EXAMINEE CONTROL OF ITEM ORDER EFFECTS ON LATENT TRAIT MODEL AND CLASSICAL MODEL TEST STATISTICS

by

MICHAEL J. SCALES

B.Ed., University of Victoria, 1978

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES
(Department of Educational Psychology and Special Education)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
July 1990

5

Department of Educational Psychology

The University of British Columbia
Vancouver, Canada

Date July 26, 1990

ABSTRACT

The purpose of this study was to determine what effect changes in the item order had on classical and on latent trait test statistics. As well, comparisons were made between students who were allowed to answer the questions in any order, and students who were required to answer the questions in the order presented in the test booklet. The results were then analyzed using the student's ability level as an additional independent factor.

Four different formats of a forty item mathematics test were used with 590 students in grade eight. Half of the booklets had the items sequenced from easiest to hardest. The other booklets were sequenced from hardest to easiest. In addition, half of the tests of each sequence had special directions which prevented students from altering the given item difficulty sequence. The classroom teachers provided a rating of each student's ability in mathematics.

The order of the items was found to have a significant effect. Tests which were sequenced from hard to easy had a lower mean score. Although students with test booklets with restrictive directions had lower scores on average, it was not a statistically significant difference. There were no significant interactions found. Classical and latent trait item difficulty statistics showed a high degree of correlation.

It was concluded that under certain circumstances, the order of the items could effect both classical and latent trait statistics. It was also recommended that care should be taken when assumptions are made about parallel forms or local independence.

TABLE OF CONTENTS

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank the members of the thesis committee for their assistance. Dr. Robert Conry, Dr. Donald Allison, and Dr. David Bateson provided me with much needed advice and support under what I am sure were conditions of very short notice. I would also like to thank the many teachers and students who provided me with their greatly appreciated cooperation. Finally, I would like to bestow my deepest gratitude to my wife, Margot Lakoduk, my son, Riordan, and my daughter, Alysha. Their sacrifices on my behalf have been sincerely appreciated.

Chapter I

Introduction

Context of the Research Problem

Ever since multiple choice tests were considered the "new-type examinations" (Ruch, 1929) to the present descriptions of computerized marking systems (Hopkins & Antes, 1985), advice has been forthcoming from many textbook authors that multiple choice tests should be arranged with the easiest questions at the beginning to the hardest questions at the end. This advice has had a great deal of intuitive appeal and assumed certainty. For example one author states:

> The level of difficulty of objective test items is used as a basis for arranging these items in a test by placing the easy ones first, the more difficult ones later, and the most difficult ones last. Such an arrangement has advantages for the average and below average pupil. With this kind of test he uses the testing time allowed more efficiently, and his morale is improved. If the difficult test items appear first, many pupils of average or low achievement will waste a great deal of time trying to answer them. They may fail to answer easier test items later in the test

> because so much time was spent on the first ones. Moreover, they may quickly become discouraged or even hostile. On the other hand, if the easier test items are listed first, these same pupils will at first make smooth progress in the test, and consequently feel encouraged. When they later encounter the more difficult test items, they no doubt will have time to attack them. Even if they fail to answer some of them correctly, as will very likely happen, the resulting disappointment will be moderated by the knowledge that they already have responded to some items in a manner that is probably correct. (Ahmann & Glock, 1963, p.115)

However, despite such conviction of what examinees will no doubt do and feel, empirical research does not support this same lack of doubt. Research over the years has been inconclusive. It is not a certainty that the arrangement of test items will make a difference to the score of the examinee.

In fact, the issue of item order effects has been an area of research for nearly forty years. As Leary and Dorans (1985) pointed out, the research has reflected the interests and statistical abilities of the times. So, while the accuracy of tests have improved, the need for more precise statistics has also increased. As a result, item order continues to be a concern due to conflicting research results; one researcher will conclude that item order has no significant effect (Allison, 1984) whereas another researcher will conclude that the effect is significant (Hambleton & Traub, 1974). In fact, Lane, Bull, Kundert, and Newman (1987) reported finding significant order effects in their first study and non-significant effects in their second study.

The first research on item order began in the early 1950s and examined the simple main effect of item order on classical test statistics. Researchers wanted to test the axiom that tests should be constructed with the easiest questions first. A variety of arrangements were tried such as easy to hard, hard to easy, random, and spiralling. Some initial studies reported a significant effect (Mollenkopf, 1950; MacNicol, 1956; Sax & Carr, 1960; Sax & Cromack, 1966; Flaugher, Melton, & Meyers, 1968; Sirotnik & Wellington, 1974; Hambleton & Traub, 1974; Kleinke, 1980; Hodson, 1984). However, some later research reported that item order did not make a significant difference (Brenner, 1964; Huck & Bowers, 1972; Monk & Stallings, 1970; Klosner & Gellman, 1973; Kestenbaum & Weiner, 1970; Allison, 1984).

In the late 1960s there was a concern about the emotional state of the exam takers and their level of anxiety, so the emphasis of the research shifted to examine these internal states. In addition, statistical techniques using factor analysis were in more common usage, and researchers could examine the interaction effect of reported anxiety level and test results. As in previous research, a variety of item arrangements were used. The results of this research were also mixed with some studies reporting significant effects (Munz & Smouse, 1968; Smouse & Munz, 1969; Towle & Merrill, 1975; Plake, Ansorge, Parker, & Lowry 1982) whereas other research found the main effects and the interaction effects to be non-significant (French & Greer, 1964; Smouse & Munz 1968; Berger, Munz, Smouse, & Angelino, 1969; Marso, 1970; Munz & Jacobs, 1971; Plake, 1980; Plake, Thompson, & Lowry, 1980; Plake, Melican, Carter, Shaughnessy, 1983; Plake, Ansorge, 1984; Klimko, 1984).

Recent research has returned to a concern about simple item order, but the researchers have begun to use a more modern computerized analysis involving latent trait models of test statistics rather than classical test statistics. The results of the limited number of studies to date have reported significant effects of item order on some item parameters (Whitely & Dawis, 1976; Yen, 1980; Kingston & Dorans, 1984).

This recent research raises some important issues. For one, if item order has a significant effect, then some of the results of previous research may be questionable since they may have lacked the power or statistical sophistication to detect an item order effect. Previous concerns and conclusions may have to be re-examined in light of new findings. Of course, this discrepancy between the latent trait model findings and some of the classical model findings may be due to fundamental differences in the types of factors under study.

This research may also bring into question a basic premise of the latent trait model that each item is locally independent. If item order has an effect such that the probability of getting one item right is effected by the probability of getting some other question right, then the assumption of local independence is violated. Therefore, any test that did show item order effects would not be perfectly suitable for using latent trait model item parameters. Hopefully, latent trait models are robust and can tolerate small violations of some basic assumptions; however, the extent and effect of this source of error needs further research.

Research into item order must investigate several areas of growing concern. Not only must the item order effect reported by latent trait model studies be examined, but also the source of the mixed results among the classical model studies must be considered. It is unclear if perhaps the classical model studies lacked the power or the sensitivity of the latent trait studies or if some of the classical models did not properly control a factor in the design of their studies which may have influenced the results they obtained. Nevertheless, comparisons between present and past research is called for to shed further light on an ongoing measurement problem.

One study, using classical statistics, that did find a significant effect from item order suggested that previous studies without significant results had been in error since they did not control for within-subject rearrangement of test item order (Hambleton & Traub, 1974). If subjects are allowed to skip hard questions and do the easy ones first then Hambleton and Traub reasoned that the effect of item order would be masked and the differences between item order arrangements would appear to be insignificant. Hambleton and Traub, as a result, developed a mathematics test with a test booklet format that prevented within-subject rearrangement. Their significant results do question the validity of the findings of previous research. However, the lack of a control group that did not have a restricted format in Hambleton and Traub's study limits the generalizability of their findings.

However, if Hambleton and Traub's findings are correct, then within-subject rearrangement is perhaps a random error factor that may have been causing the inconsistent results in this field. In addition, within-subject rearrangement may be such a significant factor that all examinees should be made aware of its potential so that they might use it when they are taking a test, just as examiners must be aware of its error causing abilities when they design a test.

There is a need to replicate Hambleton and Traub's research of examinee control of item order, but with a control group, to determine if item order and examinee control of the order are significant factors. In addition, both latent trait and classical statistics could be used in the analysis to determine if the results are duplicated with both types of statistics.

Another area of concern is the effect that item order has on low achieving students. Surprisingly, even though much of the concern over item order involved this interaction effect with low achieving students, only four studies included this as a factor (Sax & Cromack, 1966; Klosner & Gellman, 1973; Hodson, 1984; Allison, 1984). The results were inconclusive, but further research was recommended (Klosner & Gellman, 1973).

Purpose of the Study

This study replicated the procedures of the item order research of Hambleton and Traub (1974). In addition, this study examined the effects of within-subject rearrangement, as defined by Hambleton and Traub, by using an additional group of students as a control group who were not given test booklets with restrictive directions.

Another aspect studied was a comparison of the performance of high ability students with low ability students. An easy to hard arrangement has been generally believed to be of benefit to low ability students who are supposedly easily frustrated by the hard to easy arrangement. On the other hand, high ability students may be able to avoid this frustration by using within-subject rearrangement when test booklets do not prevent such changes to the item order. This study used a large sample of students with a wide range of ability levels to compare the performance of students with different ability levels under easy to hard or hard to easy item difficulty sequences and under restricted or unrestricted test booklet formats.

A final aspect studied was a comparison of the results of item difficulty statistics based on classical test theory with item difficulty statistics based on latent trait statistics. This was done to determine if studies using latent trait statistics are comparable to studies using classical based statistics.

Chapter II

Review of Literature

Initial Studies

Initial research did tend to support the view that the context of a test item could influence the score of the examinee. Mollenkopf (1950) was the first to study the effects of changing item order. In addition, his study examined the effect of reducing time limits. He used 382 grade 11 and 12 students who were divided into four groups to take one of two forms of a combined verbal and mathematics exam. Then each group was assigned to finish their test under one of two timing conditions. Item order was modified only slightly. The tests were rearranged by sections with each section arranged internally from easy to hard, and with content areas kept together. The time limits had more substantial changes. The time limits were 1 hour 45 minutes for half the students, while the other half were given only 35 minutes. The group with the short time limit was, however, allowed to complete the test with a different coloured pencil.

Most of Mollenkopf results were as he expected. For one, changing the order of whole sections did not cause any changes in the performance of the students on the mathematics test. In addition, decreasing the time limits caused a deterioration in performance. Mollenkopf recommends that to get useful test statistics, time limits should be long enough to allow at least half of the students to complete the test. One unexpected result was that with only minimal re-arrangement of the verbal test items, there was a statistically significant change in the difficulty level of those items ($p < .05$). Items placed at the end of the test had a lower proportion of correct responses. He dismissed this finding as a small insignificant error possibly related to fatigue that could be ignored by test developers.

Another early study into item order was an unpublished report by K. MacNicol in 1956. It was cited by several authors (Flaugher, Melton, & Myers, 1968; Monk & Stallings, 1970; Hambleton & Traub, 1974; Plake, 1980; Hodson, 1984; Leary & Dorans, 1985; Lane, Bull, Kundert, & Newman, 1987). According to Leary and Dorans (1985), MacNicol randomly gave 1,500 high school students one of three forms of a verbal analogies test. The mean of hard to easy arrangement was significantly lower than the easy to hard arrangement whereas the random arrangement was not significantly different from the easy to hard arrangement. Unfortunately, the 30 minute time limit on a 50 item test may have been a factor, particularly since some students reportedly did not finish the test.

Further research was conducted by Sax and Carr (1962), who used 325 college freshmen taking two forms of the Henmon Nelson Mental Ability Test for College Students. The test was arranged in two forms. One form had the test unaltered with the difficulty levels alternating among easy, medium or hard, and with the content categories intermixed. This type of arrangement is called spiral-omnibus form. The other type of arrangement was to regroup the items into their three content types, vocabulary, mathematics, and spatial relationships. All students took both forms.

Sax and Carr tried to reduce the effects of speed by increasing the time limit from the recommended 30 minutes to 40 minutes. Unfortunately, many items were not completed by the students, so speed was, unfortunately, a confounding factor.

The conclusion found by Sax and Carr was that the order of the items did make a difference ($p < .001$). Students got more answers correct with the spiral-omnibus format. In addition, students omitted fewer items at the end on the spiral-omnibus form. The most significant number of omissions occurred in the mathematics section of the content based test. They concluded that the presence of increasingly complex items tends to discourage students from responding to the more difficult items.

Critical Re-examinations

It was a classroom teacher who wanted to solve the practical problem of whether or not he could randomly rearrange test items from a test item bank to create several forms of the same test. He wanted to have two forms of the test in class to prevent cheating in crowded situations, and he wanted to change his test format over the years without writing all new items each year or jeopardizing the security of his items. M. H. Brenner (1964) used the results of his Educational Psychology 407 midterm tests to compare the reliability, discrimination and difficulty statistics of rearranged pairs of tests administered over four terms.

Brenner compared easy to hard arrangements against hard to easy arrangements. As well, he compared an easy to hard order on the first ten items with a hard to easy order on the first ten items. On both forms, the last thirty items were in random order. Unfortunately, he did not report the number of subjects involved, nor did he adequately describe the statistical tests that he used to analyze his multiple comparisons.

Brenner reported only one significant statistical differences in twelve comparisons. One pair of tests had significantly different ($p < .05$) discrimination indexes. However, without an adequate description of the type of t-test used, this could merely be a chance event. Since his results indicated that changing of the item order did not make a difference in student performance, he recommended that college instructors not bother arranging test items based on item difficulties.

The generalizability of this study is limited since fourth year education students taking a required course are a very motivated and sophisticated group of students. It seems unlikely that such students would become discouraged by any arrangement.

Other researchers also wanted to find out if item order made a difference, especially G. Sax whose first study (Sax and Carr, 1962) found a significant effect from the item arrangement. In his earlier research he had found that students writing a spiral format test did better than students who had the more traditional increasing difficulty arrangement. These findings were in contrast to the commonly held view that recommended easy to hard arrangements (Ahmann and Glock, 1963).

As a result, Sax and Cromack (1966) rearranged the Henmon-Nelson Tests of Mental Ability into the following four forms: easy to hard, hard to easy, spiral, and random.

The four forms were then administered to 467 first year college students who were allowed one of two time limits. Half the students had a generous 48 minutes which is 18 more than the manual suggests while the other half had the suggested 30 minutes. In addition, cumulative grade points of all students were used as a covariate factor in the analysis of the results to divide the group into high ability and low ability.

Predictably, Sax and Cromack found that students given more time performed better on the tests. In addition, their study found that if a restrictive time limit was imposed, then the mean of the easy to hard form was significantly higher than the mean of the hard to easy arrangement (p. < .001). The other two arrangements were not significantly different. On the other hand, if there were longer time limits, then the arrangement did not make a difference. In comparing the results of students with high grade point averages and low grade point averages there were no significant differences for time or order except when students were give greater amounts of time and were answering questions on a hard to easy format test. In this case, high achieving students performed better. This unusual interaction was significant at the .05 level.

Despite finding that under certain circumstances low achieving students do in fact have more difficulty with one order than another, Sax and Cromack concluded "...little is gained in arranging items if time limits are generous." However, while it may be true that there is not evidence that one arrangement will help low achieving students, the higher scores by high achieving students on the hard to easy arrangement may indicate a difference in attitude that is related to item order effects. In this study, although low achievers did not seem to be discouraged, high achieving students may have been challenged by the unusual format and actually performed better as a result. Item order effects may affect people differently, and as a result have both positive and negative effects. Of course, this study is not indicative of a wide population since the college students involved probably performed within a very limited, but very high, range of achievement.

Until 1968, most studies on item order involved small classroom samples. Large scale test developers such as the College Entrance Examination Board needed to know if they could rearrange small banks of items on different tests without adverse effects. Flaugher, Melton, and Myers (1968) used the C. E. E. B.'s Scholastic Aptitude Test to test 5,000 college applicants with 4 different forms. The arrangements varied the easy to hard arrangement within blocks of five similar content items, and varied the sequence of the content based blocks. Students had 30 minutes to complete 40 verbal type questions and 30 minutes to complete the 25 mathematics items.

Although this study did not involve very significant changes in item order, they found that under their somewhat speeded condition, item order did make a difference on verbal items ($p <.001$). They did not find a difference with differing arrangements of mathematics questions. They concluded that since some of the relatively easy verbal items occurred last and were omitted by some students, differing numbers of unanswered questions were a factor. Item order, they felt, was an error factor, but an error factor smaller than the tests standard error of measurement. Nonetheless, this factor would have to be considered if tests involved item rearrangement.

Initial Studies: Conclusions

To summarize the findings of the early research up to the late 1960s, time limits were shown to have a definite impact. Item statistics became more prone to item order effects as more questions are omitted by the students. One effect of time limits in some studies was that questions which were not reached or omitted were given only a random chance level of being correct. This would cause easy questions that were completed by all students at the beginning of one test to be reported as more difficult if placed at the end where they would not be reached by students taking the hard to easy test. In addition, fatigue or frustration are some other factors that might account for easy questions at the beginning of one test seeming to be hard at the end of another. The more speeded the test becomes, the more error and uncertainty develop.

Anxiety and Item Order

Researchers in the late sixties began to turn their attention to more internal responses of the students. Concerns over the anxiety and stress level of the students prompted researchers to examine these variables.

One of the first studies to consider item order and stress was by French and Greer (1964). The study involved 152 first grade students. The students were given four different versions of the Pictorial Test of Intelligence in a counter-rotated order. The four forms were either easy to hard within subtests, easy to hard within the whole test, random, or a spiral of two easy and one hard. In addition to the P.T.I., students also took the California Test of Mental Maturity, the General Anxiety Scale for Children, and the Test Anxiety Scale for Children. The children were also rated on the P.T.I. Behavioral Rating Scale and measured for skin resistance on a polygraph recorder.

