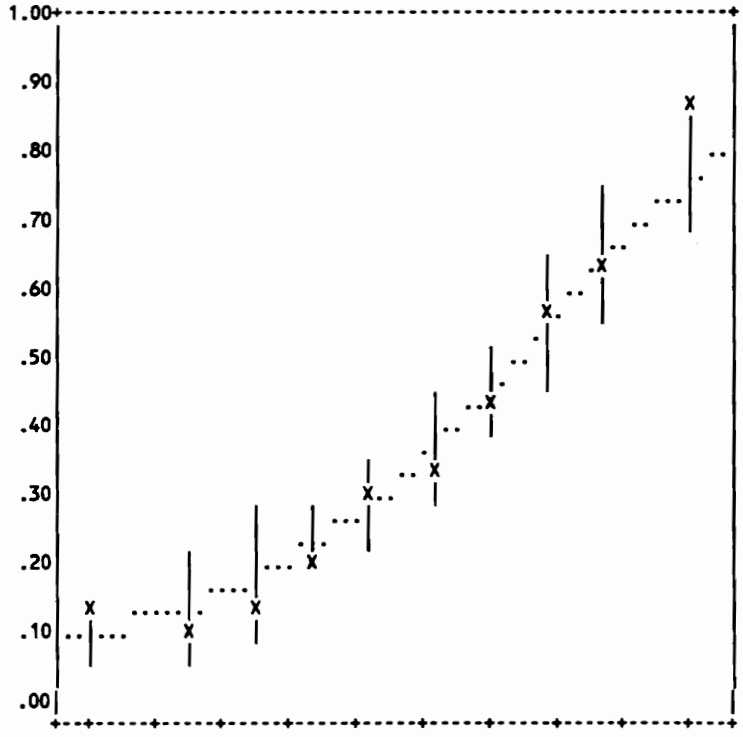
SUBTEST TF
ITEM M147 PROB< .2257



SUBTEST TF
ITEM M154 PROB< .0112

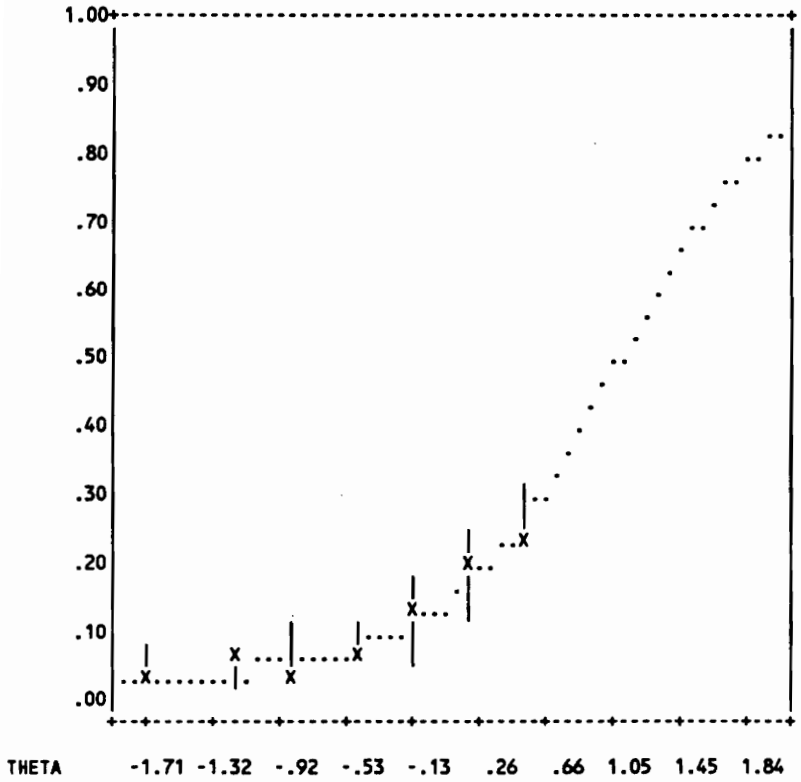


SUBTEST TF
ITEM M158 PROB< .0045

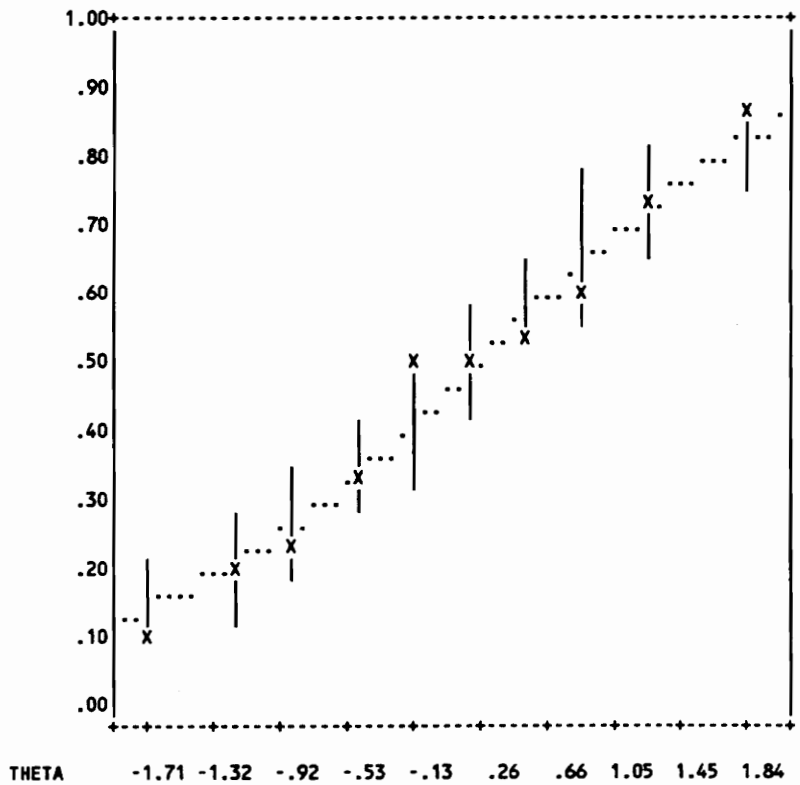


Appendix D

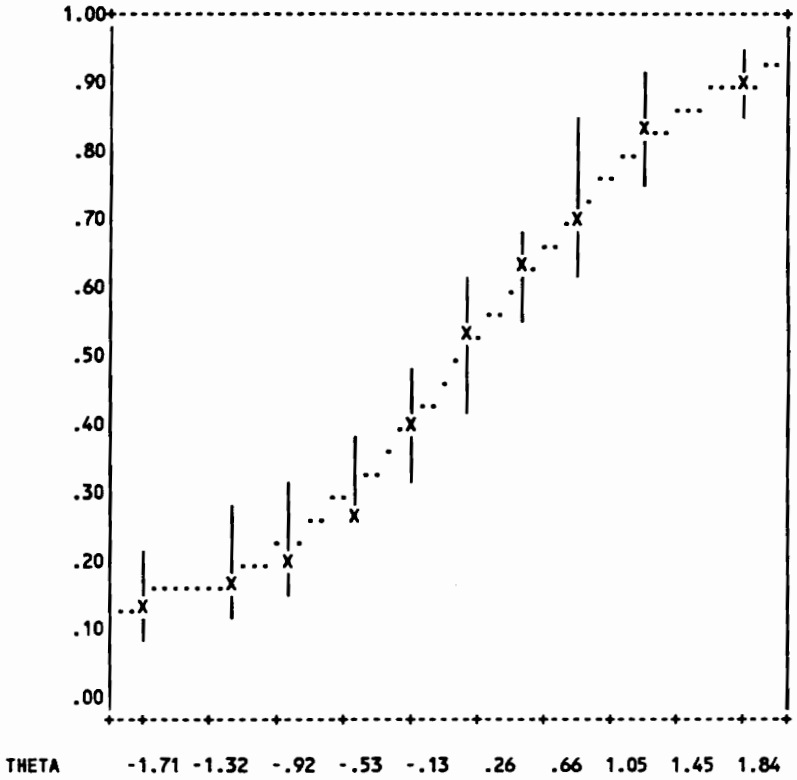
SUBTEST JP
ITEM M1 PROB< .0289



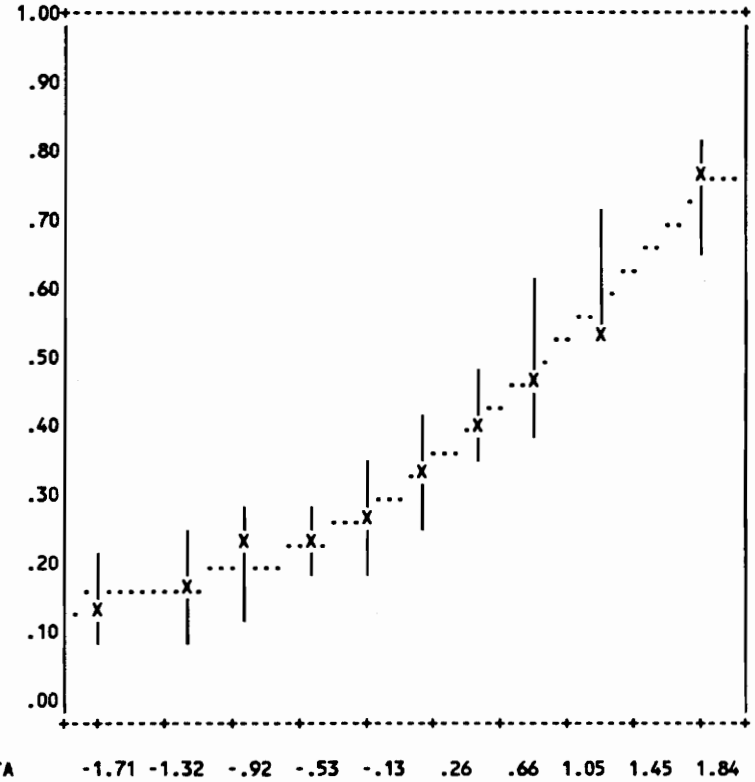
SUBTEST JP
ITEM M9 PROB< .0571



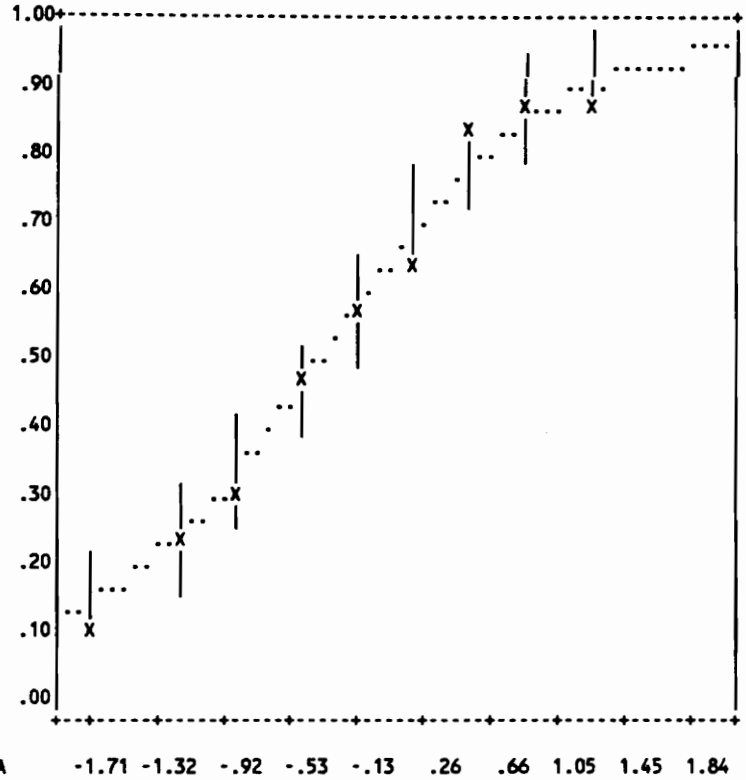
SUBTEST JP
ITEM M13 PROB< .8919



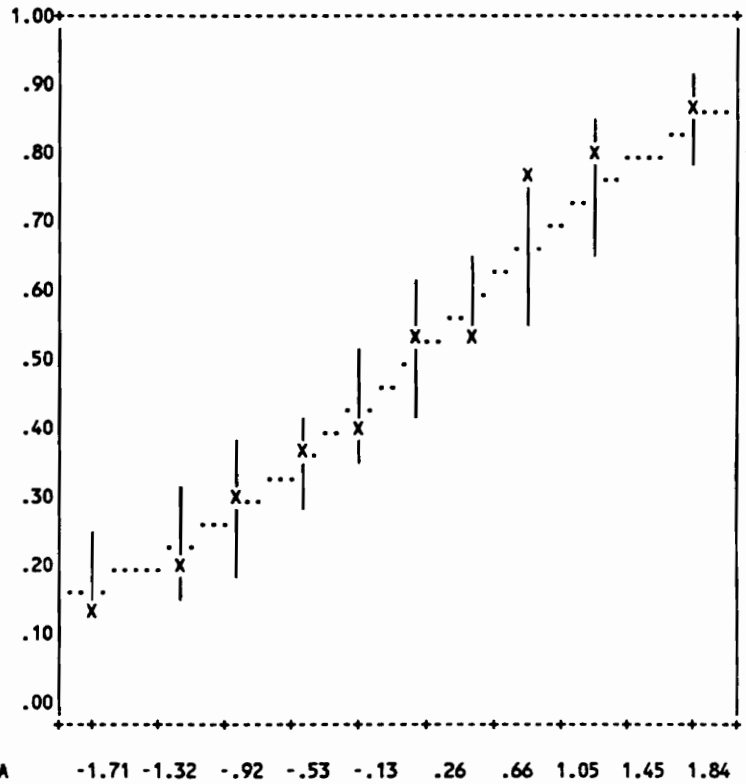
SUBTEST JP
ITEM M20 PROB< .7794



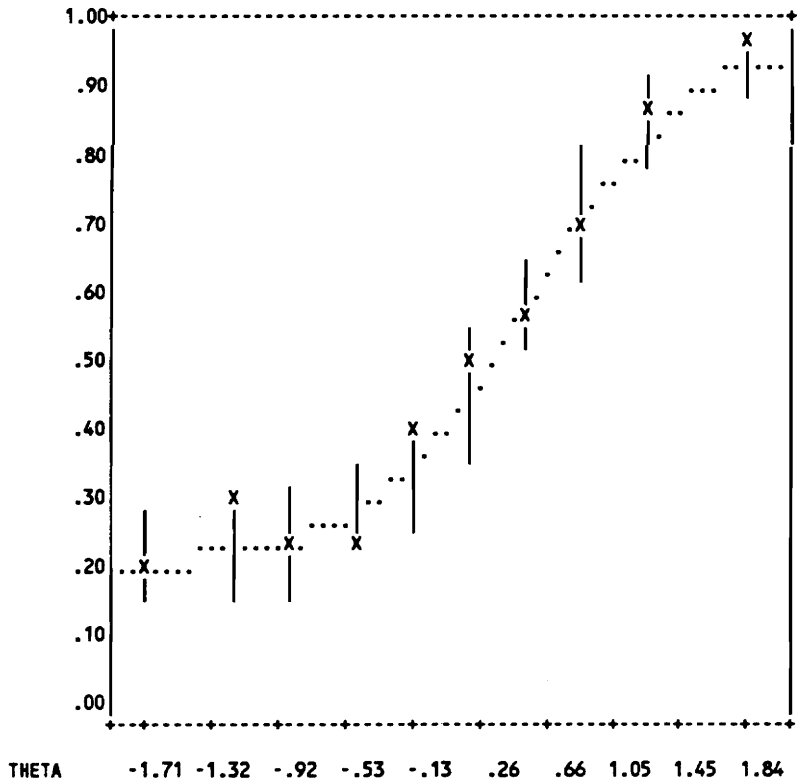
SUBTEST JP
ITEM M27 PROB< .0701



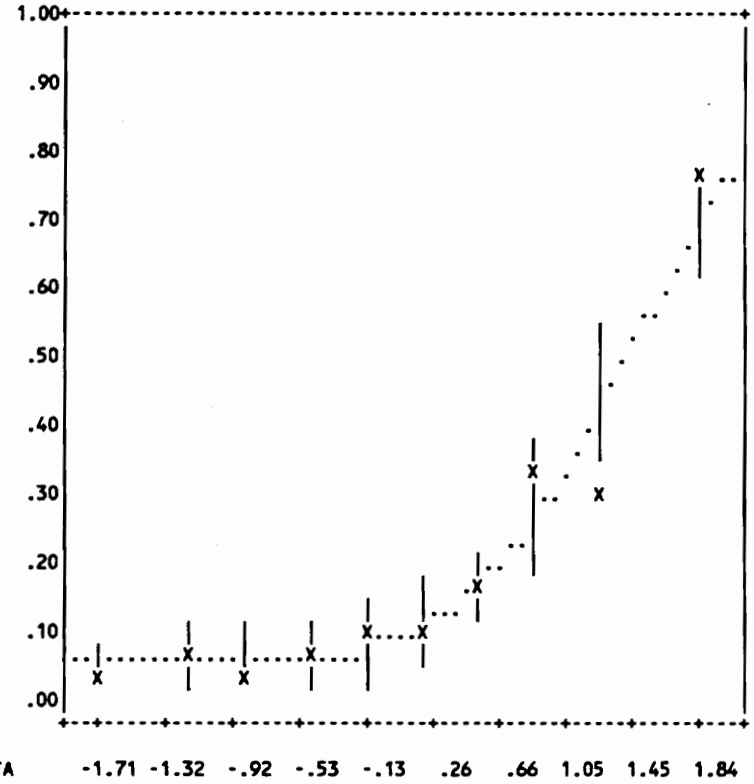