Despite a large amount of assessment techniques, not a large amount of data was reported. The only results that French and Greer reported to be significant were the data that indicated that regardless of the order, the first performance out of the four exposures was lower. There were not any increases in galvanometer readings as a result of changing the order. The authors concluded that if the P.T.I. was used with a similar group of students, they would not be sensitive to different item arrangements.

Unfortunately, their sample of students limits the applicability of the French and Greer study. For one, the reported I.Q. scores of the students ranged from 100 to 125. Most of the concern about item order involves the frustration of low ability students, but their sample did not include low ability students. In addition, first grade students may not have had sufficient school experience to find tests stressful or frustrating. Also, the item arrangements given did not involve the possibly most frustrating arrangement of hard to easy. The authors did mention that this sample may not have been sufficiently anxious enough, nor were the pictures of the test arousing enough to register any reaction on the galvanometer.

In addition to improving the generalizability of their study, French and Greer could have reported more details about their findings. For one, they did not provide the details of the difficulty level of the items used. Secondly, the results of the anxiety measures were not reported, and finally there was not a factorial analysis to determine if students with high anxiety had reactions to any particular order that were different than students with low anxiety.

Researchers not only began to take an interest in anxiety in the late 1960s, but they also began to use the F test statistics to look at both the main effects and the interaction effects. As a result, the interaction effect of high anxiety with hard to easy item difficulty sequence was not overlooked by Smouse and Munz (1968). In their study, 113 college freshmen were given one of three forms of a psychology final exam. The items on the exam were arranged by difficulty level either easy to hard (E-H), hard to easy (H-E), or random (R). In addition, the Multiple Affect Adjective Check List, a test for anxiety, was included at the end of the test. The three different item groups were randomly assigned to two different test directions groups. One group received anxiety provoking information concerning steps to prevent widespread cheating along with their directions for the test while the other group just received neutral, non-arousing test directions.

Smouse and Munz (1968) found no significant differences in test results among any of the groups. The item arrangement did not make a difference, the type of anxiety group did not make a difference, and the interactions between those factors were not significant. The results of the anxiety measure at the end of the test only showed that everyone was highly anxious. The authors were disappointed with their result and concluded that any differences that could be caused by anxiety were possibly masked by the already highly anxiety producing situation of a final exam. They also reasoned that there may be individual reactions to tests which can be effected by item sequence.

In a follow-up study Munz & Smouse (1968) used the Achievement Anxiety Test and their psychology final exam to compare 120 college freshmen. The students had to take one of three tests with the difficulty levels arranged (H-E), (E-H), or (R), as in the previous study. However in this study, the students results were divided according to their A.A.T. scores into four groups of anxiety levels for statistical analysis.

While Munz and Smouse (1968) did not show a significant main effect for item order, there was a fairly complex interaction between anxiety level and form of the test ($p < .01$). One group labelled "non-affecteds" performed lowest on the random arrangement but highest on the hard to easy arrangement. Another group, the "high-affecteds", did best on the random but worst on the easy to hard. The "debilitators" did poorly on all forms while the "facilitators" did well on all but the hard to easy form. The hard to easy form was found to have the smallest variance with no significant differences between the groups.

Munz and Smouse's conclusions were based on an inverted-U level of arousal theory. If arousal levels are increased, some people will respond well while others will respond poorly. For each person there is an optimum level of arousal that serves as the peak of an inverted-U graph of their performance. The different formats provided different levels of arousal and, as a result, different levels of performance. Increases in performance by certain anxiety types tended to be cancelled out by decreases by other types, so the main effect was not significant. However, Munz and Smouse noted that the hard to easy arrangement had the least variance, so they recommended that the hard to easy sequence would be the best format to use to minimize personality variables. In a follow-up study, Smouse and Munz replicated their study and findings, but tempered their conclusion with a caution that even though the hard to easy sequence may eliminate some unwanted variance, it may introduce other test taking contaminants not examined by their studies (Smouse & Munz, 1969).

Another follow-up study by Munz and Jacobs (1971) also essentially replicated the procedures and the results of the study by Munz and Smouse (1968). One change of procedure was to try and determine the difficulty level of each item by subjective judgement procedures rather than the typical 'p' level statistical method. This was an attempt to address the issue that if the subjective difficulty of an item differs among individuals, then arrangements of difficulty based on 'p' level may not be actually sequencing the items for the subject as the researcher intended. They asked 142 psychology students and 9 instructors to rate the difficulty level of each question which would be administered to the second group of 133 college students on their psychology final exam. The inter-observer agreement showed a moderately high relationship, and was reported as $r = .62$. The average difficulty of the group of items was reported as not differing significantly. They felt their results with this new procedure demonstrated that the difficulty of the test item was more complex than was reflected by the typical 'p' index.

Oddly, Munz and Jacobs (1971) did not report the correlation between the subjective difficulty ratings and 'p' index ratings. In addition, their findings did not seem to differ from studies which did use the 'p' index. As a result, the increased effort in obtaining subjective ratings does not seem to be necessary in examining the effects of item difficulty sequence. This is particularly the case if there is a strong correlation between the 'p' index and the subjective rating. Of course, as Munz and Jacobs concluded, further research on the relationship of subjective difficulty levels with other variables is needed.

The interaction effect that Munz and Smouse (1968) found with achievement tests and anxiety types was not found to be present with ability tests (Berger, Munz, Smouse, & Angelino, 1969). Berger et al. used a format similar to that of the previous studies by Smouse and Munz. They had the three different forms of item difficulty and identified the four different anxiety types. However, they used the Henmon-Nelson Test of Mental Abilities with 330 high school students rather than a college final exam. In this study they also had two test instruction conditions to hopefully generate 2 levels of anxiety. One group of students received instructions that the mental ability scores would be used on their permanent record while the others were told that the scores were to be only used for research purposes.

Not only did Berger et al. not find any interaction effects, but the main effects from changing item order and the main effect from giving different test instructions were also found to be non-significant. However, anxiety type was a significant factor. "Facilitators" scored highest followed by "non-affecteds" then "high-affecteds" with "debilitators" last. Berger et al. concluded that the item difficulty sequence does not effect ability tests. However, they felt differential reaction to test taking anxiety does have a significant effect on aptitude tests.

Several explanations were given by Berger et al. for this lack of effect. For one, there is the possible stability of aptitude tests. Another possible explanation is that the high school student population has a greater variation of intelligence, test taking ability, and test taking motivation. This may be true for all high school students or for just this sample, and the authors are remiss in not providing details of the intelligence scores and socioeconomic status of the sample especially since that information was reportedly gathered. Such information might also indicate if there was an unusual lack of variation in the intelligence scores of their sample. A lack of variance could mask any changes caused as a result of changing the order since those changes are supposedly most noticeable among the students with the lowest ability.

Further attempts were made to replicate the studies of Smouse and Munz. Marso (1970) reported two studies that examined anxiety and item sequence. His first study involved 122 first year college students randomly assigned to three item arrangements (E-H, H-E, and R) with item difficulties ranging from 0% to 100%. The students had to complete 139 questions of the Quick Word Test. In addition, each student completed a series of test anxiety measures prior to taking the Q.W.T. so that their test results could be grouped into high, average, or low anxiety test taking groups. There were no time limits, but a record was kept to determine how long each student took to complete his or her test.

Marso's analysis of variance found no significant effect from item order nor from an interaction of item order and anxiety level. Also, neither item order nor anxiety level had a significant impact on the length of time that students used to complete their test. However, anxiety level was found to be a significant factor in the level of performance ($p < .01$). The most anxious students had the lowest scores.

Marso's second study found similar results. Only the anxiety level was found to be a significant factor in the level of performance ($p < .01$). The second study involved 156 college students writing their psychology final exam. As in the previous study, students were grouped for analysis into high, average, and low anxiety levels based on a series of anxiety tests. This experiment was, however, quite different in the arrangement. The actual item difficulties were not used, but the order in which the topics were presented in class was used for the basis of the arrangement. One test had the items presented in the order that their content was presented by the teacher, another test was presented in a reverse order of presentation, and a final form was arranged in random order.

Marso's conclusions from the two tests were that tests without time limits do not have to be arranged in difficulty order, or in order of class presentation, or in groups of similar content. These conclusions are made even though his first experiment with the Quick Word Test may not have had subjects who were motivated enough to experience typical classroom levels of anxiety and frustration. His second study could have involved high levels of anxiety, but the tests were not arranged by difficulty levels. In particular, the hard to easy sequence was not tested.

In contrast to Marso's study is the Towle and Merrill (1975) study, which also was an attempt to replicate the work of Munz and Smouse (1968), comparing different anxiety levels and item orders. Towle and Merrill used 82 volunteers from educational psychology courses and community college mathematics courses to take the Florida Statewide Twelfth-Grade Mathematics Achievement Test. The test had a reported wide range of difficulties, and the items were arranged in the three common difficulty patterns (E-H, H-E, and R). The students were also reported to have a wide range of mathematics abilities. In addition to the mathematics test, the students completed the Achievement Anxiety Test and the State-Trait Anxiety (S.T.A.I.) prior to the mathematics test and the S.T.A.I. after the mathematics test. Towle and Merrill found no significant interactions, and anxiety levels were not a significant factor. However, unlike the previous studies, they did find that the order of the item difficulties was significant ($p < .05$).

Towle and Merrill concluded that the significant effect of item order was a result of the slight time limits placed on the exam. Many of the students did not have time to consider every problem. Another significant result was the increase in anxiety level as recorded on the S.T.A.I. post test. This anxiety increase was however not related to the different item orders, so none of the orders seemed to increase anxiety more that any other. While the inverted-U theory of Munz and Smouse (1968) was not supported with a significant interaction, Towle and Merrill felt that the data showed a tendency that would indicate the presence of such an effect. They felt that the effect may have been masked by the fact that the anxiety level groups are based on sample norms in each study rather than standardized norms. As a result, an individual could be placed in a different anxiety level group in each study depending on the sample used in the study.

In summary, research on item order and its relation to anxiety was not proving to be a fruitful line of research. For one, the lack of anxiety level norms limited the generalizability of the findings. Further, only three studies found a significant interaction (Munz & Smouse, 1968; Smouse & Munz, 1969; Munz & Jacobs, 1971). The other five studies, including two with Smouse and Munz, failed to find an interaction. One, however, did find a significant effect of order alone, but as with other studies, it involved a test with time limits.

Achievement and Item Order

An important area of interaction research is with achievement levels and item sequencing. It has been theorized that students with average or below average achievement would perform most confidently and efficiently with an easy to hard arrangement (Ahmann & Glock, 1963). Few studies have considered the effect of this factor. Sax and Cromack (1966), as previously reviewed, reported that high ability students did better on a hard to easy test if they were also given a generous amount of time.

Klosner and Gellman (1973) ranked 54 graduate students on the basis of their midterm marks as either high or low achievers. Students were randomly assigned to take one of three forms of the final exam. The item format was arranged using item difficulty in a typical classroom manner. One test was an easy to hard order within similar content groupings. Another was random order within similar content order. The third format was the more common easy to hard arrangement. There were no time limits or item difficulties reported.

Klosner and Gellman did not show a significant interaction or main effect of item order. However, the authors felt that the interaction was almost significant ($p < .15$) and therefore showed a trend. The low achieving students seemed to do best on the easy to hard with the content groups arrangement. Their suggestion that further research should proceed is indeed warranted since there is reason to believe that their study lacked power due to the small sample size. In addition, a more generalizable study should be done since the sample of this study was high ability graduate students who probably do not demonstrate some of the typical behaviour patterns of students normally considered low achievers.

One typical characteristic of low achievers is their low achievement motivation. Although Kestenbaum and Weiner (1970), did not examine specifically various achievement levels, they did examine achievement motivation. In addition, they examined the relationship between item order, test anxiety and achievement motivation. They used 79 seventh and eighth graders who were administered a reading test in either random or easy to hard sequence. Also, the students were administered the Test Anxiety Scale for Children and the Children's Achievement Motivation Scale. There were no significant effects as a result of different order, but there was a significant correlation between motivation, anxiety, and performance scores. They did not report how motivated the students were in the study, but they did conclude that highly motivated students with low anxiety tend to persist at endeavours despite failure. It would have been interesting to see the interactions that may have resulted in this study if it had included the possibly most frustrating sequence of hard to easy.

Hodson (1984), however, did use various achievement levels and a hard to easy sequence. He compared 157 students between the ages of 16 to 19 who were taking the British school system's A-level exam in chemistry. Unfortunately, due to the highly academic and competitive nature of the exams, the students who took the exam were reported to be high ability students and highly motivated students. Nonetheless, the students were grouped into three different ability levels based on their previous O-level exam results. The students were given one of three tests arranged in the typical formats: easy to hard, random, or hard to easy.

The students who took the hard to easy arrangement had the lowest average mean, 26.6 on the fifty item test. In addition, most of the students who failed to complete the test within the time limits had taken the hard to easy test. Item order had an effect since there was a significant difference between the means of all the tests ($p < .01$). The mean of the easy to hard test was the highest, 31.5, and the mean of the random format was 29.6. The ability level was also reported as a significant main effect ($p < .01$), but there were no significant interaction effect with the item order and the ability levels. Sex of the student was also examined as a factor, but no significant main effects or interactions were found.

Surprisingly, Hodson concluded, "Apart from a slightly inflated mean score, which might have some motivational value, there was no evidence to support the practice of presenting multiple choice chemistry questions in an easy to hard sequence." However, his findings show more than a slightly inflated mean score. They show a significant difference between test scores. The three tests are clearly not equivalent. From an academic student's point of view there is also a very significant difference between receiving a score of 31.5 on the test as compared with a 26.6 on the test and possibly not having enough time to finish. Such a difference on a crucial exam could have very significant effects on a student's future as well. Despite the suggestions by Hodson, chemistry teachers should in fact spend the effort to sequence items in a way which at the very least avoids the theoretical discouragement of the hard to easy sequence.

Other Interactions

The interactions of item order with various other unusual factors were also of particular interest to B. S. Plake. Plake was involved in five studies. Although other factors were examined, the primary area of interest in three studies was the effect of item order and student's knowledge of that order, with the student's performance and perceptions.

Plake's first study (Plake, 1980) used 104 psychiatric nurses taking three forms of their midterm exam, easy to hard, spiral, or random. Half of the nurses were given information in the test instructions about the order and strategies to deal with the order. The other half received no such extra information. All students had to complete a questionnaire at the end of the test describing their perceptions of the test order, their performance, and their expected score. Unfortunately, an adequate description of his measure of student perception was not provided, so there are some questions about the validity of this measure.

In this first study, Plake (1980) did not find that item order or knowledge of the item order significantly effected either the students' test scores or their reported perceptions of the test. Plake did admit that the type of examination and the type of student used limited the generalizations possible from this study. However, she did theorize that anxiety may have been an interacting factor. She proposed that knowledge of easy to hard order may have caused the performance of the highly anxious to drop and offset the rise in performance by the less anxious student.

Anxiety along with item order and the knowledge of that order was the focus of the next study (Plake, Thompson, & Lowry, 1980). Anxiety was measured as in earlier research by Towle and Merrill (1975). They used the Achievement Anxiety Test before the exam and the STATE and TRAIT Anxiety inventories as pre-tests and post-tests. Item order, knowledge of order and student perceptions were the same as in Plake's first test. In addition, two different scoring methods were used, number right or elimination. Knowledge of the scoring method was also a factor in the study with the half of the students who did not receive information about the order receiving the information about the scoring system instead. Unfortunately, this procedure does not provide for a control group to not receive information about scoring or order. The subjects were 97 educational psychology students who volunteered to take the A.C.T. College Mathematics Placement Program test. The students received course credit for volunteering.

With so many factors, some of the results of Plake, Thompson et al. were quite difficult to interpret. For example, the interaction of anxiety condition with knowledge of order with using the number right marking system was significant. The authors admitted that such a result may not be meaningful since the interaction between anxiety and the knowledge of order was already shown to be non-significant. The implications of the other findings were a bit more clear. The main effect of order was not found to be significant. However, order did interact with the pre-test and post-test anxiety scores to produce a result that was almost significant ($p < .10$). The lower post-test anxiety scores hints at a trend which might indicate that some item arrangements increased pre-test anxiety whereas others may have caused a decrease. The authors state:

> Trends in the data do support the presence of some possible effects. Therefore, conclusions based on the results of this study should be tempered by the knowledge that the lack of significant effect may have been due in part to insufficient power and/or

> motivation in the research design. (Plake, Thompson, & Lowry, 1980, p. 218)

It is the concern with motivation that characterizes her third study on the effects of the knowledge of order (Plake, Ansorge, Parker, & Lowry, 1982). To get a larger and more motivated sample, the authors used 170 senior and graduate students enrolled in an introductory statistics course. The study was arranged in a similar manner as the previous study by Plake, Thompson, and Lowry, there were 3 common item arrangements (E-H, R, S), half the students were given information about the order. In addition, the same mathematics test, anxiety tests, and perception questionnaires were used. The differences between the two studies were that the students were also given a Revised Mathematics Anxiety Rating Scale, there was only one scoring system used, and to motivate the students, they were told that the results of the mathematics test would be used to determine which students would qualify for extra remedial mathematics classes. An additional change was that all of the tests were grouped by sex for factor analysis. This was done to examine the influence of sex on mathematics test performance and to examine the interactive effects of anxiety, sex of the subject, item arrangement, and knowledge of arrangement.

Even though the students in the Plake, Ansorge, et al. study had 75 minutes to complete a 48 item test, and even though the volunteers in the previous study by Plake, Thompson, and Lowry easily finished the test in the allotted time, these highly motivated students did not complete all items. Twenty percent of the students failed to finish the mathematics test. Due to some problems with directions, the two perception questions were also left blank by many students. As a result, a planned power test became a speeded test. The effect of time limits on order has been well established. As previous research has shown, when there are significant time limits, item order can have an effect.

So, with time limits involved, Plake, Ansorge et al. found an item order effect. While there was no significant main effect from item order, there was an unusual but significant interaction effect ($p < .007$). Not only did males out perform females on all mathematics tests combined ($p < .002$), but factor analysis showed that males actually did best on the easy to hard order and the random order while both males and females performed equally well on the spiral arrangement. The authors suggested that further research would be needed using non-mathematics tests and longer time limits.

Item order was also a significant factor in the self-reported perceived performance and perceived test difficulty. Although none of the pairwise comparisons were significant, the random form showed a tendency to have the lowest perceived difficulty rating and the highest performance rating.

One lesson that might be learned from the Plake, Ansorge et al. study is to not assume that a power test will in fact be a power test. Sometimes, it is the students who will determine if there will be time limits. Despite providing what one might consider to be generous time limits, and therefore using assumptions about power test, if a number of students do not finish, then the assumptions about time limits and item order effects might apply.

The unusual findings of Plake, Ansorge et al. that item order interacted with the sex of the subject on a mathematics test was further examined in two follow-up studies (Plake, Melican, Carter, & Shaughnessy, 1983; Plake & Ansorge, 1984). Both tests used college students who were writing a psychology final exam. Plake, Melican et al. had 167 students write tests composed of three sections with each section having questions that would be arranged in one of the following three ways: easy to hard, spiral, or random. While the item order and sex of the subject interaction was not found to be significant at the .05

level, it would have been significant at the .10 level. There was a significant effect in connection with the sex of the subject with the 128 females receiving higher scores than the 39 males ($p < .05$). On the other hand, Plake & Ansorge did not find any such sex of the subject effect with their 279 female and 73 male students.

Both of these studies concluded that the findings of Plake, Ansorge et al., which involved higher scores by males on a mathematics test, were not applicable to non-quantitative tests. It was noted that comparisons were difficult to make since the tests in all the studies were not equally difficult. For example, the test used by Plake and Ansorge had a difficulty rating of .67, but the tests used by Plake, Melican et al. had difficulty ratings between .32 and .48.

Another area which causes difficulty in making comparisons, which was not mentioned by the authors of either study, was the non-random nature of their samples. Unfortunately, their samples were based on students who took certain courses. It does not seem appropriate to equate males and females who take a statistics class with males and females who take a psychology class. For a variety of reasons, these two samples may be different. The males or females in either course might not be the same as the males and females in the total population. Conclusions about gender differences must be taken with extreme caution.

Another item order study involving female and male students taking an educational psychology examination was by Klimko (1984). Additional factors that were considered were anxiety level and cognitive entry level. Cognitive entry level was defined as prerequisite types of knowledge, skills and competencies which are essential to the learning of a particular new task or set of tasks. Cognitive entry was measured at the beginning of the course using a forty-five item test which was designed by Klimko. There were 93 female and 18 male college students who were randomly assigned to midterm examination formats containing the three common item arrangements (E-H, H-E, R).

Klimko found that cognitive entry characteristics was the only factor with a significant relationship to the performance score ($p < .0001$). Item order, sex of the subject, and anxiety level were not significant factors. His main conclusions was that item order does not influence achievement examination performance. He also concluded that cognitive entry was a meaningful predictor of achievement performance. Unfortunately, he provided little in the way of specific information detailing the parameters of cognitive entry characteristics. His forty-five item test of cognitive entry characteristics had no validity data other than its results were similar to a psychology midterm examination.

Another factor based on cognitive theory was by Lane, Bull, Kundert, and Newman (1987). They used Bloom's taxonomy to subjectively determine the "cognitive difficulty" of every test item in a forty item education course examination. The test items were arranged in five different formats. Two forms had the items with increasing cognitive difficulty and either increasing or decreasing statistical difficulty. Two other forms had the items with decreasing cognitive difficulty and either increasing or decreasing statistical difficulty. A fifth format had the questions in random order. These tests were used in two different studies.

In the first study by Lane et al., 59 male and 96 female college students wrote one of the five different examination formats. Although item order was not a significant factor in the total scores, some subscores did seem to be effected by the item order ($p < .05$). Students who had the easy knowledge items first followed by questions of increasing statistical and cognitive difficulty had the highest mean scores. The lowest mean scores were received by students who had exams with decreasing cognitive difficulty and increasing statistical difficulty. Another unusual finding was that gender was a significant interactive and main factor with females scoring higher than males.

In the second study, Lane et al. used only three of his original five formats. They used the form with increasing cognitive and statistical difficulty, the form with decreasing cognitive and statistical difficulty, and the form with the random order of difficulty. In addition, half of the tests had labels on each item to indicate its cognitive difficulty. The six forms were randomly given to 78 male and 169 female college students as an exam in an education course.

The results of the second study differed from the first since item order had no significant effect with either the total score or with the subscores. However, as in the first study, females had higher scores than males. The presence or absence of labels was also a significant factor. Both males and females had higher scores when the labels were present. In addition, there was an interaction effect between gender and labelling. Although all students did better when labels were included, the discrepancy between males and females decreased when the labels were included.

Lane et al. concluded that the presence of labelling was beneficial and could possibly negate the effect of item ordering based on item difficulty. However, the generalizability of this study may be limited by the fact that only subjects who understood Bloom's cognitive levels took the test. Lane et al. also suggested that further research was needed in the area of performance by females since this research contradicted the results of Plake and Ansorge (1984) on the performance of females on non-quantitative tests.

The effects of item order and sex of the subject were studied by Kleinke (1980), but speed was also included as a factor. He had 484 fourth grade students complete a thirty six item social sciences test. The test was presented in either an easy to hard arrangement or a uniform spiral arrangement. The students were given twenty minutes to complete the test, but after only ten minutes, students were asked to draw a line to indicate which question they had just finished. This provided data on each student's speed and accuracy under very speeded conditions. Unfortunately, only 314 students followed the directions and drew the line.

An additional area of study by Kleinke was response location. The position of item responses were placed on the left side of half the booklets and on the right side of the other half of the booklets. This response positioning was compared to the handedness and sex of the student to see how these factors related to the student's performance on the test.

Kleinke found that the easy to hard item arrangement had a higher mean under the speeded ten minute condition ($p < .01$). However, the total scores were equal under the twenty minute condition. He concluded that if an examinee had ample time to complete a test, they will persist no matter what the arrangement. He was not able to draw conclusions on some of his other findings. For one, males had higher mean total scores, higher scores after ten minutes, and more questions complete after ten minutes. He also had no conclusion for his finding that left handed pupils had higher scores after ten minutes and more questions complete after ten minutes. Other effects and interactions were not found to be significant.

It is unfortunate that 170 subjects failed to draw a line at the ten minute point of their test. Conclusions must be limited by the fact that the remaining 314 are a less than random sample. None the less, it can be concluded that some students will get more questions correct at the beginning of a test if they begin the test with easy questions in comparison with some students whose test begins with items of varying difficulty. However, this study did not address the issue of what effects would a series of difficult questions have on a student's total score.

Interaction Research: Conclusions

The research into the interaction of item order with other measures did not produce clearly significant results. For example studies involving anxiety showed some initial success (Munz & Smouse, 1968; Smouse & Munz, 1969), but the effects were not confirmed by other researchers (Smouse & Munz, 1968; Berger et al., 1969; Marso, 1970; Towle & Merrill, 1975).

Although anxiety has been one of the more common factors included in item order research, other factors have also been studied in conjunction with item order. However most have had results of limited significance and limited applicability. For one, knowledge of arrangement was one line of inquiry that did not produce any significant results (Plake, 1980; Plake, Thompson et al., 1980; Plake, Ansorge et al., 1982). Klimko (1984) included cognitive entry characteristics along with sex, test anxiety, and item order. He found cognitive entry characteristics to be the only predictor of examination performance. Lane et al. (1987) had very mixed results when they included cognitive and statistical item difficulty along with knowledge of the item arrangement and gender. In one study, the ordering of the test items based on cognitive and statistical methods showed a significant effect. However, in their other study, when the test items were labelled with their cognitive level, there were no significant order effects, but there were significant differences related to the presence of labels and the sex of the subject.

The sex of the subject was also a factor in several other studies. Plake, Ansorge et al. (1982) found an interaction effect between item order and sex of the subject. Several studies did not find an interaction effect, but did find a significant main effect (Plake, Melican, et al., 1983; Kleinke, 1980; Lane et al., 1987). Other studies that examined the sex of the subject factor did not find it to be a significant factor (Plake & Ansorge, 1984; Klimko, 1984). The mixed results of these studies would indicate that one must be cautious in making conclusions about gender differences since there may be other factors involved with any reported observations.

Achievement level as an interacting factor did show a tendency toward significance in the Klosner and Gellman (1973) study, but unlike Sax and Cromack (1966), it was still found to be a non-significant factor with the sample used. On the other hand, Hodson (1984) did not find an interaction effect with ability level, but he did find a significant main effect for item order. Considering the importance of ability levels to the justification of sequencing questions from easiest to hardest, the lack of research in this area is surprising.

Any study using slightly speeded tests found significant results; results which confirmed previous findings. The speeded tests either caused an outright main effect for item order (Towle & Merrill, 1975; Kleinke, 1980; Hodson, 1984) or an interaction effect with the sex of the student (Plake, Ansorge et al., 1982).

Simple Effect Re-examined

The interaction of variables with item order was not the only concern of later researchers investigating the effects of item order. The simple effect of item order remained a concern since textbooks of the day continued to suggest that items be arranged from easy to hard and since classroom teachers still had questions about the equivalency of multiple forms. In addition, the use of multifactor analysis procedures, which allowed interaction analysis, could also be used for multiple comparisons of the simple main effect among several similar groups.

Huck and Bowers (1972) reported the results of two very similar experiments that compared the effect of several different random arrangements on item difficulty. In the first experiment Huck and Bowers used ten different random variations to a psychology final exam with 120 psychology students. In the second study they used six different random forms of a psychology midterm with 162 students. The difficulty ratings ranged from .00 to 1.00 in the first study and .15 to 1.00 in the second. The average of difficulty rating ranged from .54 to .61 in the first study's 10 tests and .66 to 70 in the second study's 6 tests. Huck and Bowers found no significant difference in the item difficulty ratings from any of the test forms. They concluded that the sequence effect hypothesis might not be a valid one and criticized other writers, "comments concerning a sequence effect should be somewhat qualified as compared with presently appearing statements."

This study was limited in its applicability. For one, as Huck and Bowers briefly mention, college students enrolled in a psychology class do not represent a random sample for the purpose of generalization. In theory, item order effects are most applicable to low ability students, and if low ability is a category reference relevant to any sample, then a college sample would have "low ability" students. However, if "low ability" is in reference to a large, age based population, then a college sample probably does not include any low ability students for whom item order might make a difference. Secondly, the effect of item order is supposedly the result of a lack of success (Ahmann and Glock, 1963). Therefore various random orders would have approximately equal amounts of easy questions at the beginning and correspondingly equal amounts of success at the beginning for the easily frustrated low ability student. Huck and Bowers' findings are therefore limited to only one type of item arrangement, random, and one type of ability level, high.

Sirotnik and Wellington (1974), on the other hand, did try to have their item order study generalize to a larger population, and they used a content based arrangements as well as random. The content based arrangements had items grouped according to their basic subject areas--either mathematics, social studies, science, language arts, or reading. In addition, they used a large grade based sample of 2,463 eighth grade students, so, by most definitions, low ability students were included. For an item pool they used the final exam questions from the five basic subject areas and systematically divided them into four tests arranged by content or four tests arranged in random sequence. This multiple matrix sampling design allowed them to rearrange one four hour test into eight one hour tests which they hypothesized would be equivalent. The one hour time limit did introduce a slight speeding effect that the authors felt was negligible but typical to the school system. The time limit was another attempt to allow their results to have a broad and practical application.

Most of Sirotnik and Wellington's tests of significance supported their hypothesis that there was no difference in the means, variance, item difficulties, or KR-20 reliability for most of their tests. However, they did find that the means of the reading tests were significantly different. The content based reading test was 1.5 percent higher than the random arrangement. Although this result was statistically significant, it was dismissed by Sirotnik and Wellington as not being of practical significance.

Once again, the effect of time limits comes into play with item order. Research involving typical classrooms and large numbers of students must contend with the very real problem of time limits. As a result, item order is likely to have an effect. In Sirotnik and Wellington's research, item order only seemed to effect the results in one subject area. There may have been other interactions that were not analyzed in this study such as sex or achievement level. Of course, the significance of the main effect could have been just a chance anomaly from the increased alpha error level due to repeated tests of significance. The study could have been improved with the use of appropriate significance tests for multiple comparisons. None the less, it is notable that the mathematics tests did not show an effect of item order under time limits which is contrary to some previous research (Towle & Merrill, 1975). Unfortunately, it is difficult to make accurate comparisons of the mathematics tests because Sirotnik and Wellington did not provide any details as to the difficulty rating of their tests and item, nor did they have the hard to easy variation of item order.

A study similar to the one by Sirotnik and Wellington is the study by Feldt and Forsyth (1974). They were, however, only concerned with the context effects caused by the matrix sampling techniques used with content based item groups. They used two forms of the Iowa Test of Educational Development on 530 students in grades 9 to 12. One group of language questions or one group of three different groups of mathematics questions were drawn from one of the two regular test forms and included as a special section in the other test form. As a result, eight different test packages were developed and tested. They found significant differences between the means of the mathematics questions ($p < .05$) and no significant differences between the means of the language questions. The means of mathematics questions in the special sections were higher than the means of the same questions when they were presented in the regular section.

They cited several possible reasons for their mixed results. For one, as with other studies, time limits seemed to have an effect. They said that even though students had enough time to finish the test, they may have felt more rushed to complete the questions when they were given in the regular form rather than when they were given in the shorter special section. Another possibility given was that the students may have been slightly more fatigued when answering questions when they were presented in the longer regular section. These conclusions are tempered by the fact that the effects were observed only with the mathematics sections but not with the language section. They felt that the mathematics sections may have been more rigorous. A final reason suggested was that item sequence effects may have been present in some subtle form since the item order of the questions in the special section was not identical to the order in the regular form.

It is unfortunate that Feldt and Forsyth did not control for the possible subtle effects of item sequence by administering different item arrangements of each special section. They also did not clearly address the issue of whether or not their tests were power tests or if they were speeded tests. They state that the non-completion of the tests was not a problem, but they also state that the omission of mathematics items was common. This results in their being able to claim that time was a factor when there was a significant effect, and claim that time was not a factor when there was not a significant effect. However, it is also possible that item sequence effects can be the cause of what seems to be a time problem rather than the effect. Results similar to speeded tests may occur with different item orders since students frustrated by one arrangement may tend to omit or fail to complete more questions than students experiencing a different order without the frustration. Time limits must be more carefully controlled and standardized to make research on item order effects more replicable and conclusions more certain.

Controlling Other Factors

Some of the shortcomings of previous research were more adequately controlled by Hambleton and Traub (1974). After conducting a fairly careful review of previous research, Hambleton and Traub concluded that several of previous studies which had not shown an effect from item order had not controlled three fairly important factors.

The first factor that Hambleton and Traub felt was not well controlled was the examinee's control of item order, or the within-subject rearrangement factor. It is commonly assumed that an easy to hard test or a hard to easy test are taken in that order. However, it may not be the case that students, in fact, do the items in the order that was intended by the researcher. Particularly in the case of hard to easy arrangements, many students may try to do the easy ones first and therefore statistically mask the effect of item order. While Hambleton and Traub did not have an estimate of how extensive this within-subject rearrangement was in the high school population they sampled, they attempted to prevent examinee control of the item order by instructing the students to do the questions in the order presented in their booklets. The booklets were also specially printed with only one question printed on each page to discourage students from searching for easy questions to do first and hard questions to do last.

The second reason that Hambleton and Traub felt other studies had failed to find an effect was due to the limited range of the item difficulties used in the test. They pointed out that no information on the variation of item difficulties was published. Unfortunately, Hambleton and Traub's article reported only the rank correlation between item position in the test and the position the item should have been in, based on item difficulty level as estimated from the data of their study. However, their test was a standardized and published test with item difficulty information available in the test manual.

A third factor Hambleton and Traub dealt with was student motivation. Hambleton and Traub contended that some of the previous studies may have lacked realistic or effective motivation by the subjects. They pointed out that most studies gave no clear description of how important the student perceived the tests to be. Item order effects, they felt, would be directly related to the importance a student attaches to the test. To attempt to insure student motivation, Hambleton and Traub reported that they told the students that the results would be used by their classroom teacher to arrive at a final grade. This point is also significant since Plake, Thompson, and Lowry. (1980) used volunteers and found no significant effects, but when Plake, Ansorge, Parker, and Lowry (1980) used motivated students in a similar study, they did get significant results.

With these three factors controlled, Hambleton and Traub felt that item order should have an effect. They hypothesized that the effect was caused by two factors. For one, fatigue, as suggested by Mollenkopf (1950), could cause the easier questions at the end of a hard to easy arrangement seem harder. Another possible cause was personality trait such as anxiety. As Munz and Smouse (1968) suggested, some arrangements such as hard to easy improve the performance of those students who need to have a higher anxiety level on the test while at the same time performance is lowered for those who are debilitated by the higher anxiety of the hard to easy arrangement.

To test their theories, Hambleton and Traub used 106 eleventh grade mathematics students enrolled in a mathematics summer school program. Their mathematics ability was considered low to average. They were given an Achievement Anxiety Test two weeks prior to the mathematics test to determine the highest scoring 25 percent and the lowest scoring 25 percent for analysis of anxiety trait reactions. On the day of the test all students were given the Cooperative Mathematics Test Algebra II published by the Educational Testing Service. The students were randomly assigned two forms, either easy to hard or hard to easy arrangements. The test booklets were designed to discourage students from changing the arrangement order. The students were told that the exam could be used for marks. To measure stress levels students had their pulse measured every 10 minutes. Unfortunately, in one class the pulse meter did not work, so pulse rates were collected for only half of the students. The test was reported as a power test, but some of the students did not finish, so the test must be considered as slightly speeded.