SUBTEST JP
ITEM M35 PROB< .2734



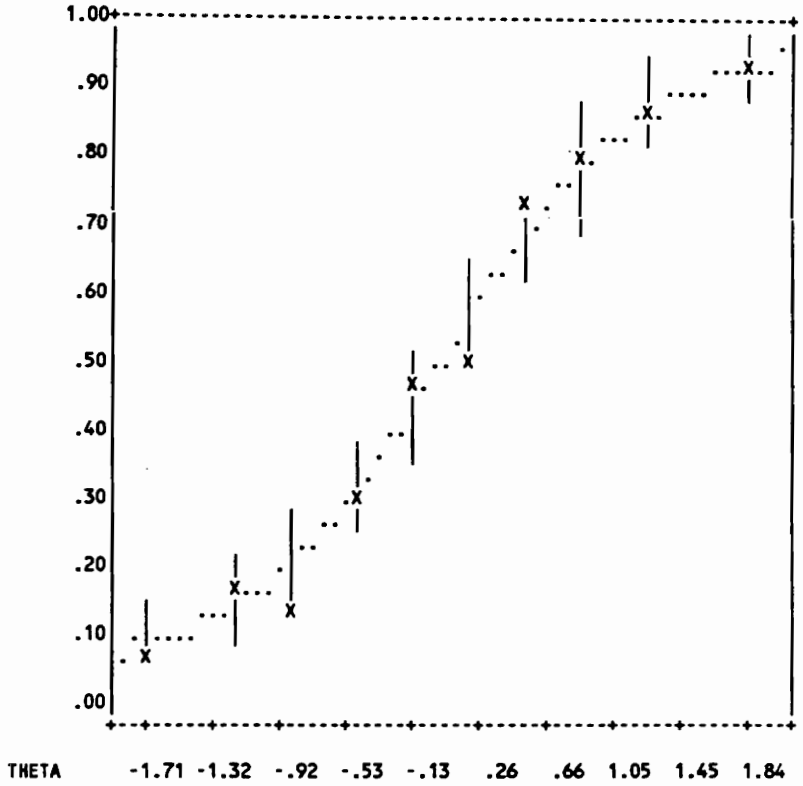
SUBTEST JP
ITEM M42 PROB< .0468



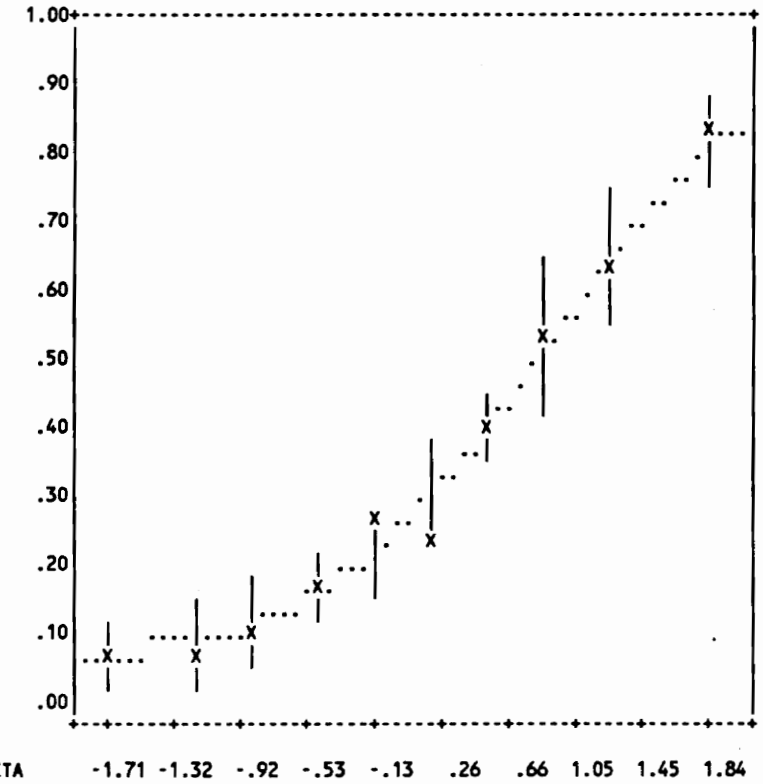
SUBTEST JP
ITEM M49 PROB< .0074



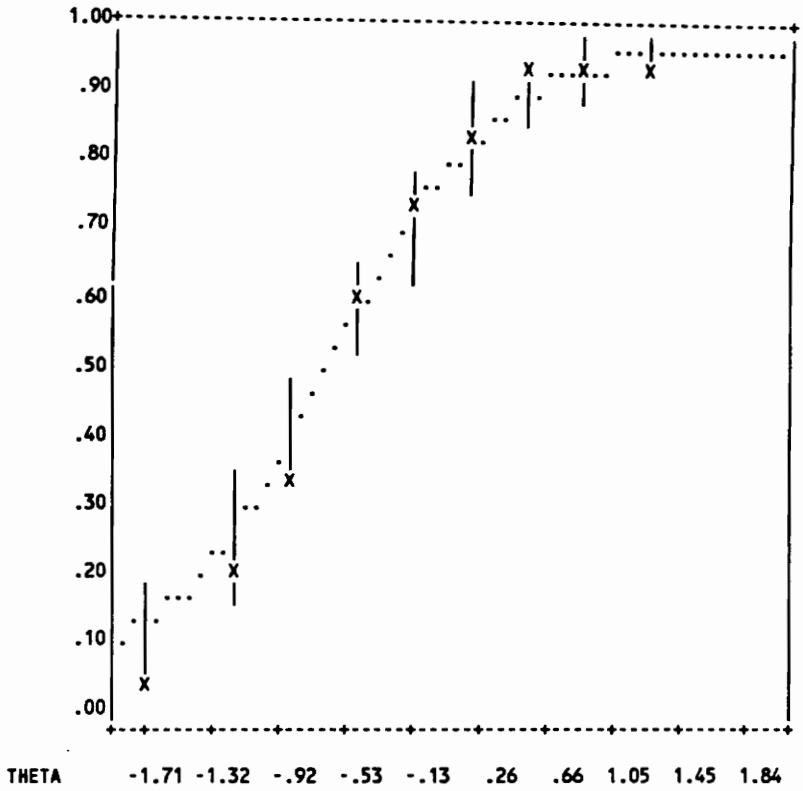
SUBTEST JP
ITEM M55 PROB< .0198



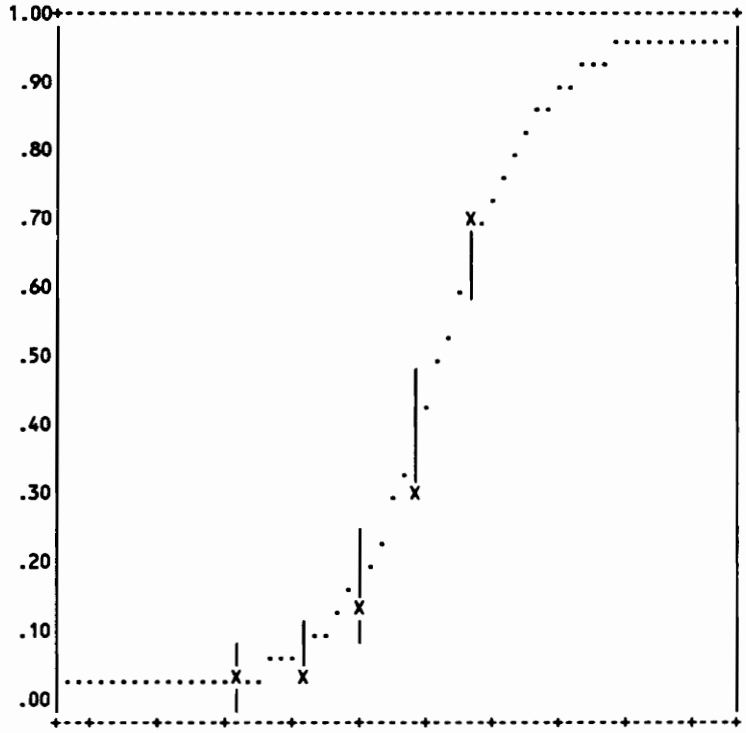
SUBTEST JP
ITEM M60 PROB< .3186



SUBTEST JP
ITEM M74 PROB< .0000

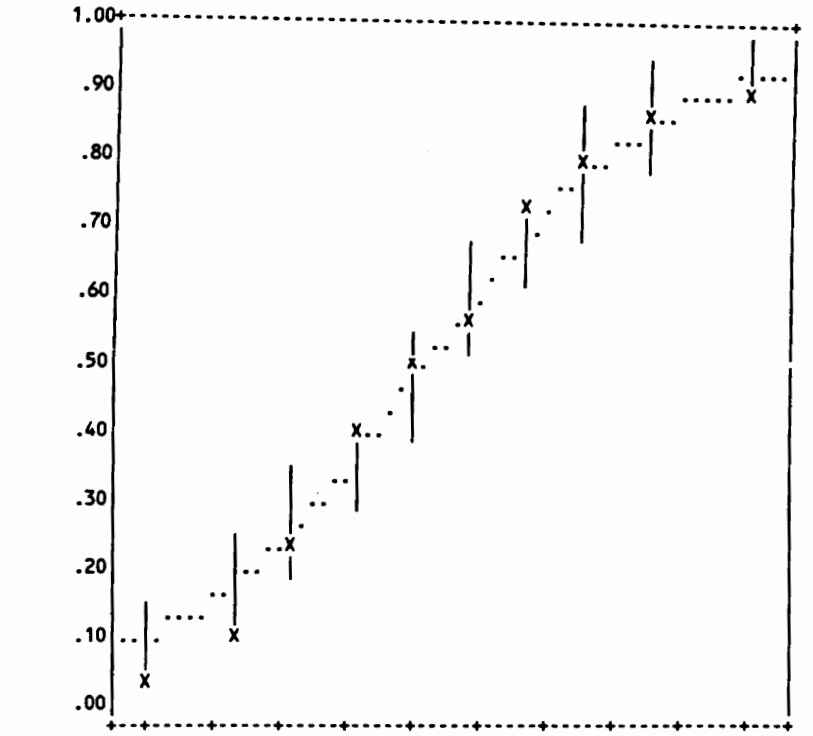


SUBTEST JP
ITEM M85 PROB< .0000

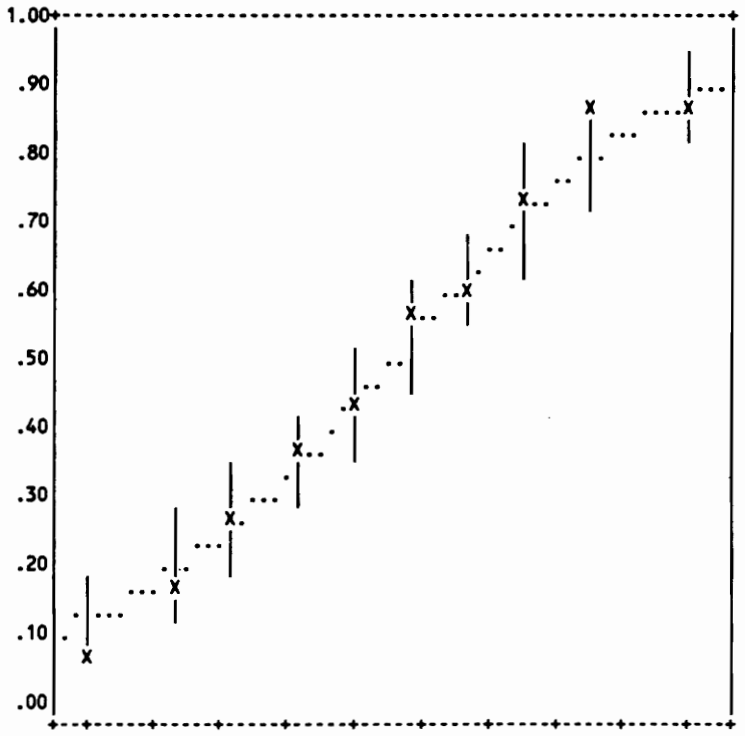


THETA -1.71 -1.32 -.92 -.53 -.13 .26 .66 1.05 1.45 1.84

SUBTEST JP
ITEM M94 PROB< .0014

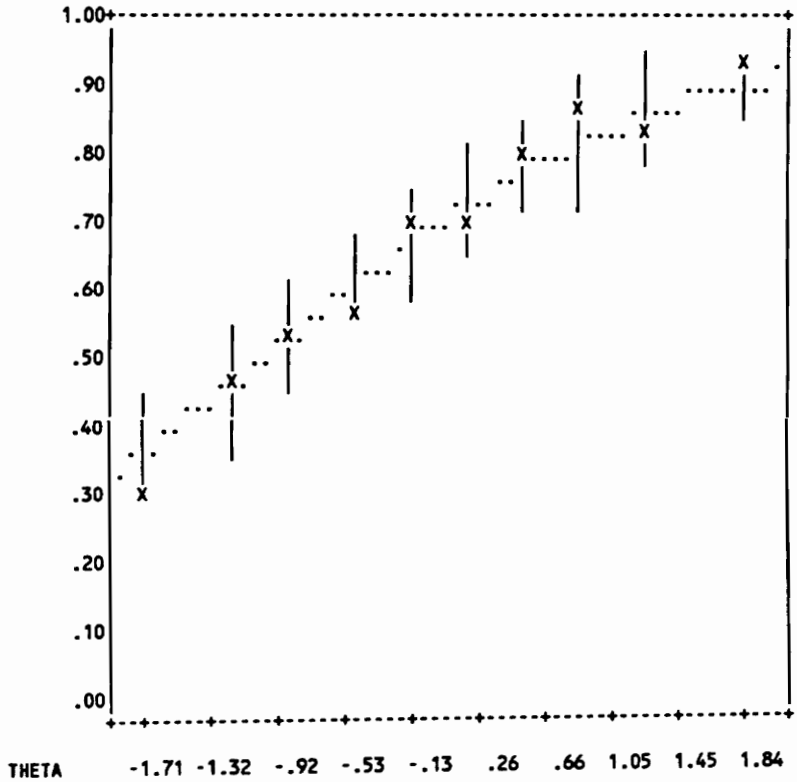