Hambleton and Traub's results indicated a significant effect due to item order ($p < .05$). Scores on the hard to easy arrangement were lower than on the easy to hard arrangement. While some students did not finish their tests the authors felt it was an insignificant number based on a chi squared test for contingency. They further analyzed the types of questions not completed and felt that the results would not be substantially enough different if all students had finished the test. As a result, they felt that speediness was not a plausible explanation of the results.

In addition, item order affected test anxiety. The hard to easy item order produced more stress as measured with the pulse meter than the easy to hard item order. This result is marred by the small numbers, and the analysis was difficult since the two samples had different means to begin with, but Hambleton and Traub did tentatively conclude that the order does effect the stress level. The performance of the two anxiety levels as identified with the A.A.T. did not show a significant main effect nor a significant interaction with item order. The results did show a trend with a level of significance between .05 and .10; however, the results of Munz and Smouse (1968) were not replicated.

Interestingly, this was one of the first studies to report an analysis based on sex. Hambleton and Traub did not find any significant difference between male and female students in this study even though in a later study by Plake, Ansorge, Parker, and Lowry (1980) there was an effect related to sex and a sex by item order interaction.

Hambleton and Traub concluded that item order does make a difference, and they caution against the practice of making several forms of a test in a class to reduce the chance of cheating. Two tests with different orders are two tests with different properties and therefore make comparisons invalid. They felt that the cause for this effect was from what Cronbach (1946, 1950) called response set. In a multiple choice format, students expected an easy to hard arrangement, so when a student begins with the hardest items, he becomes more anxious since his expectation is to have even harder questions as he progresses through the test.

Monk and Stallings (1970) were concerned about Hambleton and Traub's recommendation against using multiple forms in a class to prevent cheating. They examined the results of twenty-two tests they had administered over a three year period in their geography course. The tests were made from items chosen from an item pool and administered in pairs. Each pair had identical questions grouped together by content categories. There were no arrangements based on item difficulty, but each pair had their content categories randomly rearranged to discourage cheating. Nearly 2000 subjects were involved since each form had approximately 100 students writing it. The tests were slightly speeded as is typical in large scale testing situations.

They tested the significance of their result by repeated t-tests, and found that nine out of eleven pairs were not significantly different. They concluded that equivalent tests could be produced by rearranging items. However, they did concede that two tests were significant. One was at the .01 level of significance, and the other was at the .001 level. As a result, they cautioned that item order effects may in fact be present in large scale testing programs.

Monk and Stallings' analysis is marred by the repeated use of the t-test without a description of the type of t-test used. If an inappropriate t-test were used then, from chance alone there could be at least one significant result in the ten comparisons at the .05 level of significance. Comparisons with other studies are also limited by the absence of data on item difficulties and the absence of data on the ordering of easy questions in relationship to hard questions. These weaknesses hide the possibility that if the item difficulty range was minimal, if the changes in item sequence were minimal, and if the appropriate t-tests were used, then the fact that two tests showed any effect may be evidence that item order is a significant factor.

The research findings of Hambleton and Traub (1974) were also of concern to Allison (1984). It was noted by Allison that most of the previous research into item order which had involved students who were probably mature enough to use the item rearranging strategy cited by Tuck (1978). While Hambleton and Traub had tried to control this factor with a restricted test format, Allison simply chose a sample of young students. He felt that students in grade six would generally not yet use this strategy. In addition to being one of the few studies to examine item order outside of a college classroom, Allison's study was also one of the few studies to examine the interaction effect of item order with low and high ability students.

In his study, Allison randomly gave 364 grade six students a science exam arranged in one of the three common orders, easy to hard, hard to easy, and random. The items had a wide range of difficulty as recommended by Hambleton and Traub (1974). The items ranged in difficulty from .178 to .981 with a mean of .673. Students were given an ample 90 minutes to complete the 64 item test. Motivation was a factor that was taken into account since the students were told that the test was an important part of their program. High and low ability students were identified from I. Q. scores in the students school records. Thirty-five students did not have such information on file so only 327 students were used in the analysis. Their mean I. Q. was 113.31. Finally, as so many other studies have done, the scores of the 160 boys were compared with the scores of the 167 girls to see if item order has any interactive effect with gender.

In contrast to Hambleton and Traub's findings, Allison found no significant difference between the means of the three different test formats. Nonetheless, there were significant main effects associated with I. Q. and gender.

Boys and students with high I. Q. scores had higher means on this science test. Allison drew no conclusions regarding the two significant main effect. In examining the interactions, Allison reported that there were no significant interaction effects involving the factors of sex, I. Q., and item order. These results are similar to other studies where the tests were not speeded. Allison concluded that measurement specialists should hesitate before recommending any one item arrangement over another.

Since both the study Allison and the study by Hambleton and Traub were very thorough in controlling many of the factors involved in item order research, it leaves open the question why these two studies differed in their results. For one, where Hambleton and Traub did not have a control group of unrestricted students, Allison did not have a an experimental group with restricted students. Although his subjects were young, they may still have used the technique that was controlled by Hambleton and Traub. Another difference is that while Hambleton and Traub used a math test, Allison used a science test. Students may be more sensitive to a series of difficult math questions than to a series of difficult science questions. Science questions which are statistically difficult may not be subjectively as difficult for students. On the other hand, math questions may have a better match between subjective and statistical difficulty. As a result, difficult science questions may produce less of an item order effect. Of course the issue of speeded test as compared to power test may be a factor. Although both studies claim to be power tests, there was some suggestion that time may have been a factor in the Hambleton and Traub study. If time was a significant factor, then both studies merely confirm previous research.

Test Wiseness

The studies by Hambleton and Traub (1974) and Allison (1984) were the only studies of item order effects to suggest that subjects may possess some sort of skill in modifying the effects of item order. They suggested that the examinees could control the item order by answering the easiest questions first rather than answering the questions in the order presented by the researcher. This skill is one of several under the general category described as "Test-wiseness". Test-wiseness is defined by Millman and Bishop (1965) as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score." Hambleton and Traub reasoned that if too many subjects did the easiest questions first by omitting the hard questions until the end of the test, then the effect of item order on the students who did not use this strategy would be masked by the results of those who did use the strategy.

This conclusion of Hambleton and Traub is significant in light of a later studies. For one, Tuck (1978) questioned ninety psychology students about the strategies they used on multiple choice tests. He found that 69% of the students reported that they would seek out the easy questions first and leave the difficult questions until last. Further evidence of within-subject rearrangement was uncovered by Klimko (1984) in his item order study. Klimko included a self-report questionnaire at the end of a midterm examination with his 111 psychology students. He found that 43 subjects took the test strictly in the order given. On the other hand, 58 students went in order, but skipped over the hard questions to work on them at the end of the test. In addition, 5 students skipped around looking for any easy questions to do first, and 3 students flipped through the test to begin at a random point, and 2 students did not use any of the methods listed on the questionnaire.

In a very extensive study, Allison and Thomas (1986) gave a short questionnaire on the student's preferred strategy for answering achievement test items to 415 students from grade four through to fifth year university. All grades reportedly had some students who would use one of the three different types of strategies, either easy to hard, as presented, or hard to easy. The easy to hard within-subject rearrangement strategy was used by 58.4% of the students in the intermediate grades, 62.7% of the students in junior high school, 49.6% of the students in senior high school, and 59.6% of the students in their third to fifth year of university. Most of the remaining students used the "as presented" strategy. Although these results are evidence that students may use rearrangement strategies, Allison and Thomas conclude:

> Whether the students' own test-taking strategies supersede the item-difficulty sequence intended by the examiner is not clear. There does not seem to be enough evidence to doubt the majority of the studies on item-difficulty sequence simply because the actual sequence of responding to items was not controlled. In fact, it may be that the results of studies involving item-difficulty sequence can be more readily generalized to typical testing situations when students are allowed to use whichever strategy they usually choose. It is also possible to argue that students perceive the questions they can answer to be the easy items and the questions they cannot answer easily to be the hard items. In other words, it may be that the majority of students say, quite reasonably, that they answer the questions they can answer first and the questions they cannot answer are left until later. (Allison & Thomas, 1986, p.869)

It is certainly not clear if students can improve their scores by answering the more difficult questions later. Rindler (1980) had 160 college volunteers write a thirty item verbal aptitude test with special scoring sheets designed to identify if items were skipped. Students were also put into one of three different timed conditions, 20 minutes, 25 minutes, or 65 minutes. Grade point averages of all students were also obtained to divide the students into high, medium, or low ability rankings for a comparison of their performance.

The results indicated a complex interaction between ability groups and skipping questions. While there were some students who skipped in every ability group and under most timing conditions, only the high ability students who skipped questions had consistently higher scores under every timing condition. On the other hand, middle ability students who did not skip had the higher scores under every timing condition. In contrast, low ability students who skipped had higher scores only when they had 25 minutes for the test. Low ability students did not skip any questions when they had the ample 65 minute time limit.

Rindler concluded that some students may be better advised to spend more time on the easier questions which are usually at the beginning of the test. This tactic would help to make sure that those students would spend the most time on the questions they were most likely to get correct. She felt it was unfair to suggest to all examinees that their time would be used effectively if they skipped the difficult questions.

Another fact shown by this research is that using a test-wiseness strategy and using it well are two different skills. One of procedures associated with skipping difficult questions is to return to the skipped question. When low ability students skipped questions, they returned to only 5% of the questions skipped. High ability students, depending on the amount of time available, would return to between 20% to 98% of their skipped questions.

Latent Trait and Context Effects

The research into item order has been problematical. There are not enough clearly significant results to conclusively predict or dismiss item order effects. This uncertainty began to have an impact in the late 1970s as another new statistical technique gained increasing popularity.

Item response theory or latent trait model theory had several advantages over classical test theory that test constructors found attractive, but its statistics involved an assumption that each test item was locally independent. In other words, to be statistically effective, the context of questions before and after an item could not change the probabilities of how students responded to that item. Therefore, latent trait theory required that item order was not a significant factor.

Whitely and Dawis (1976) were the first to test this assumption of local independence and the effect of context on item parameters. They administered sixty verbal analogies tests to 1,568 junior and senior high school students. There were seven forms which had a common core of fifteen questions with each item in the same position on any one of the test. The other forty-five questions were unique questions developed for each of the seven tests to provide seven totally different contexts around the core items.

While this study was not designed to test if specifically the item order has an effect on students, it does address the issue of the effect that one question might have on another question. As a result, Whitely and Dawis' results do not indicate if some students do best with a test that starts with the easy questions. However, their results did show a significant difference in the rasch item parameters obtained for nine of their core items ($p < .05$). In addition, they found a significant difference between the means of the core items on two of their tests ($p < .05$). They concluded that item-parameter-invariant models must have their assumptions tested before developing equivalent measures.

Unfortunately, rasch item parameters are usually established in relation to the other items on the test. Since each test was different, the rasch parameter would be different for the core group of items. So Whitely and Dawis had to use an uncommon statistical method to establish a common point of reference for the core group of items before they could conclude that there was in fact a significant difference.

Whitely and Dawis raised some concerns by test constructors who were using latent trait models to pretest questions at one test administration and then use them later at another sitting using what were assumed to be invariant item parameters. As a result, Kingston and Dorans (1984) tested context effects within the pre-operational calibration of Graduate Record Examinations General Test. Their research design involved 1500 examinees who took one of two forms of the G.R.E. General Test. Each test had different questions divided into four operational sections of similar content. A fifth pre-operational section was composed of a random selection of questions from the other test form. Six versions of the first form were made with different questions from the second form in the fifth pre-operational section, and six versions of the second form were made also with different questions from the other form. As a result, every operational question of each form was used in one version of the other form in the pre-operational fifth section. Rasch model item difficulty parameter estimates for verbal, quantitative, and analytical items could then be compared between their operational and their pre-operational placements.

Kingston and Dorans did find some items to be affected by context while many were unaffected. Of verbal item types, items involving antonyms displayed a slight practice effect when they were placed after a series of such questions, while "reading comprehension" showed a slight fatigue effect if located in the final section ($p < .05$). Quantitative items showed little change with the exception that one item on one form in "data interpretation" was significantly more difficult when placed at the end of the test ($p < .01$). By comparison, in the analytical section, "analysis of explanation" and "logical diagrams" both showed significant effect from practice on both forms ($p < .01$). "Analytical reasoning" in that same section did not show any significant differences.

As with the previous research, Kingston and Dorans study does not show the specific effects of item order. However, it does indicate that some items can be affected by placement at either the beginning or the end of the test.

Of course, these findings are limited by the fact that students who take the Graduate Record Examination are not a random cross section, but may be unusually motivated and competent students. None the less, Kingston and Dorans felt that, for their purposes, they could conclude that location effects were specific to certain types of items, and the elimination, or the proper placement of those types of items would solve the problem.

Context Effects and Item Order

One study using latent trait item parameters did examine the specific effect of item order as well as more general context effects. Yen (1980) had 1,300 sixth grade students complete the California Achievement Test mathematics section while 1,100 fourth grade students completed the appropriate reading section. Each person used one of the seven different forms for either the mathematics group or the reading group. Each form was a different combination of five different sets of questions. Six of the seven forms contained set A questions which had a range of difficulties and relatively good model fit which were used as the common core of questions for item parameter anchoring purposes. Also randomly intermixed were set X on four of the forms and set Y on the other three forms. These sets were of primary interest and were questions with good model fit and discrimination, but more limited in their range of difficulty. The final variation of forms was that one test with sets A and X also had a set V while one form with sets A and Y also had a set W. Sets V and W had relatively poor fit and extreme difficulties or low discriminations. All items from the sets that were included in the form were intermingled so that some forms contained all identical questions but in randomly different orders while other forms had many identical questions in a different order but also had some additional questions that changed the context of the test items.

Yen found that changing the order or including additional questions did significantly alter the difficulty parameters and the discriminating parameters ($p < .05$). Item order effects were demonstrated by the fact that the greater the correlation of sequence the greater the similarity of item parameter estimates. Speediness was not considered a factor since 93 percent of the students did answer the last question of one booklet examined. Of course, some students did omit some items. As a result, their computer analysis gave only those students who omitted an item a chance level of answering the question, but ignored the missing answers of students who did not reach an item. This is one reason that Yen concluded that fatigue or impatience to finish rather than a computer scoring anomaly were possible causes of some questions near the end of the test seeming to be more difficult. However, not all questions at the end were more difficult, so Yen felt that other factors such as the content of the preceding item may have caused the instability.

Although Yen's study does indicate an effect, it is similar to other recent studies using latent trait analysis and does not address the issue of whether it is best to sequence test from easiest to hardest for the benefit of the low ability students. Yen's study did use a large and heterogeneous sample of elementary students, but her complex altering of context added other factors such as test length into her analysis that limit the conclusions about item order.

Summary and Conclusions

Thirty-seven studies published over the last 40 years show the continued interest in item order research and demonstrate the difficulty that any researcher would have in drawing conclusions. The definitive study has not yet been published, but the many attempts to settle the item order controversy have served to create a large pool of conflicting results.

Table 1 presents the research results in a format similar to the analysis by Leary and Dorans (1985). Unfortunately, results of item order effects do not always fall simply into the category of non-significant main or interaction effects (non) or the category of significant main or interaction effects (sig). Those studies that do show some findings that report results which to a limited extent show some possible effect of item order will be indicated as being non-significant with additional information (non +).

The tests in Table 1 were also divided according to type based on whether they were power tests, and reported all students finished the test, or timed tests, and reported that some students did not finish the test. This also is a difficult judgement to make since some studies gave no indication one way or the other. Studies which do not clearly state the effects of their time limits are indicated as power tests. In addition, a study which involved a large number of students could be essentially a power tests for most of the students but a speeded tests for a few. Such a test is listed as a timed test.

Finally in Table 1, although the number of students involved is one of the easiest judgments to make, compromises between brevity and accuracy were made. In the case of the study by Monk and Stallings (1970) the same experiment was repeated 11 times with 11 different small samples, but only the total aggregate number is listed.

Table 1

List of Item Order Research: Results, Types, and Samples

Results	Authors	Year	Type	Sample	n
Non	French & Greer	1964	p	elem.	152
	Smouse & Munz	1968	p	col.	113
	Berger et al.	1969	p	sec.	330
	Marso (a)	1970	p	col.	122
	Marso (b)	1970	p	col.	156
	Kestenbaum & Weiner	1970	p	sec.	79
	Huck & Bowers (a)	1972	p	col.	120
	Huck & Bowers (b)	1972	p	col.	162
	Plake	1980	p	col.	104
	Plake & Ansorge	1984	p	col.	352
	Klimko	1984	p	col.	111
	Allison	1984	p	elem.	327
	Lane et al. (b)	1987	p	col.	247
Non +	Brenner	1964	p	col.	-
	Munz & Smouse	1968	p	col.	120
	Smouse & Munz	1969	p	col.	181
	Monk & Stallings	1970	p	col.	2,000
	Munz & Jacobs	1971	p	col.	133
	Klosner & Gellman	1973	p	col.	54
	Plake, Thompson et al.	1980	p	col.	97
	Plake, Melican et al.	1983	p	col.	167
Sig	Mollenkopf	1950	t	sec.	382
	MacNicol	1956	t	sec.	1,500
	Sax & Carr	1962	t	col.	335
	Sax & Cromack	1966	t	col.	467
	Flaugher et al.	1968	t	sec.	5,000
	Sirotnik & Wellington	1974	t	sec.	2,463
	Feldt & Forsyth	1974	t	sec.	530
	Hambleton & Traub	1974	t	sec.	106
	Towle & Merrill	1975	t	col.	82
	Whitely & Dawis	1976	t	sec.	1,568
	Yen	1980	t	elem.	2,400
	Kleinke	1980	t	elem.	484
	Plake, Ansorge et al.	1982	t	col.	170
	Kingston & Dorans	1984	t	col.	4,000
	Hodson	1984	t	sec.	157
	Lane et al. (a)	1987	p	col.	155

Note non = not significant; non + = additional relevant effects; sig = significant; p = power; t = timed; elem. = elementary; sec. = secondary; col. = college; - = not given

The results of Table 1 show sixteen studies reporting significant effects related to item order and twenty-one reporting non-significant effects of item order. As Leary and Dorans (1985) concluded, those studies that indicated that speediness was a factor, due to students not finishing the test within the given time limits, were studies that reported significant order effects. Item order seems to be a factor whenever speed is a factor. One suggested reason is that students with the easiest questions first will get more questions correct before they run out of time. This conclusion is supported by the fact that, with one exception, any study which had no effect from item order was a power test, and any study which had a significant item order effect had students who did not finish.