SUBTEST JP
ITEM M97 PROB< .2061

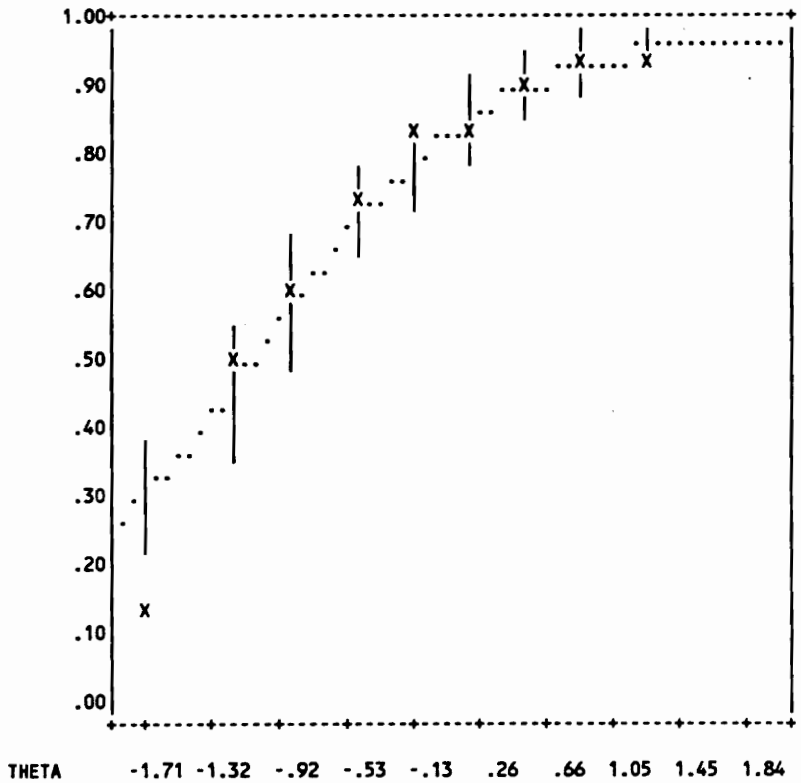


THETA -1.71 -1.32 -.92 -.53 -.13 .26 .66 1.05 1.45 1.84

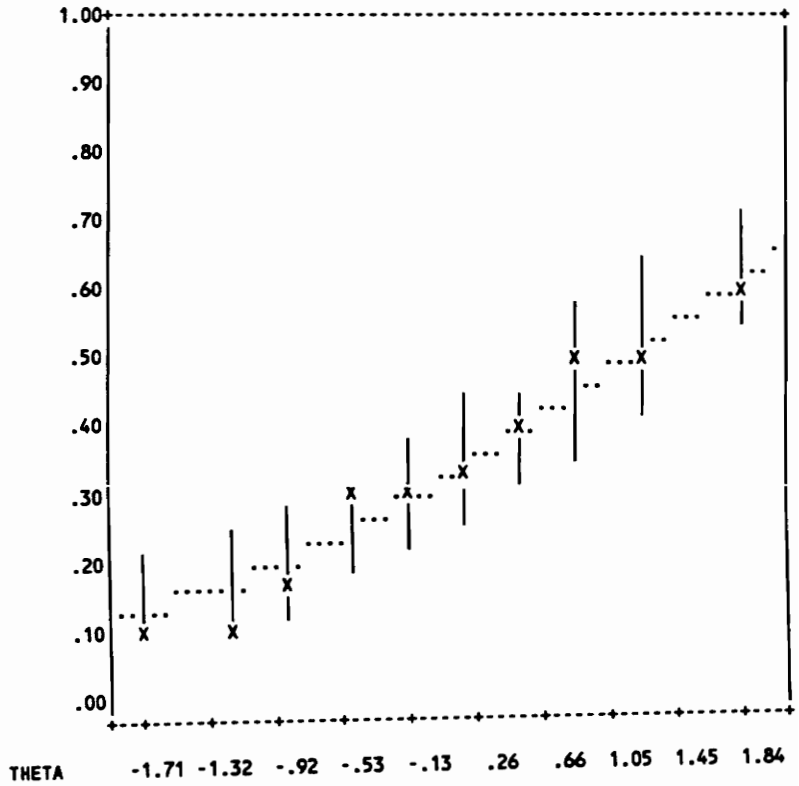
SUBTEST JP
ITEM M99 PROB< .2052



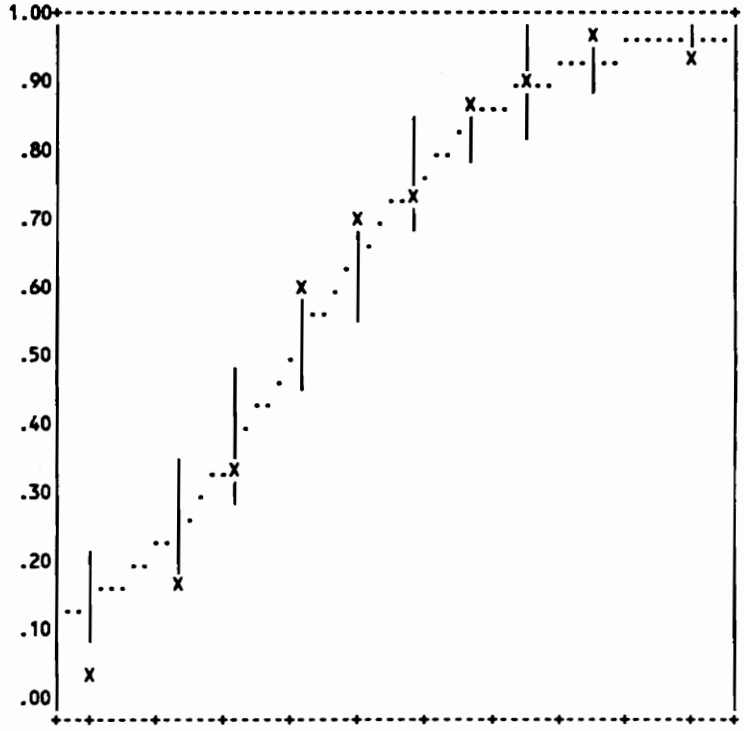
SUBTEST JP
ITEM M109 PROB< .0000



SUBTEST JP
ITEM M113 PROB< .1843

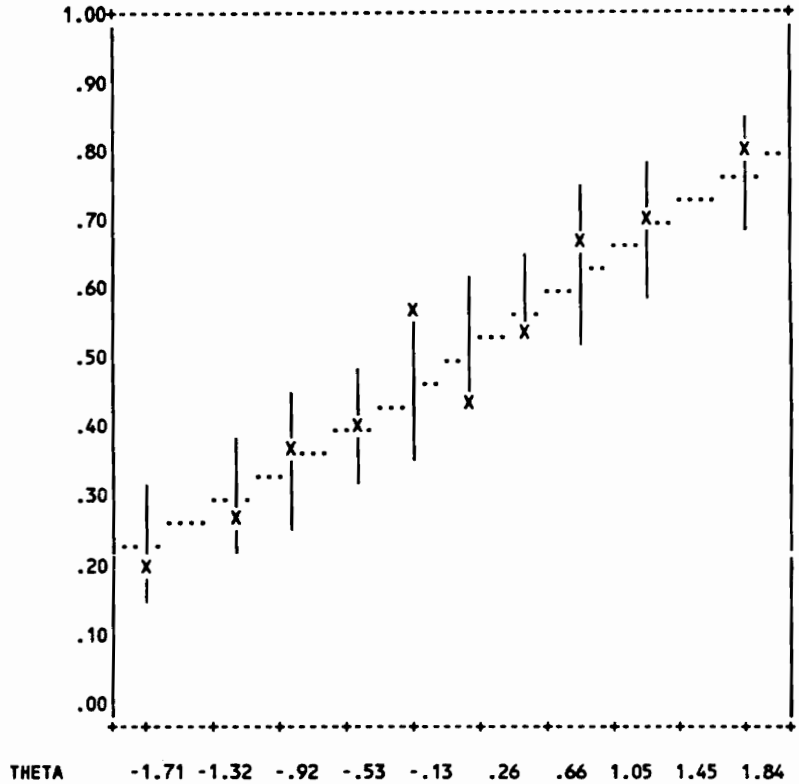


SUBTEST JP
ITEM M118 PROB< .0000

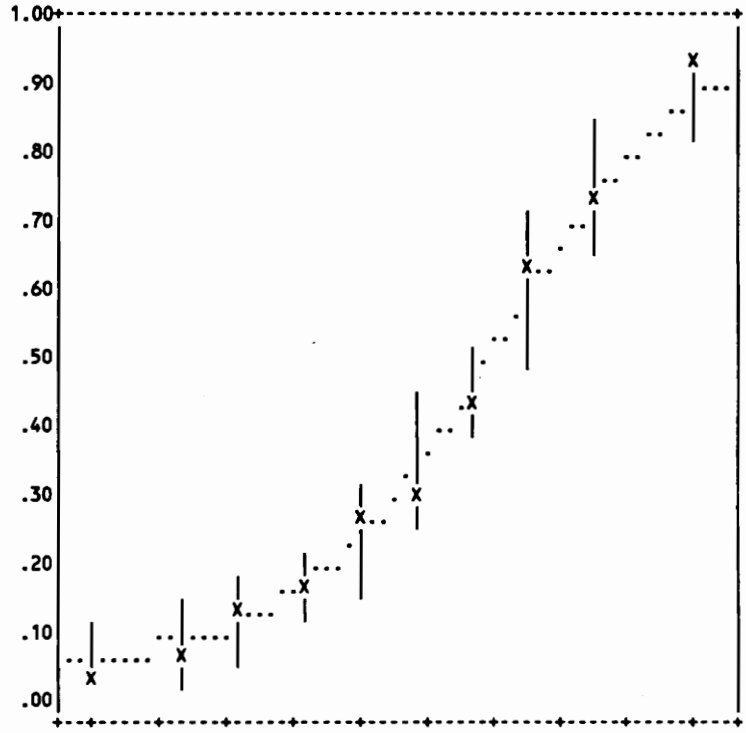


THETA -1.71 -1.32 -.92 -.53 -.13 .26 .66 1.05 1.45 1.84

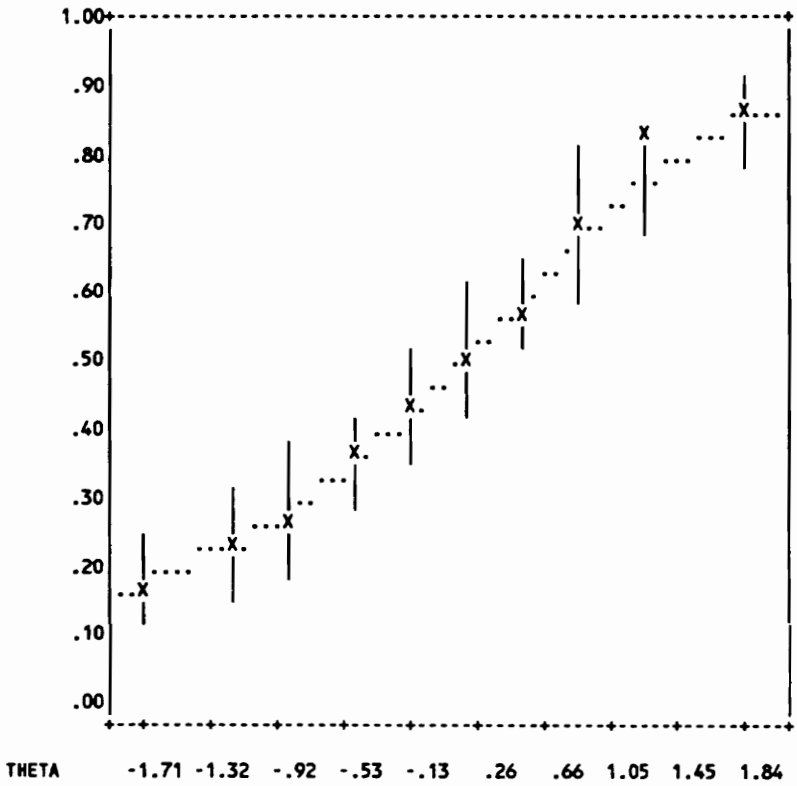
SUBTEST JP
ITEM M124 PROB< .0367



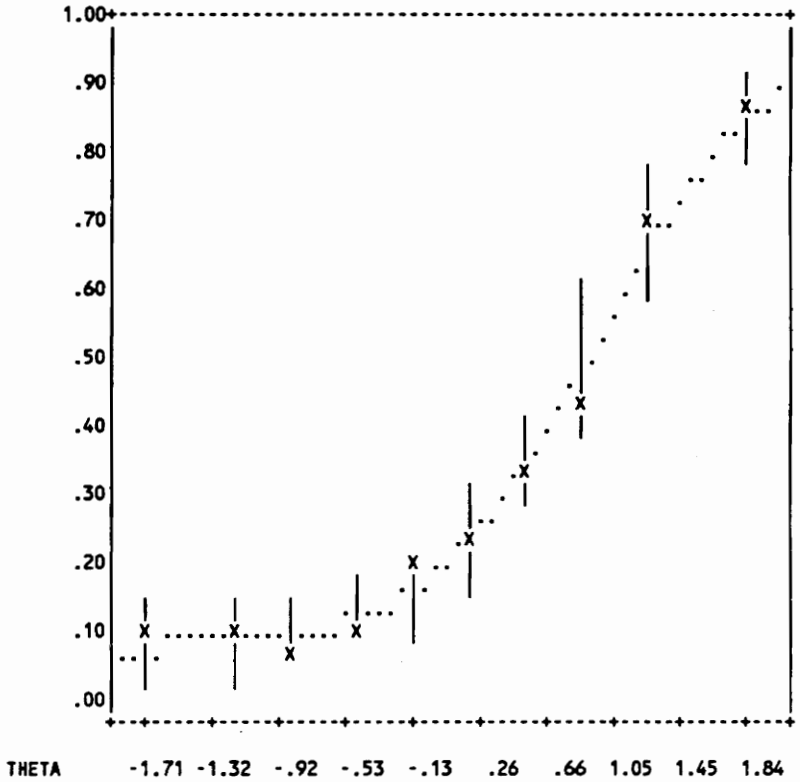
SUBTEST JP
ITEM M132 PROB< .1495



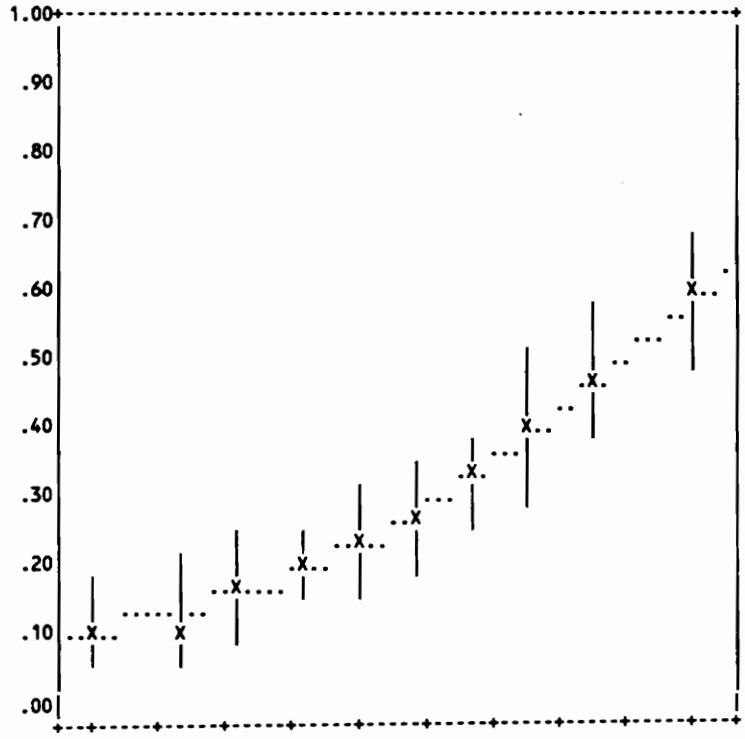
SUBTEST JP
ITEM M142 PROB< .7293



SUBTEST JP
ITEM M151 PROB< .4574



SUBTEST JP
ITEM M153 PROB< .9149



THETA -1.71 -1.32 -.92 -.53 -.13 .26 .66 1.05 1.45 1.84

Stuart Elliott Greenberg

7513-A Northcrest Drive
Indianapolis, IN 46256

EDUCATION

Ph.D. 1992	Industrial/Organizational Psychology Virginia Tech Blacksburg, Va
M.S. 1990	Industrial/Organizational Psychology Virginia Tech Blacksburg, Va
B.A. 1987	Psychology Lafayette College Easton, Pa

PROFESSIONAL EXPERIENCE

Boehringer Mannheim Start 11/92
Indianapolis, IN
Market Research Analyst

-assess perceptions of consumers in multiple markets