Another explanation for this coincidence is that studies that attempted to increase their statistical power to find significant effects or studies that were using latent trait statistics needed to have larger numbers. With larger numbers of students, the need for time limits increased as well as the likelihood that some students would not finish the exam. In addition, attempts were made to improve the generalizability of the study, so samples with a wider range of abilities than the typical first year college student sample were used. This resulted in samples from the public school system and therefore samples with a wider variation in performance speed. Time limits became a necessary and typical factor of studies that were powerful enough to find an item order effect.

Table 1 shows that the sample type and the sample size seem to be a factors, but they are factors related to time. A study with over 300 students will probably have a significant item order effect or context effect, but it will probably also have time limits on such a large number of subjects. Eleven out of fifteen studies show such a pattern. As well, ten out of fifteen studies involving secondary or elementary students, reported significant item order effects, but they also reported that time limits were a factor for some students. Finally, three studies involving latent trait models indicated significant item order effects, but since latent trait models require large numbers, they also involved time limits as a factor (Yen, 1980; Whitely & Dawis, 1976; Kingston & Dorans, 1984)

The obvious relationship between the speediness of the test and the item order effects may involve other factors. Whether a test is considered a timed or a power test involves a judgmental problem related to correlation as opposed to causation. While there is a correlation between item order and time, item order may, in fact, cause some students not to finish on time rather than time causing some students to be effected by the item order. A strong correlation could also be a result of researchers who unexpectedly found an item order effect and chose to dismiss it as a side effect of their time limits; as a result, they reported students who failed to complete the test and declared that their test was speeded. On the other hand, researchers who found no effect from item order may have chosen not to mention that students did not complete the test. One related example is Hambleton and Traub (1974) who expected to find an item order effect and considered their test to be a power test. However, since some of their subjects did not finish the test, their study is considered to involve speeded tests (Leary & Dorans, 1985).

Clearly, there is a need for further research about the factors that influence item order effects. It seems that item order effects may involve a question of statistical power. So, a sample size over 400 would seem to be an appropriate size to ensure adequate power. In addition, item order effects may also involve subject variance. Therefore, sampling should be from a population with a wide range of abilities. Further, item order research may also involve controlling such confounding variables as within-subject rearrangement, gender, test content, subject motivation, and the item difficulty range. All of which have been shown by the research to be potential sources of error. Finally, the research must use a power test to avoid problems with test scores that are the results of speediness.

Chapter III

Problem

Statement of the Problem

The main problem that was addressed in this study was whether or not a test with the items arranged from hardest to easiest is more difficult than if it is arranged from easiest to hardest. Two tests were used. One test had the questions arranged in the ascending difficulty sequence, and the other test had the questions arranged in the descending difficulty sequence. However, since research would indicate that several conditions can alter the size of the item order effect, several other problems were addressed to determine the influence of these other variables.

For one, the problem of within-subject rearrangement was addressed by replicating the restrictive test booklet directions used in the Hambleton and Traub (1974) study. However, to determine if the test booklet directions are in fact a significant factor, a control group with unrestrictive directions was also used.

A third problem was to determine whether or not low ability students performed differently than high ability students with the two different item order arrangements or with the two different test booklet formats. A large and diverse sample was used to ensure an ample range of ability levels.

Finally, the problem of whether or not studies using latent trait statistics are comparable to studies using classical statistics was addressed by using both types of statistics in the analysis of the data.

Rationale

Forty years of research into item order arrangements has still not resolved the issue of whether or not the item order sequence will effect test scores. This study was intended to help clarify this controversy. In addition, several variables have been suggested by the research as factors which might influence the presence or absence of an item order effect.

One possible variable is indicated by the research of Hambleton and Traub (1974). In their study they did not allow students to rearrange the order of the items by doing only the easy ones first regardless of the item's location in the test. This prevented students from using a test taking strategy that some students seem to possess (Millman & Bishop, 1965; Tuck, 1978; Rindler, 1980; Klimko, 1984; Allison & Thomas, 1986). However, as Hambleton and Traub concluded, when any students who may have this strategy are not able to use it, then item order has an effect that is statistically significant. They may be prevented from using their strategy by the test booklet as Hambleton and Traub did, or they may be hindered in using their strategy by the presence of time limits which tend to discourage skipping back and forth between items.

If this within-subject rearrangement technique is a significant factor, then much of the previous research on item order must be more seriously questioned for its failure to control this variable. The validity of previous item order studies can be questioned since the study may be measuring an additional factor of knowledge and usage of a test-wiseness strategy.

A second variable is the possibility that when low achieving students are involved in the study, these students may lack the skill to effectively use the strategy of within-subject rearrangement and are therefore affected by the item order. This is in keeping with the more common historically held view that supports the need for arranging items from easy to hard to avoid frustrating low achieving students. Their frustration may be caused by a lack of this test taking strategy.

If some examinees can control the detrimental effects of item order, then tests should be arranged from easy to hard to obtain the best results from students who are not adept at omitting the hard questions. Also, students who are unaware of this strategy could be taught how to use the within-subject rearrangement strategy to perform better on tests.

A third variable could be that latent trait item parameters are more sensitive to item order effects than classical model statistics. Although Yen (1980) found both classical and latent trait statistics to be sensitive to the item difficulty sequence, this could be one reason why the studies involving item order and latent trait statistics reported that changes in the context were associated with significant changes in the item parameters (Whitely & Dawis, 1976; Kingston & Dorans, 1984). It needs to be determined if non-significant results with classical statistics would be significant if latent trait statistics were used instead. If latent trait statistics indicate different conclusions than classical statistics, then it would be inappropriate to compare results from a test using latent trait statistics with one that used classical statistics. In addition, if latent trait statistics are particularly sensitive to item order changes then the latent trait model assumption of local independence must certainly be questioned.

Hypotheses

Hypothesis #1: A test with the items arranged from easy to hard will be easier than the same test with the items arranged from hard to easy. This effect will be evident from the easy to hard arrangement having a higher mean, a higher average of classical item p-level difficulty indexes, and a lower average of latent trait model item parameter difficulty indexes.

Hypothesis #2: A test with test booklet directions which restrict within-subject rearrangement will be more difficult than tests without such restrictions. This effect will be evident with the restricted test having a lower mean and classical difficulty index. The latent trait model difficulty indexes will be higher.

Hypothesis #3: There will be an interaction effect between the ability of the student, the test arrangement and the test format. Students with high ability will have the lowest mean and classical difficulty index on the restricted format, hard to easy arrangement. The latent trait model difficulty index will be highest. On the other hand, low ability students will have their lowest mean, their lowest classical difficulty index, and the highest b-value parameter on any hard to easy arrangement. Figure 1 and Figure 2 display this hypothesis.

Figure 1

Low Ability Students

	Unrestricted	Restricted

Easy to Hard	Mean = high
	p-level = high
	b-value = low
		Mean = high
		p-level = high
		b-value = low

Hard to Easy	Mean = low
	p-level = low
	b-value = high
		Mean = low
		p-level = low
		b-value = high

Figure 2

High Ability Students

	Unrestricted	Restricted

Easy to Hard	Mean = high
	p-level = high
	b-value = low
		Mean = high
		p-level = high
		b-value = low

Hard to Easy	Mean = high
	p-level = high
	b-value = low
		Mean = low
		p-level = low
		b-value = high

Hypothesis #4: Classical and latent trait test statistics will indicate similar patterns of results to each other. Comparisons between classical p-level statistics and latent trait b-value statistics will have a high degree of correlation.

Chapter IV

Method

Design

The design was a post-test only control group design. Four different test booklets were used as the four different treatment groups. The booklets had either restrictive or unrestrictive directions, and they had either an easy to hard item order or a hard to easy item order. The students were placed randomly into either the control group of unrestricted, easy to hard test booklets or into one of the other treatment groups that used the other test booklets. The mathematics test itself served as the treatment and as the post-test.

The independent variables were item order, test format, and ability level. Item order was either the easy to hard arrangement or the hard to easy arrangement. Test format was either restricted or unrestricted. Ability level was rated by the teacher on a scale from "1" to "6". The 10% of students with the lowest ability in mathematics received a "1". A "2" was used for the next 15%, a "3" was used for the next 25%, a "4 was used for the next 25%, a "5" was used for the next 15%, and a "6" was used for the 10% of students with the highest mathematical ability.

The dependent variables were the test scores and the item difficulties. Item difficulty was with comparisons using classical p-level statistics, and with comparisons using latent trait b-value item parameters.

Subjects

Students in grade eight from approximately 25 different classrooms in three different secondary schools in two suburban school districts near Vancouver B.C. were used . The 590 students were from every grade eight class in each of the three schools, so a variety of socioeconomic backgrounds and ability levels were included the sample. Students with learning problems significant enough to not be enrolled in a regular classroom, or students who had an excused absence on the day the test was administered did not participate in the study.

Instrument and Tasks

A grade 8 mathematics test was developed using 40 randomly chosen items from the 150 items in the Second I. E. A. International Study of Mathematics (Robitaille & Garden, 1987). The difficulty levels had a range from $p = .13$ to $p = .89$ with the mean of difficulty levels at .492. There was no attempt to reflect the provincial curriculum in the items chosen since individual schools differed in the timing of their instruction of the curriculum.

Four mathematics test booklets were prepared using these 40 randomly chosen questions. One half of the test booklets had the questions arranged from easy to hard while the other two booklets were arranged from hard to easy. In addition, both arrangements were presented on one of two test booklet formats. One format had instructions which directed the students to only do the questions in the order that they were presented. The other format had instructions which allowed the student to look back and forth through the test booklet. Both booklet formats were printed with one question per page, and both booklet formats were colour coded to allow test takers and test supervisors the ability to be sure that the proper restricted or unrestricted instructions were being followed.

In order to make a practical assessment of student's ability, teachers were asked to complete a simple six point rating scale. This scale asked the teacher to use a normal curve distribution to rank the students' mathematics ability based on the teacher's personal judgement. A rating of "1" indicated the students with the lowest ability in math, and a rating of "6" indicated the students with the highest ability in math.

Procedure

All four test booklet types were distributed alternately to the students in each classroom. It is assumed that this systematic distribution produced an essentially random sample. Students were asked to put their name, school, and mathematics class identification information on the test booklet answer sheet. Next, students were given directions for completing the test and the restrictions for those with the restricted format booklets. Both the motivation and the cooperation of the students was sought in the directions. The students were asked to cooperate since this was part of an experiment to see if going back and forth in a test booklet would help students who had to take tests. In addition, the students were told that even though it was part of an experiment, the results might be used by the teacher in determining the students final grade. Students were then allowed the remaining fifty minutes of class time to complete the test.

Analysis

The analysis of the test scores was with a 2 x 2 x 6 Anova format using the S.P.S.S.-X program. It was assumed that the means were normally distributed, homogeneous in variance, and independent. It was also assumed that the factors of school and classroom were non-significant random factors with equal means.

The analysis of the item difficulties was with a Manova repeated measures format using the S.P.S.S.-X program. It was assumed that each item's difficulty was measured four times with each test booklet being presented to equivalent samples of the population. Further, it was assumed that the means were normally distributed, homogeneous in variance, and independent. It was also assumed that the factors of school and classroom were non-significant random factors with equal means.

The alpha level chosen for the tests of significance was .05. This level was chosen to replicate the statistical conditions established in the study by Hambleton and Traub (1974).

Chapter V

Results

A mathematics test of forty questions was administered to 590 grade eight students. The forty questions were presented in four different test booklets with the four booklets randomly given to the students. Two of the booklets had the items arranged from easiest to hardest and the other two booklets had the questions arranged from hardest to easiest. Further, each sequencing format was presented in two different test booklets. One booklet had directions which restricted subjects from rearranging the item order, and the other booklet had no such restrictions. The four tests were designed to answer the following questions:

1) Does altering the order of test items result in changes in means of the test scores?

2) Do directions on the test booklet which restrict the within-subject rearrangement of test items result in greater effects associated with the changes in item order?

3) Is their an interaction effect between a students ability, the item order and the test booklet format? Specifically, do low ability students have their lowest score whenever they have a test which begins with hard

items, and do high ability students only have low scores when the test begins with hard items and when they are not allowed to alter the item difficulty sequence?

4) Are any changes in p-level difficulty statistics similar to changes in b-value item parameter statistics?

Main Effects

The means of the two different item sequences were found to be significantly different ($p < .001$). The 297 students who had an exam with an easy to hard sequence had an average score on the 40 item test of 18.56. The 293 students who had a test which began with hard questions only had an average score of 15.90.

The means of the two different test formats were not found to be significantly different. The 296 students with the unrestricted test format booklet had an average score of 17.58. On the other hand, the 294 students with the restricted sequence did about the same with an average score of 16.89.

The four formats demonstrated a high level of reliability. Cronbach's coefficient alpha was calculated for each format. The first format with an easy to hard sequence and unrestricted directions had a reliability of .845. The second format with a hard to easy sequence and unrestricted directions had a reliability of .855. The third format with an easy to hard sequence and restricted directions had a reliability of .829. Finally the fourth format with a hard to easy sequence and restricted directions had a reliability of .835.

The means of the different teacher-rated ability levels were found to be significantly different ($p < .001$). The correlation between teacher rating and the student's score was significantly correlated ($p < .001$) with a Pearson correlation coefficient of .6340. The results are presented in Table 2.

Table 2

Test Means and Sample Sizes of Student Ability Levels

Level	Ability Label	Mean	n
1	Lowest 10%	11.27	44
2	Next 15%	11.76	90
3	Next 25%	14.96	178
4	Next 25%	18.92	112
5	Next 15%	21.63	115
6	Highest 10%	26.41	51
Total		17.37	590

Interactions

There were no significant interaction effects. Low ability students did not seem to be any more likely to receive a lower score on a test which began with hard questions than did high ability students. The effects associated with changes of item order effected all ability groups equally.

As well, test directions did no have any interaction effects. So, in addition to the fact that there was no overall main effect associated with different test directions, different ability groups were not more likely to receive higher scores if they had a different type of test direction.

The results of the 2 x 2 x 6 analysis of variance are presented in Table 3.

Table 3

Summary of Analysis of Variance of Test Scores by Ability (Ab), Item Order (Or), and Test Directions (Dir)

Source	Sum of Squares	DF	Mean Square	F	Prob.
Main	12743.090	7	1820.441	65.715	< .001
Ability	11629.609	5	2325.922	83.962	< .001
Order	694.458	1	694.458	25.069	< .001
Directions	32.034	1	32.034	1.156	.283
Ab x Or	255.123	5	51.025	1.842	.103
Ab x Dir	80.120	5	16.024	.578	.717
Or x Dir	6.453	1	6.453	.233	.630
Ab x Or x Dir	179.829	5	35.966	1.298	.263
Explained	13249.491	23	576.065	20.795	< .001
Residual	15679.289	566	27.702		
Total	28928.780	589	49.155		

Item Difficulties

Due to the close mathematical relationship between the mean and p-level, it is not surprising that a significant difference was also found between the mean of the p-levels of each test. A multivariate analysis of variance for repeated measures was used to determine if there was a significant difference between the item difficulties on each test. Each item had p-level values from four tests. The samples were assumed to be equivalent with the variance primarily a result of the explained variance between the tests or ability levels as reported in the previous Anova results in Table 3.

As expected, a significant difference was found between the different item difficulties of each test as shown in Table 4.

Table 4

Summary of Multivariate Analysis of Variance of Test Item P-levels

Source	Sum of Squares	DF	Mean Square	F	Prob.
P-levels	.10	3	.03	10.26	< .001
Within Cells	.18	57	.004		

The same type of analysis with rasch item parameters was used. The rasch item parameters were estimated using the Microcat Testing System (1988) with ability levels standardized. Table 5 shows that the rasch item parameters also have a similar level of significance as the classical item difficulties in Table 4.

Table 5

Summary of Multivariate Analysis of Variance of Test Item B-values

Source	Sum of Squares	DF	Mean Square	F	Prob.
B-values	3.14	3	1.05	9.09	< .001
Within Cells	6.56	57	0.12		

The similarity of results is also apparent by comparing the means of the test scores, the p-values, and the b-levels for each test as given in Table 6. The changes in the p-levels and the test means are inversely related to the changes in the b-values.

Table 6

Test Format P-level Means, B-value Means, and Score Means

Format Number	Item Order	Test Type	P-level Mean	B-value Mean	Score Mean
1	Easy - Hard	Unr.	.474	.128	18.946
3	Easy - Hard	Res.	.454	.295	18.169
2	Hard - Easy	Unr.	.405	.493	16.190
4	Hard - Easy	Res.	.390	.652	15.603

Note Unr. = unrestricted directions; Res. = restricted directions

The Pearson correlation coefficients in Table 7 also demonstrates a strong relationship between an item's p-levels and its b-values regardless of which test format is used with the item.

Table 7

Pearson Correlation Coefficients of P-levels and B-values for Easy to Hard (EH), Hard to Easy (HE), Restricted (R) and Unrestricted (U) Test Formats

B-value Test Formats	P-level Test Formats EH, U	HE, U	EH, R	HE, R
EH, U	-.9652	-.9385	-.8796	-.9125
HE, U	-.9001	-.9050	-.8558	-.8820
EH, R	-.9225	-.9350	-.8818	-.9099
HE, R	-.8141	-.8294	-.7697	-.8181

Note All correlations are significant at $p < .001$.

Finally, Table 8 further demonstrates the similarities between classical and latent trait statistics. Table 8 uses theta values calculated from b-values standardized for difficulty.

Table 8

Summary of Analysis of Variance of Theta Values by Ability (Ab), Item Order (Or), and Test Directions (Dir)

Source	Sum of Squares	DF	Mean Square	F	Prob.
Main	239.955	7	34.279	55.127	< .001
Ability	217.250	5	43.450	69.875	< .001
Order	14.997	1	14.997	24.117	< .001
Directions	.212	1	.212	.341	.559
Ab x Or	5.052	5	1.010	1.625	.151
Ab x Dir	2.614	5	5.523	.841	.521
Or x Dir	.002	1	.002	.004	.951
Ab x Or x Dir	4.160	5	.832	1.338	.247
Explained	251.773	23	10.947	17.604	< .001
Residual	351.954	566	.622		
Total	603.727	589	1.025		

Chapter VI

Summary and Conclusions

Purpose of The Study

This study was an attempt to determine if the sequence of test items has an effect on the performance of students. Further, this study tried to determine if some examinees were able to mitigate any such effects by personally rearranging the item order as presented by the researcher in the test booklet. Whether or not the item order can have an effect on test scores has been an area of research for forty years. Recently, with the increased usage of latent trait statistics, the issue of context effects and local independence has become more of a concern. As a result this study also examined the effect of item order on latent trait statistics.

Four different test booklets were used and given randomly to 590 grade eight math students. Two booklets had the items arranged in sequence from easy to hard questions, and the other two booklets had the items arranged in sequence from hard to easy. Both types of sequences had one booklet which allowed the students to rearrange the order of item presentation by skipping back and forth between questions, and there was one booklet of each sequence type that did not allow such within-subject rearrangement.

The results of the study supported several of the hypotheses about the effects of item order and the factors associated with it.

Sequence

The sequence of the test items can affect the performance of the students. Students who took the test with the items arranged from easy to hard had a significantly higher mean than the students who had the hard to easy arrangement ($p < .001$). The students who took the easy to hard tests had a mean score of 18.6 as compared to the students who took the hard to easy test and had a mean score of 15.9. The mean of the students who took the hard to easy exam had scores 7% lower than students who took the other test.

Although there was no actual measure of the students theoretical frustration and discouragement while they were taking the test, the common concern about beginning a test with too many hard questions may have some justification. It is clear that a hard to easy arrangement can result in lower scores for the students. As a result, caution should be exercised when developing two forms of the same test. It is possible to create formats with significantly different test characteristics even though the items are identical.

Directions

The difference between the two types of directions was not significant. However, there was a trend toward significance since mean of a test with restricted directions was lower than a comparable test with unrestricted directions. In comparing the overall test scores, those students who were allowed to do the questions in any order and to go back and forth between the questions had a mean of 17.6. The students who could not rearrange the order of the exam had a mean score that was not significantly lower at 16.9.

The effort to prevent students from rearranging the item order does not seem to greatly improve the likelihood of finding a significant item order effect. The evidence for the widespread and effective use of this test-wiseness strategy was not clearly demonstrated. Therefore, this study supports the conclusions of Allison and Thomas (1986) that there is not enough evidence to doubt that the majority of item order studies would have had different findings if this factor had been controlled. However, in light of the trends indicated in the data, this factor may be a problematical variable in some situations as Hambleton and Traub (1974) suggested.

Ability

The main effect associated with ability levels was significant. The fact that teachers are able to predict how well their students will do on a mathematics test is not an unexpected finding. As a result, no conclusions will be drawn from this finding.

Interactions

A more significant fact is that there were no significant interactions between any of the factors including ability. Students who were considered to have limited mathematical ability were effected by the sequence and the directions to the same degree as the students who were considered to have high mathematical ability. This calls into question one of the justifications for the concern over item order. The easy to hard order does not appear to help the low ability student any more than it helps the high ability student. So while a concern for the feelings of low achieving students is admirable, there is no special justification for arranging the test items from easy to hard based on the results of this study.

Another lack of significant difference involves the interaction between the test directions and the ability levels. The lack of any significant differences contradicts the findings of Rindler (1980) that all ability levels possess the test-wiseness strategy of skipping questions, but they use it with varying degrees of success as demonstrated by the complex interactions found in Rindler's study. However, the results of these studies do not preclude the value of teaching test-wiseness strategies to possibly help students to learn to use the item rearrangement strategy effectively.

Latent Trait

Studies which have used latent trait statistics in their analysis of item order or context effects have all found significant differences in their difficulty parameters. Only one study (Yen, 1980) used both latent trait statistics and classical statistics, and that study found the same significant effect of item order using both types of statistics.

Both classical based difficulty levels and latent trait difficulty levels were found to be significantly correlated in this study, and both demonstrated the same significant effect associated with changes to the item order. It can be concluded that the studies which used latent trait statistics and found a significant effect from changes in context would probably have found similar results had they used classical based statistics. Further, it is also possible that if some previous classical based studies had used larger samples, then their results may have been similar to the latent trait based studies.

The assumption of local independence can not be supported by the results of this study. Latent trait difficulty parameters were affected by the difficulty parameters of preceding items. Caution must be exercised when comparing tests by using the latent trait statistics since a test which begins with harder questions cannot be assumed to be the parallel to a test which begins with easier questions.

Limitations

The conclusions of this study must be limited to comparisons between a test arranged from easy to hard and one arranged from hard to easy. Other formats such as random or spiral were not included. It is open to conjecture and future research if a random arrangement that began with primarily hard questions would have significantly lower scores than a random arrangement that began with primarily easy questions.

A second area of limitation involves the content of the tests. This study confirms many of the results found with studies that used quantitative type tests (Hambleton & Traub, 1974; Feldt & Forsyth, 1974; Towle & Merrill, 1975; Yen, 1980; Plake, Ansorge et al. 1982; Kingston & Dorans, 1985). However, this study may only generalize to mathematics tests. One possible reason is that the difficulty of an item may be highly subjective. The statistical difficulty level may only be an estimate of the actual difficulty that is perceived by the individual encountering the item. In the case of a mathematics question, the statistical difficulty level may be a good predictor of how difficult each individual perceives the question to be. On the other hand, a science question may be statistically very difficult because most subjects answer it incorrectly, but it is perceived as a very easy question by the respondents due to the effectiveness of the distractors. As a result, a series of statistically difficult science questions may not result in the same effects as a series of difficult mathematics questions.

A third limitation is the result of the definition of mathematical ability used in this study. The ability levels used in this study were based on a teacher rating system and would be strongly influenced by classroom behaviour, student personality, and the errors of teacher judgement. Even though the teacher rating scale had a significant .6340 correlation with the mathematics test scores ($p < .001$), the results of this study may differ from a study which uses a measure of student ability with a more valid criterion of mathematical ability.

The issue of difficulty involves another limitation. Items used in this study do not have difficulty levels which are identical to other studies. The conclusions of this study are based on a series of items whose pre-tested difficulty levels had an average p value of .49 with a standard deviation around that mean of .21. The range of p levels was from .13 to .89. Unfortunately it is not clear if the results of this study compare with the results of other studies since, as Hambleton and Traub (1974) pointed out, many studies do not give information about their item difficulty. The degree to which the mean and variation of difficulty levels influence item order effects is a subject for future research.

A definite limitation of this study is that the results only generalize to children enrolled in the intermediate or secondary school programs of Canadian public schools with a wide diversity of student ability levels. This study may not be applicable to a college setting where some previous research has indicated, changes in item order do not result in significant changes in the scores of college students. However, the results of this study do call into question some of the generalizations of previous research which used college students to conclude that item order does not have an effect.

Conclusions about latent trait statistics are restricted by the small sample size. Although there were 590 students in total, there were only about 147 students taking each test. The latent trait parameters for each test format were established just with the students who were given that particular test booklet.

Generalizability may also be limited by the possible interaction of selection process and the tests used in this study. This limitation is outlined as a possible weakness of post-test only control group designs by Campbell and Stanley (1963). While the three schools involved in the study are hopefully representative of the typical junior secondary school, it is possible that the three schools involved were atypical. For one, they were the only three out of the five schools asked which agreed to participate in the study. The two schools which declined to participate did so because they felt that the district's labour difficulties had already significantly shortened their instructional time. It should be noted that two of the participating schools did not feel that the shortened instructional time was a hindrance to their participation. Therefore, since the factors involved with participation seem to be unrelated to the factors under study, it can be concluded that there probably was not any interaction between the selection process and the tests used in the study.

Campbell and Stanley also point out that the design is limited by the possible effect of reactive elements. To a certain extent students were affected by the unusual nature of the testing procedure. For one, students who received the unrestricted test booklets may have reacted more positively to the testing situation since some expressed pleasure at having received the unrestricted test booklet. Unfortunately, the testing situation may have also limited the possible effect related to directions since some students may not have fully cooperated with the directions to not rearrange the item order. While the majority of students were cooperative, a few students in each school seemed to be uncooperative since they assumed that it was really just some type of an experiment. Those students passively resisted teacher attempts to have them follow the directions and do their best. This limits the accuracy of any conclusions about the effect of the test directions. On the other hand, the item order effects were less likely to be influenced by such a factor since the students were not told that the tests were also prepared with different item sequences. In fact, some teachers said that the hard to easy sequence was more likely to cause uncooperative behaviour rather than the uncooperative behaviour limiting the item order effects.

Nevertheless, despite the limitations of this study, the conclusions of this study should not be in any way limited to speeded tests. Every attempt was made to have this test be a power test within the limitations of testing 590 students in the public school system. Most students easily finished the test within the time allotted. The teachers who administered the test stated that the time limits were ample and generous. Nonetheless, there are students who did not complete the test, and there are items that were omitted at the end of the test and technically classified as "not reached".

However, whether or not a test is a power test because all students completed the test and only omitted the most difficult or whether a test is a speeded test because questions were not reached by some students is a difficult distinction. It is not realistically possible for tests of the type used in this study to not have a small percentage of not reached questions. For example, one student after trying the first question of the hard to easy sequence booklet threw his test across the room and refused to complete the rest of the test. As a result thirty-nine questions on his test can erroneously be scored as not reached rather than omitted. As another example, students were observed to use a test taking strategy of doing the questions at the beginning and the end first while questions in the middle were left until last. If these students did not have enough time to complete the test their not reached questions would then be scored as omitted. The difference between not reached and omitted questions is also not clear since students who took the easy to hard format had 1.7% of their questions not reached, but the students with the hard to easy format had .9% of their questions not reached. Students with hard questions at the end of the test omitted more questions at the end of the test which increases the number of technically not reached questions. While omitting questions on a power test is very different from not reaching questions on a speeded test, it is not accurate to make a statistical distinction between the two under the conditions of this study. For all intents and purposes the tests used in this study were power tests with some students choosing to omit questions.

Implications

Caution should still be expressed by writers in the measurement field about item sequencing. Under certain circumstances, it is possible for the context of the items to influence the statistics of the items. Care must be taken in the development of parallel forms of a test to prevent significant differences in scores as a result of differences in the item sequencing.

Future Research

Many areas remain as subjects for future research. For one, the differentiation between the six different ability groups could be the basis of further research. Students could be classified into different groups based on a pre-test that measures intellectual ability, mathematical achievement, or both. The scores from those tests could be used to identify more or fewer groups as needed for the analysis of any interaction between ability level and other factors.

A second area of research is to determine if there is a significant difference between subjective item difficulty and statistical item difficulty. Students could rate the subjective difficulty of tests, and those ratings could be compared with statistical ratings to determine the correlation. Different subject areas could be used to compare the correlation between content areas to determine if the types of questions with the highest subjective and statistical correlation are the content areas with the greatest item order effects.

Variations in item and test difficulty could also be examined. For one, the number of difficult items at the beginning of a test could be varied to determine the maximum number of difficult items that could be tolerated by students without resulting in lower test scores. Variations of the mean and range of item difficulty would also give evidence to the sensitivity of students to item difficulty.

References

Ahmann, J. S., & Glock, M. D. (1963). Evaluating pupil growth (2nd ed.). Boston: Allyn & Bacon.

Allison, D. E. (1984). The effect of item-difficulty sequence, intelligence, and sex on test performance, reliability, and item difficulty and discrimination. Measurement and Evaluation in Guidance, 16, 211-217.

Allison, D. E., & Thomas, D. C. (1986). Item-difficulty sequence in achievement examinations: Examinees' preferences and test-taking strategies. Psychological Reports, 59, 867-870.

Berger, V. F., Munz, D. C., Smouse, A. D., & Angelino, H. (1969). The effects of item difficulty sequencing and anxiety reaction type on aptitude test performance. Journal of Psychology, 71, 253-258.

Brenner, M. H. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. Journal of Applied Psychology, 48, 98-100.

Campbell, D. T., & Stanley, J. C. (1963) Experimental and quasi-experimental designs for research. Boston: Houghton Mifflin.

Cronbach, L. J. (1946). Response sets and test validity. Educational and Psychological Measurement, 6, 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. Educational and Psychological Measurement, 10, 3-31.

Feldt, L. S., & Forsyth, R. A. (1974). An examination of the context effect in item sampling. Journal of Educational Measurement, 11, 73-82.

Flaugher, R. L., Melton, R. S., & Myers, C. T. (1968). Item rearrangement under typical test conditions. Educational and Psychological Measurement, 28, 813-824.

French, J. L., & Greer, D. (1964). Effect of test-item arrangement on physiological and psychological behavior in primary-school children. Journal of Educational Measurement, 1, 151-153.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. Journal of Experimental Education, 43, 40-46.

Hodson, D. (1984). The effect of changes in item sequence on student performance in a multiple-choice chemistry test. Journal of Research in Science Teaching, 21, 489-495.

Hopkins, C. D., & Antes, R. L. (1985). Classroom measurement & evaluation (2nd.). Itasca, IL: Peacock.

Huck, S. W., & Bowers, N. D. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. Journal of Educational Measurement, 9, 105-111.

Kestenbaum, J. M., & Weiner, B. (1970). Achievement performance related to achievement motivation and test anxiety. Journal of Consulting and Clinical Psychology, 34, 343-344.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 147-154.

Kleinke, D. J. (1980). Item order, response location and examinee sex and handedness and performance on a multiple-choice test. The Journal of Educational Research, 73, 225-229.

Klimko, I. P. (1984). Item arrangement, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. Journal of Experimental Education, 52, 214-219.

Klosner, N. C., & Gellman, E. K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. Educational and Psychological Measurement, 33, 413-418.

Lane, D. S., Bull, K. S., Kundert, D. K., & Newman, D. L. (1987). The effects of knowledge of item arrangement, gender, and statistical and cognitive item difficulty on test performance. Educational and Psychological Measurement, 47, 865-879.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. Review of Educational Research, 55, 387-413.

MacNicol, K. (1970). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. Journal of Educational Measurement, 7, 113-118.

Microcat testing system (3rd Ed.). (1988). St. Paul, MN: Assessment Systems Corporation.

Millman, J., & Bishop, C. H. (1965). An analysis of test-wiseness. Educational and Psychological Measurement, 25, 707-726.

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika, 15, 291-315.

Monk, J. J., & Stallings, W. M. (1970). Effects of item order on test scores. The Journal of Educational Research, 63, 463-465.

Munz, D. C., & Jacobs, P. D. (1971). An evaluation of perceived item-difficulty sequencing in academic testing, British Journal of Educational Psychology, 41, 195-205.

Munz, D. C., & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement-anxiety reaction on academic performance. Journal of Educational Psychology, 59, 370-374.

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. Journal of Experimental Education, 49, 56-58.

Plake, B. S., & Ansorge, C. J. (1984). Effects of item arrangement, sex of the subject, and test anxiety on cognitive and self-perception scores in a nonquantitative content area. Educational and Psychological Measurement, 44, 423-430.

Plake, B. S., Ansorge, C. J., Parker, C. S., & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test performance. Journal of Educational Measurement, 19, 49-57.

Plake, B. S., Melican, G. J., Carter, L., & Shaughnessy, L. C. (1983). Differential performance of males and females on easy to hard item arrangements: Influence of feedback at the item level. Educational and Psychological Measurement, 43, 1067-1075.

Plake, B. S., Thompson, P. A., & Lowry, S. (1980). Effect of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. Journal of Experimental Education, 49, 214-219.

Rindler, S. E. (1980). The effects of skipping over more difficult items on time-limited tests: Implications for test validity. Educational and Psychological Measurement, 40, 989-998.

Robitaille, D. F., & Garden, R. A. (Eds.) (1987). The second international mathematics study: Vol. 2. Context and outcomes of school mathematics. Albany, NY: International Association for the Evaluation of Educational Achievement.

Ruch, G. M. (1929). The objective or new-type examination. Chicago: Scott Foresman.

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. Educational and Psychological Measurement, 22, 371-376.

Sax, G., & Cromack, T. R. (1966). The effects of various forms of item arrangements on test performance. Journal of Educational Measurement, 3, 309-311.

Sirotnik, K., & Wellington, R. (1974). Scrambling content in achievement testing: An application of multiple matrix sampling in experimental design. Journal of Educational Measurement, 11, 179-188.

Smouse, A. D., & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. Journal of Psychology, 68, 181-184.

Smouse, A. D., & Munz, D. C. (1969). Item difficulty sequencing and response style: A follow-up analysis. Educational and Psychological Measurement, 29, 469-472.

SPSS-X user's guide. (1983). New York, NY: McGraw-Hill.

Towle, N. J., & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 12, 241-249.

Tuck, P. J. (1978). Examinees' control of item difficulty sequence. Psychological Reports, 42, 1109-1110.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

APPENDIX I

TEACHER INSTRUCTIONS

MATHEMATICS 8 EXAMINATION

THESIS PROJECT OF M. J. SCALES

I. General Directions

II. Thesis Project Background

III. Detailed Directions (optional)

  a. Start Explanations
  b. Booklets
  c. Answer Sheets
  d. Identification Number
  e. Name Section
  f. Gender
  g. Grade
  h. Birth Date
  i. Answer Sheet Usage
  j. Start Examination
  k. End Examination
  l. Collect Testing Materials
  m. Math Ability Rating
  n. Return Testing Materials

IV. Appendix

  a. Student Identification Sample

I. GENERAL DIRECTIONS

Materials Required by the Examiner

A. A copy of these instructions.

B. A class set of mixed test booklets, complete with an answer sheet and some scrap paper to give one booklet to each student.