United Airlines 6/90-12/90
Chicago, IL
Human Resource Planning Intern

Revised Flight Attendant selection system
- -conducted criterion oriented test validation
- -revised group exercise
- -researched and developed rating dimensions
- -developed structured interview protocol

Completed designated projects
- -conducted criterion oriented test validation for Special Service Representatives
- -supervised pilot testing of new selection instruments
- -analyzed data with the SAS system
- -consulted internally for statistical analysis
- -administered job analysis focus groups
- -conducted training sessions
- -administered tests

Bell Atlantic 12/91-1/92
Arlington, Va
Contracted as Test Administrator

-Administered computerized tests on-site for validation of a selection battery

Management Consulting Team 1/91-5/91
Blacksburg, Va 5/89-1/90
Consulting Team Member

-conducted on-site job analysis interviews
-wrote job descriptions
-researched literature on selection tests and performance appraisal instruments concerned with customer service
-administered tests
-scored, compiled and analyzed test results

Virginia Tech 9/87-5/92
Department of Psychology
Blacksburg, Va
Graduate Teaching Assistant

-assisted in grading and evaluation of students
-SAS computer consultant for Graduate statistics class

-lecture assistant for: Introductory Psychology class
Sensation and Perception class
-taught: Social Psychology lecture class
Introductory Psychology labs
Advanced Social Psychology labs
Psychometric Theory labs

PRESENTATIONS

Greenberg, S. E. & Smither, J. W. (1988). *Union commitment of teachers: A replication and validation of the Friedman and Harvey Scale*. Paper presented at the annual meeting of the Eastern Psychological Association, Buffalo, N.Y.

Greenberg, S. E. (1992). *A sequence of SAS programs used to teach psychometric theory*. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

RESEARCH INTERESTS

Thesis: Measuring Absence Cultures: An Examination of Absence Perceptions of Males and Females

Dissertation: Differential Item Functioning on the Myers-Briggs Type Indicator

PROFESSIONAL MEMBERSHIPS

American Psychological Association

Society for Industrial and Organizational Psychology

References

Dr. Neil Hauenstein
Department of Psychology
Virginia Tech
Blacksburg, Va 24061-0436
(703) 231-5716

Dr. R. J. Harvey
Department of Psychology
Virginia Tech
Blacksburg, Va 24061-0436
(703) 231-7030

Dr. Jeff Klawsky
United Airlines
EXOHR
P.O. Box 66100-0100
Chicago, Il 60666-0100
(708) 952-7539

Stuart E. Greenberg