Test Format 1 (orange)
Test Format 2 (yellow)
Test Format 3 (blue)
Test Format 4 (green)

C. A supply of sharpened soft-lead pencils.

D. An extra supply of booklets, complete with answer sheets and scratch paper.

1. A class period of one hour should be sufficient to explain (15 min.) and administer (45 min.) the test.

2. Explain to the students that today they will be taking a test as part of a study to determine if skipping back and forth between test questions will help students to do better on tests. Some students will receive tests which allow them to skip back and forth. Other students will receive booklets which require that they do not skip ahead but must do the questions in the same order as in their booklet.

   Do not discuss the order of the items in the test or the test taking strategy of skipping the hard questions to do the easy questions first.

3. Be sure all students have a pencil (No. 2 or HB).

4. Caution students not to open their test booklet until till they are told to do so.

5. Distribute one test booklet with appropriate answer sheet and scrap paper to each student. Alternate evenly between the four different types of colour coded test booklets.

6. Have the students carefully remove the answer sheet and the piece of scrap paper from the test booklet. Have them check to see if box A in the

"Identification No." section of their answer sheet has been marked with the number that corresponds to the test format number on the front cover.

7. Have the students complete the name, sex, birth date, and grade sections of their answer sheets. If the class is unsure how to complete these sections, read the appropriate "Detailed Directions" of this booklet to the class or use the sample sheet in the appendix as a guide.

8. To get the students to be realistically motivated, please tell them that these test results may be used to calculate their final marks.

9. Encourage the students to read the remaining directions from number 5 to the end of the page. If necessary, read them to the whole class.

10. Remind those students with "Special Instructions" that they may not skip ahead to new questions or go back to old ones. They must do the questions in the same order as they are presented in the test. If they can't answer a question, they may omit that question and go on to the next one. Nonetheless, they should at least try their best to answer every question on the test.

11. When you are sure that all students understand the directions, begin the test (45 minutes).

12. During the testing period, students might ask for help. Encourage them to read and respond to each item to the best of their abilities. Do <u>NOT</u> change the wording of any items, or explain specific terms, or discuss the ordering of the questions. Treat this testing situation as normal and as serious as any other examination.

13. After 45 minutes, or sooner if all students are finished, end the test. Collect the test booklets in colour coded groups. Collect the answer sheets and check to see that all of the identification sections of the answer sheets have been completed correctly.

14. On a class list, rate each students ability to do mathematics. Using a six point scale, record a number from 1 to 6 that represents your best estimate of each students mathematical abilities. This rating should be somewhat independent from

overall intelligence or classroom behaviour. Use a "1" for those students with the lowest 10% of mathematical ability, a "2" for the next 15% of students with higher mathematical ability, a "3" for the next 25%, a "4" for the next 25%, a "5" for the next 15%, and a "6" for the 10% of students with the highest mathematical ability.

II. THESIS PROJECT BACKGROUND

Ever since multiple choice tests first came out in the early 1920s, most textbook authors have suggested that these tests should be arranged with the easiest questions at the beginning and the hardest questions at the end. One justification for such an arrangement is to help low ability students avoid early frustration with the test. However, much of the research over the last 40 years has generally found that the item order does not make much of a difference to the final results of the tests.

One purpose of this study is to examine the discrepancy between what research has statistically found and what teachers and textbook writers have intuitively found. Since most of the past research has used college students, this study will involve a younger and more diverse sample of high school students. It is the hypothesis of this study that students who are in fact affected by the item order are more likely to be found in a typical public school rather than in a college classroom.

If students of low ability are, in fact, easily discouraged by starting tests with the more difficult questions, then another purpose of this study is to examine one of the skills that high ability students may use to avoid that discouragement. One possible skill of the more able students is the tactic of omitting the hard questions until they have finished the easy questions. This may be a skill that the low ability students are either unaware of or just fail to use.

A third purpose of this study is a more esoteric one which involves examining the results of this test using two types of test statistics. The studies which have found no differences as a result of item order have used classical statistics to examine their data. However, recent studies which have found some effects of item order, have used the more modern latent trait statistics. This study would compare the results obtained from each type of statistical method.

A final reason, of course, is to complete the requirements to obtain a Master of Arts degree in the Faculty of Education at the University of British Columbia in the department of Educational Psychology and Special Education with a specialization in measurement, evaluation and research methodology.

III. DETAILED DIRECTIONS (optional)

All directions that you can read to the students are indented so that they stand out. You may read them exactly as they are written, using a natural tone and manner. If necessary, you may supplement the directions with your own explanations, but do not give help on specific test questions.

Try to maintain a natural classroom atmosphere during the test administration. Encourage students to do their best, and advise them not to spend too much time on any one question. Check periodically to make sure that students are recording their answers properly, are following instructions, and are working to the end of the test, or as far as they can.

The scoring machine used to process the answer sheets is capable of almost 100% accuracy if the answer sheets are marked correctly and kept in good condition. Remind the students to handle the sheets with care; to record their answer with heavy, dark marks; and to avoid making stray marks on their answer sheets. Answer sheets should never be folded, clipped, or torn.

a. Start Explanations

(Have all desks cleared, and see that each student has a soft-lead pencil (No. 2 or HB), and an eraser. Say:)

> "You are going to take a special math test today. Don't open your test book or make any marks on it until I tell you what to do."

b. Booklets

(Give one test booklet to each student. As you hand out the tests, alternate between the four different types of test booklets to evenly distribute the four types among your students. Place the booklet with the front cover up. Also, make sure each student has an answer sheet and a piece of scratch paper in his booklet. When the booklets have been distributed, say:)

> "Please don't open your booklets until you are told to do so by me."
>
> "Four different booklets have been distributed as part of a special experiment to see if students

can do better on tests if they are allowed to skip around between test questions. Those of you with the yellow or orange tests are allowed to go back and forth in the test booklet and do the questions in whatever order you wish. Those of you with blue or green test booklets are requested not to skip ahead to a new question and then go back to an old one. You must answer the questions in the order they appear on the test. Students with the blue or green test booklets will also find that they have some special instructions on their test booklets and on their test questions to remind them of these special rules."

(Pause and answer questions. Do not discuss the special order of the items. Try to maintain your normal testing routine. Try to obtain the cooperation and motivation of the students.)

c. Answer Sheets

(Say:)

"Carefully remove the answer sheet from the inside front cover of your test booklet. Your answer sheet is going to be scored by machine, so be careful with it. Keep it as clean as possible, and don't bend it or fold the corners."

d. Identification Number

"I may be using the results of this exam to help me determine your final grade at the end of the year. It is therefore important that you do your best. It is also important that I know which test you took. I want everyone to find the box marked 'Identification No.' on their answer sheet and the test format number on the front of their test booklets."

(Show the location of the 'Identification No.' section on the back of the answer sheet and the test format number on the front of the test booklet.)

"In the box labelled 'A' in the identification number section, make sure that the number of the test format of your test is in that box. The orange test is format one. The yellow test is format two. The blue test is format three. The green test is format four.

(Pause)

e. Name Section

"Find the spaces for your name."

(Demonstrate)

"First in the boxes print as many letters of your last name as you can. Use one box for each letter. Then, leave one box as a space. Next, print as many letters of your first name as you can. Then, leave another space. Finally, print your middle initial. If you cannot fit your full name in the space provided, try to print at least most of your last name, a space, your first initial, a space, and finally your middle initial."

(Pause)

"Now in the column below each box, fill in the circle that has the same letter or space as the letter or space in the box above it. Be sure that you mark only one circle in each column. Fill in the blank circle at the top of every column in which you have left a space. Be sure to make heavy, shiny marks that cover the whole circle. If you make a mistake, erase your mark completely. If you have any questions, raise your hand."

(Pause until all students have finished filling in the name section. Then say:)

"You should have 19 circles filled in under the name boxes. Count and make sure."

(Pause)

f. Gender

(After students have checked the name section, say:)

"Now look at the box below the columns you filled in for your name."

(Demonstrate)

"Fill in the circle next to 'Male' if you are a male or next to 'Female' if you are female."

(Pause)

g. Grade

"Now look at the box and circled numbers to the right labelled 'Grade or Education'. Just fill in the circle with an 8 since this is a grade 8 course.

(Pause)

h. Birth Date

"Now look at the columns underneath the box labelled 'Birth Date'."

(Demonstrate)

"Fill in the circle next to the month in which you were born."

(Pause)

"Fill in the boxes labelled 'Day' with two numbers for the day of your birth. For example, if you were born on the seventh of the month, you would write zero seven."

(Pause)

"Fill in the circles in the columns underneath the boxes labelled day to show the number in the box above the column. Be sure to only fill in one circle in each column."

(Pause)

"Now fill in the boxes labelled year with the two numbers for the year you were born in, and fill in the circle under each box to indicate the number in the box."

(Pause)

"Now check to make sure that you have correctly filled in all the required information."

(Pause)

i. Answer Sheet Usage

> "Before I tell you to open your test booklet and start, I am going to tell you how to properly mark your answer sheet. Listen carefully so that you will know how to mark your answers. You are to mark all your answers on your answer sheet. Don't make any stray marks on it and do not write in your booklet. You should already have some scratch paper for any figuring that you might have to do. For each question, choose the best answer. Then, on your answer sheet, find the number for the question, and mark the space for your answer. Be sure to mark the space for your answer. Be sure to mark only one answer space for each question. Make your mark heavy and shiny, and see that it completely fills the answer space. If you change your mind after you've marked an answer, erase the wrong mark completely; then make your new mark."

(On the chalkboard, show students how to fill in an answer space. Answer all questions.)

> "You will have 45 minutes to work on this test. If you have any trouble reading a question, raise your hand and I will help you. Of course, you may not use a calculator. If you're not sure about the answer to a question, do the best you can, but don't spend too much time on any one question. You may omit a question if you are sure that you cannot answer it."
>
> "Make sure that you have turned your answer sheet over to side one, so the name section is face down, so the side with the picture of the pencil is face up, and so the answer space for question one is face up."

(Demonstrate and check to make sure everyone is starting on side 1.)

j. Start Examination

(When you feel that everyone understands the directions, say:)

> "You may start working now."

(Record the starting and ending times on the chalkboard. While students are working, walk around the room to make sure that the students are following directions. Try your best to make sure that students do not change the order of the exam questions if they are in the blue or green booklets with the special instructions. If you see that a student is having difficulty reading a problem, you may help the student read the problem; however, do not give help in answering any of the questions.)

k. End Examination

(After 45 minutes, or sooner if all students have finished, say:)

> "Stop! Put your pencil down now, and close your booklet so that the front cover is up. I will collect your test booklets and answer sheets."

l. Collect Testing Materials

(Collect the test booklets into the four colour coded groups. Collect the answer sheets and check to make sure that the student identification sections have been correctly filled out. Collect the scratch paper and dispose of it. Collect any of the extra pencils loaned to the students.)

m. Math Ability Rating

On a class list, rate each students ability to do mathematics. Using a six point scale, record a number from 1 to 6 that represents your best estimate of their mathematical abilities. This rating should be somewhat independent of overall intelligence and general classroom behaviour. Use a "1" for those students with the lowest 10% of mathematical ability, a "2" for the next 15% of students with higher mathematical ability, a "3" for the next 25%, a "4" for the next 25%, a "5" for the next 15%, and a "6" for the 10% of students with the highest mathematical ability.

This information will be kept strictly confidential and used only to identify which students, from a teacher's point of view, may be either frustrated by the arrangement of the test questions or hindered by the directions of the test booklets.

Please keep answer sheets grouped in classes with their class list. The results for each of your students will be sent to you at your request.

n. Return Testing Materials

Please return testing material, the tests, the answer sheets, the pencils, and the rating lists to Michael Scales.

The test will be scored and analyzed by Michael Scales, graduate student at the University of British Columbia, and teacher at Aldergrove Secondary. The results will be kept strictly confidential with the results of individual students only being sent to that student's classroom teacher if so requested. It is not the intention of this study to make comparisons between individual classes, schools, teachers, or students.

Thank you for your cooperation and your efforts.

APPENDIX II

MATHEMATICS 8 EXAMINATION

TEST FORMAT 1

INSTRUCTIONS

1. Do NOT open the test booklet until you are told to do so. You will have 45 minutes to complete this test.
2. Carefully remove the answer sheet from inside the front cover and make sure there is a 1 marked in box A of the Identification No. section of the answer sheet.
3. Be sure you have a pencil, an eraser, and some scratch paper.
4. Fill in your answer sheet with your name, sex, grade, and birth date.
5. Do NOT use a calculator or a protractor.
6. For each question, select the best answer. Mark your choice on the answer sheet by filling in the bubble under the correct letter. Make sure the question number is the same as the question number in the test booklet.
7. Do not spend too long on any one question. Try your best pick a good answer to every question.
8. If you make a mistake, completely erase your first choice and fill in the bubble of your new choice.
9. Do NOT write in the test booklet. Mark only your answer sheet. If your booklet already has any inappropriate marks, ask for a clean booklet.

FINISHED?

Close your test booklet.

Make sure you have filled in your answer sheet with your name, sex, grade, and birth date.

Make sure that the Identification No. Box A has a 1 in it.

Turn in your test booklet and answer sheet.

THANK YOU

MATHEMATICS 8 EXAMINATION

TEST FORMAT 2

INSTRUCTIONS

1. Do NOT open the test booklet until you are told to do so. You will have 45 minutes to complete this test.
2. Carefully remove the answer sheet from inside the front cover and make sure there is a 2 marked in box A of the Identification No. section of the answer sheet.
3. Be sure you have a pencil, an eraser, and some scratch paper.
4. Fill in your answer sheet with your name, sex, grade, and birth date.
5. Do NOT use a calculator or a protractor.
6. For each question, select the best answer. Mark your choice on the answer sheet by filling in the bubble under the correct letter. Make sure the question number is the same as the question number in the test booklet.
7. Do not spend too long on any one question. Try your best pick a good answer to every question.
8. If you make a mistake, completely erase your first choice and fill in the bubble of your new choice.
9. Do NOT write in the test booklet. Mark only your answer sheet. If your booklet already has any inappropriate marks, ask for a clean booklet.

FINISHED?

Close your test booklet.

Make sure you have filled in your answer sheet with your name, sex, grade, and birth date.

Make sure that the Identification No. Box A has a 2 in it.

Turn in your test booklet and answer sheet.

THANK YOU

MATHEMATICS 8 EXAMINATION

TEST FORMAT 3

INSTRUCTIONS

1. Do NOT open the test booklet until you are told to do so. You will have 45 minutes to complete this test.
2. Carefully remove the answer sheet from inside the front cover and make sure there is a 3 marked in box A of the Identification No. section of the answer sheet.
3. Be sure you have a pencil, an eraser, and some scratch paper.
4. Fill in your answer sheet with your name, sex, grade, and birth date.
5. Do NOT use a calculator or a protractor.
6. For each question, select the best answer. Mark your choice on the answer sheet by filling in the bubble under the correct letter. Make sure the question number is the same as the question number in the test booklet.
7. Do not spend too long on any one question. Try your best pick a good answer to every question.
8. If you make a mistake, completely erase your first choice and fill in the bubble of your new choice.
9. Do NOT write in the test booklet. Mark only your answer sheet. If your booklet already has any inappropriate marks, ask for a clean booklet.

SPECIAL INSTRUCTIONS

1. You must begin with question 1. When you have chosen the best answer and marked your answer sheet, then you must go to question 2. When you have finished question 2, then you must go on to question 3, then question 4, then question 5, and so on to the end of the test.
2. Try each question once and only once. If you can't answer a question, go on to the next one.

<u>Do NOT skip ahead to new questions or go back to old ones.</u>
<u>Try to answer each question in its proper order.</u>

END OF TEST

Close your test booklet.

Do not go back to any of the questions.

Make sure you have filled in your answer sheet with your name, sex, grade, and birth date.

Make sure that the Identification No. Box A has a 3 in it.

Turn in your test booklet and answer sheet.

THANK YOU

MATHEMATICS 8 EXAMINATION (141)

TEST FORMAT 4

INSTRUCTIONS

1. Do NOT open the test booklet until you are told to do so. You will have 45 minutes to complete this test.
2. Carefully remove the answer sheet from inside the front cover and make sure there is a 4 marked in box A of the Identification No. section of the answer sheet.
3. Be sure you have a pencil, an eraser, and some scratch paper.
4. Fill in your answer sheet with your name, sex, grade, and birth date.
5. Do NOT use a calculator or a protractor.
6. For each question, select the best answer. Mark your choice on the answer sheet by filling in the bubble under the correct letter. Make sure the question number is the same as the question number in the test booklet.
7. Do not spend too long on any one question. Try your best pick a good answer to every question.
8. If you make a mistake, completely erase your first choice and fill in the bubble of your new choice.
9. Do NOT write in the test booklet. Mark only your answer sheet. If your booklet already has any inappropriate marks, ask for a clean booklet.

SPECIAL INSTRUCTIONS

1. You must begin with question 1. When you have chosen the best answer and marked your answer sheet, then you must go to question 2. When you have finished question 2, then you must go on to question 3, then question 4, then question 5, and so on to the end of the test.
2. Try each question once and only once. If you can't answer a question, go on to the next one.

Do NOT skip ahead to new questions or go back to old ones. Try to answer each question in its proper order.

END OF TEST

Close your test booklet.

Do not go back to any of the questions.

Make sure you have filled in your answer sheet with your name, sex, grade, and birth date.

Make sure that the Identification No. Box A has a 4 in it.

Turn in your test booklet and answer sheet.

THANK YOU

APPENDIX III

P-level Item Analysis Data

	Easy to Hard As Given Order			Hard to Easy Reversed Order		
Item	Seq. No.	Form 1 Unr.	Form 3 Res.	Seq. No.	Form 2 Unr.	Form 4 Res.
A	1	.913	.919	40	.728	.678
B	2	.906	.838	39	.707	.562
C	3	.866	.831	38	.762	.651
D	4	.832	.723	37	.619	.521
E	5	.678	.635	36	.558	.548
F	6	.765	.723	35	.599	.589
G	7	.812	.750	34	.565	.527
H	8	.624	.696	33	.531	.445
I	9	.604	.520	32	.537	.527
J	10	.705	.676	31	.660	.575
K	11	.651	.669	30	.449	.459
L	12	.470	.466	29	.435	.384
M	13	.631	.595	28	.483	.452
N	14	.617	.541	27	.503	.527
O	15	.503	.493	26	.442	.404
P	16	.369	.385	25	.374	.288
Q	17	.362	.399	24	.272	.418
R	18	.416	.351	23	.313	.342
S	19	.490	.527	22	.435	.438
T	20	.456	.392	21	.361	.397
U	21	.376	.358	20	.367	.411
V	22	.477	.419	19	.429	.308
W	23	.523	.527	18	.435	.500
X	24	.389	.378	17	.367	.329
Y	25	.329	.291	16	.374	.432
Z	26	.456	.459	15	.306	.377
AA	27	.315	.284	14	.293	.267
BB	28	.356	.351	13	.354	.390
CC	29	.275	.324	12	.272	.253
DD	30	.201	.257	11	.204	.267
EE	31	.362	.297	10	.313	.281
FF	32	.369	.439	9	.367	.370
GG	33	.255	.257	8	.218	.240
HH	34	.275	.209	7	.238	.219
II	35	.242	.223	6	.252	.151
JJ	36	.228	.250	5	.272	.288
KK	37	.242	.189	4	.238	.137
LL	38	.262	.182	3	.190	.205
MM	39	.161	.169	2	.122	.199
NN	40	.181	.176	1	.245	.247

Note Unr. = Unrestricted; Res. = Restricted

B-value Item Parameter Estimates

	Easy to Hard As Given Order			Hard to Easy Reversed Order		
Item	Seq. No.	Form 1 Unr.	Form 3 Res.	Seq. No.	Form 2 Unr.	Form 4 Res.
A	1	-2.743	-3.028	40	-1.257	-1.004
B	2	-2.653	-2.096	39	-1.134	-0.357
C	3	-2.206	-2.037	38	-1.476	-0.845
D	4	-1.909	-1.255	37	-0.641	-0.139
E	5	-0.915	-0.741	36	-0.325	-0.284
F	6	-1.427	-1.255	35	-0.535	-0.504
G	7	-1.752	-1.429	34	-0.360	-0.175
H	8	-0.633	-1.089	33	-0.188	0.262
I	9	-0.531	-0.128	32	-0.222	-0.175
J	10	-1.063	-0.970	31	-0.861	-0.430
K	11	-0.772	-0.931	30	0.233	0.189
L	12	0.126	0.155	29	0.292	0.602
M	13	-0.667	-0.520	28	0.051	0.226
N	14	-0.599	-0.234	27	-0.051	-0.175
O	15	-0.037	0.014	26	0.257	0.487
P	16	0.633	0.592	25	0.611	1.178
Q	17	0.669	0.517	24	1.195	0.411
R	18	0.392	0.783	23	0.950	0.839
S	19	0.028	-0.163	22	0.292	0.299
T	20	0.192	0.555	21	0.684	0.525
U	21	0.598	0.744	20	0.647	0.449
V	22	0.093	0.407	19	0.327	1.047
W	23	-0.135	-0.163	18	0.292	-0.030
X	24	0.529	0.630	17	0.647	0.921
Y	25	0.850	1.148	16	0.611	0.337
Z	26	0.192	0.191	15	0.990	0.641
AA	27	0.925	1.191	14	1.070	1.314
BB	28	0.704	0.783	13	0.721	0.563
CC	29	1.162	0.941	12	1.195	1.409
DD	30	1.657	1.369	11	1.658	1.314
EE	31	0.669	1.106	10	0.950	1.223
FF	32	0.633	0.299	9	0.647	0.680
GG	33	1.287	1.369	8	1.558	1.506
HH	34	1.162	1.710	7	1.416	1.660
II	35	1.374	1.608	6	1.325	2.259
JJ	36	1.465	1.415	5	1.195	1.178
KK	37	1.374	1.872	4	1.416	2.403
LL	38	1.245	1.929	3	1.762	1.768
MM	39	1.981	2.047	2	2.389	1.824
NN	40	1.812	1.987	1	1.370	1.457

Note Unr. = Unrestricted; Res. = Restricted

APPENDIX IV

<u>MATHEMATICS 8 EXAMINATION</u> (147)

TEST FORMAT 3

INSTRUCTIONS

1. Do NOT open the test booklet until you are told to do so. You will have 45 minutes to complete this test.

2. Carefully remove the answer sheet from inside the front cover and make sure there is a 3 marked in box A of the Identification No. section of the answer sheet.

3. Be sure you have a pencil, an eraser, and some scratch paper.

4. Fill in your answer sheet with your name, sex, grade, and birth date.

5. Do NOT use a calculator or a protractor.

6. For each question, select the best answer. Mark your choice on the answer sheet by filling in the bubble under the correct letter. Make sure the question number is the same as the question number in the test booklet.

7. Do not spend too long on any one question. Try your best pick a good answer to every question.

8. If you make a mistake, completely erase your first choice and fill in the bubble of your new choice.

9. Do NOT write in the test booklet. Mark only your answer sheet. If your booklet already has any inappropriate marks, ask for a clean booklet.

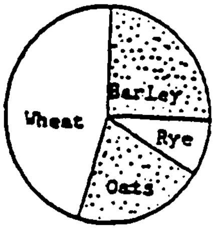
SPECIAL INSTRUCTIONS

1. You must begin with question 1. When you have chosen the best answer and marked your answer sheet, then you must go to question 2. When you have finished question 2, then you must go on to question 3, then question 4, then question 5, and so on to the end of the test.

2. Try each question once and only once. If you can't answer a question, go on to the next one.

<u>Do NOT skip ahead to new questions or go back to old ones.</u>
<u>Try to answer each question in its proper order.</u>

ITEM A: QUESTION 1



The circle graph shows the proportions of various grain crops produced by a country. Which of the following statements is TRUE?

● A More oats than rye is produced.

B The largest crop is barley.

C Equal quantities of wheat and barley are produced.

D The smallest crop is oats.

E Wheat and oats together make up less than half the total grain crop.

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM B: QUESTION 2

162 x 45 is equal to

A 1378

B 1458

C 5890

D 6290

● E 7290

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM C: QUESTION 3 (150)

A team scores an average of 3 points per game over 5 games. How many points altogether were scored in the 5 games.

A $\frac{3}{5}$

B $\frac{5}{3}$

C 3

D 5

● E 15

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM D: QUESTION 4 (151)

In a discus-throwing competition, the winning throw was 61.60 metres. The second place throw was 59.72 metres. How much longer was the winning throw than the second place throw?

A. 1.12 metres

[✓ B. 1.88 metres]

C. 1.92 metres

D. 2.12 metres

E. 121.32 metres

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM E: QUESTION 5

If $10^2 \times 10^3 = 10^n$ then n is equal to

A 4

• B 5

C 6

D 8

E 9

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM F: QUESTION 6

A group of children was divided into 7 teams with nine in each team. Later, the same group of children was divided into teams with seven in each team. How many teams were there then?

A 7

B 8

[✓ C 9]

D 16

E 63

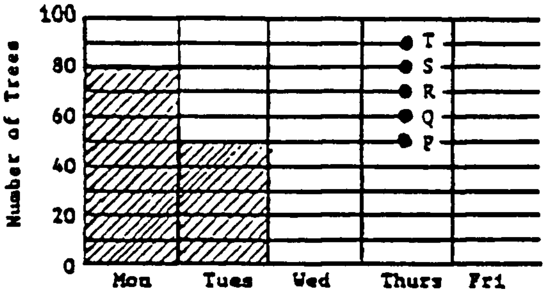
PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM G: QUESTION 7 (154)

Here is a table that shows the number of trees planted along a highway in a week.

Days of the Week	Mon	Tues	Wed	Thurs	Fri
Number of Trees Planted	80	50	60	90	75

On the diagram below, the graph for the first two days' plantings has been drawn.



If the graph were completed, which point would indicate the top of the bar on Thursday?

A P

B Q

C R

D S

E T

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM H: QUESTION 8

What is the volume of a rectangular box with interior dimensions 10 cm long, 10 cm wide, and 7 cm high?

A 21 cm^3

B 70 cm^3

C 140 cm^3

D 280 cm^3

[✓ E] 700 cm^3

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM I: QUESTION 9 (156)

If the ratio of 2 to 5 equals the ratio of n to 100, then n is equal to

A 10

B 20

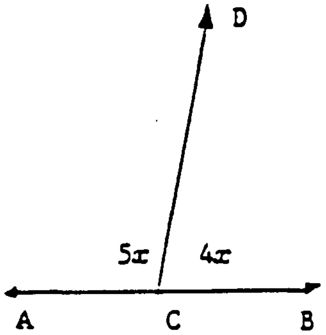
● C 40

D 150

E 250

<u>PLEASE DO NOT TURN BACK TO THIS PAGE.</u>

ITEM J: QUESTION 10



If AB is a straight line, what is the measure in degrees of angle BCD?

A 20

B 40

C 50

[✓ D 80]

E 100

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM K: QUESTION 11

In a school of 800 pupils, 300 are boys. The ratio of the number of boys to the number of girls is

A. 3 : 8

B. 5 : 8

C. 3 : 11

D. 5 : 3

● E. 3 : 5

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM L: QUESTION 12

Which of the following equals
$7 \times (3 + 9)$?

● A $(7 \times 3) + (7 \times 9)$

B $(7 \times 9) + (3 \times 9)$

C $(7 \times 3) + (3 \times 9)$

D 7×27

E $21 + 9$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM M: QUESTION 13 (160)

0.40 x 6.38 is equal to

A. .2552

B. 2.452

[✓ C. 2.552]

D. 24.52

E. 25.52

<u>PLEASE DO NOT TURN BACK TO THIS PAGE.</u>

ITEM N: QUESTION 14

When $x = 2$, $\frac{7x + 4}{5x - 4}$ = is equal to

A 11

[✓ B 3]

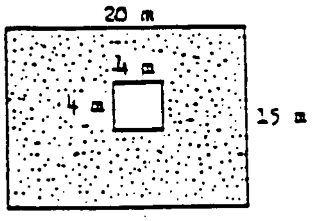
C $\frac{11}{5}$

D $\frac{9}{5}$

E $\frac{7}{5}$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM O: QUESTION 15



A square is removed from the rectangle as shown. What is the area of the remaining part?

A. 316 m^2

B. 300 m^2

[✓ C. 284 m^2]

D. 80 m^2

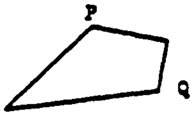
E. 16 m^2

PLEASE DO NOT TURN BACK TO THIS PAGE.

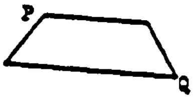
ITEM P: QUESTION 16

If segment $\overline{PQ}$ were drawn for each figure shown below, it would divide one of the figures into two congruent triangles. Which figure?

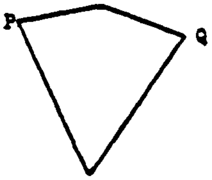
A



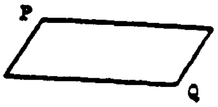
B



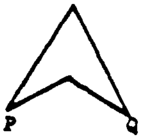
C



• D



E



PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM Q: QUESTION 17

$7\frac{3}{20}$ is equal to

A. 7.03

[✓ B. 7.15]

C. 7.23

D. 7.3

E. 7.6

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM R: QUESTION 18 (165)

What is the square root of 12 x 75?

A 6.25

[✓ B 30]

C 87

D 625

E 900

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM S: QUESTION 19 (166)

If $x = -3$, the value of $-3x$ is

A -9

B -6

C -1

D 1

[✓ E] 9

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM T: QUESTION 20 (167)

How many pieces of pipe each 20 metres long would be required to construct a pipeline 1 kilometre in length?

A 5

[✓ B 50]

C 500

D 5000

E 50,000

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM U: QUESTION 21

2 metres + 3 millimetres is equal to

A. 2.0003 metres

[✓ B. 2.003 metres]

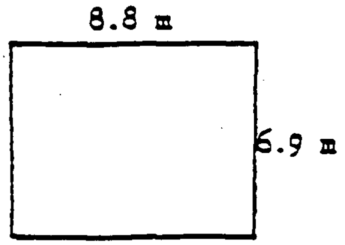
C. 2.03 metres

D. 2.3 metres

E. 5 metres

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM V: QUESTION 22



Which of the following is the closest approximation to the area of the rectangle with measurements given?

A. 48 m^2

B. 54 m^2

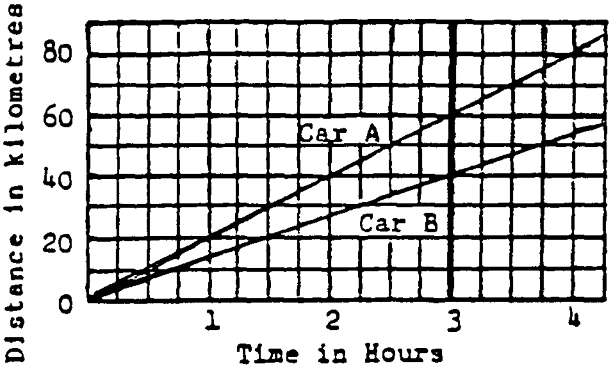
C. 56 m^2

[✓ D. 63 m^2]

E. 72 m^2

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM W: QUESTION 23



Three hours after starting, car A is how many kilometres ahead of car B?

A 2

B 10

C 15

[✓ D] 20

E 25

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM X: QUESTION 24

The arithmetic mean (average) of:
1.50, 2.40, 3.75 is equal to

A 2.40

[✓ B] 2.55

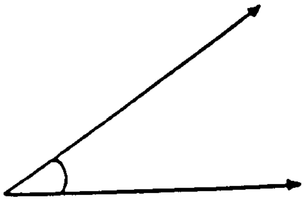
C 3.75

D 7.65

E None of these

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM Y: QUESTION 25 (172)



The measure of the angle shown is nearest to

A 155°

B 145°

C 50°

[✓ D 35°]

E 15°

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM Z: QUESTION 26

If $x = y = z = 1$,

then $\frac{x - z}{x + y}$ is equal to

A -2

B -1

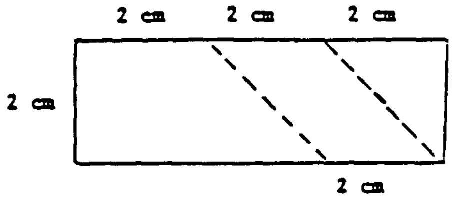
[✓ C 0]

D $\frac{1}{2}$

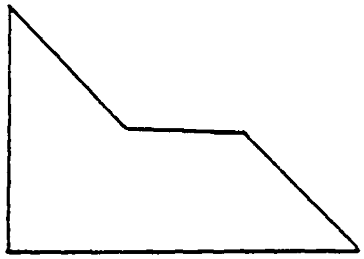
E 1

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM AA: QUESTION 27 (174)



The rectangle shown above is cut along the dotted lines and the three parts put together, without overlapping, to give the figure shown below.



The area in square centimetres of this figure is

A $8\ \text{cm}^2$

B $10\ \text{cm}^2$

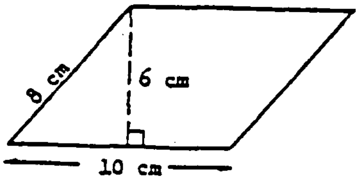
● C $12\ \text{cm}^2$

D $14\ \text{cm}^2$

E $16\ \text{cm}^2$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM BB: QUESTION 28



What is the area of the above parallelogram?

A 30 cm^2

B 36 cm^2

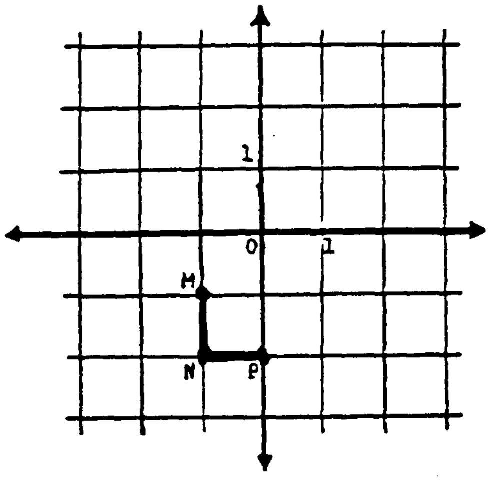
C 48 cm^2

● D 60 cm^2

E 80 cm^2

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM CC: QUESTION 29



Suppose you start at point M(-1,-1), move a distance of one unit to N(-1,-2), then turn left and move one unit to the point P(0,-2). If you again turn left and move one unit, you will now be at the point with coordinates

A (1, -2)

B (0, -3)

[✓ C] (0, -1)

D (-1, -2)

E None of the above

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM DD: QUESTION 30

0.00046 is equal to

A. 46×10^{-3}

[✓ B. 4.6×10^{-4}]

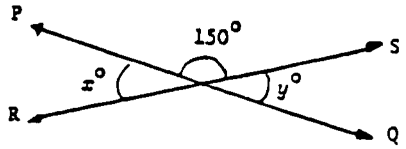
C. 0.46×10^{3}

D. 4.6×10^{4}

E. 46×10^{5}

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM EE: QUESTION 31



If, in the given figure $\overleftrightarrow{PQ}$ and $\overleftrightarrow{RS}$ are intersecting straight lines, then $x + y$ is equal to

A 15

B 30

● C 60

D 180

E 300

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM FF: QUESTION 32

The table below compares the height from which a ball is dropped (d) and the height to which it bounces (b).

d	50	80	100	150
b	25	40	50	75

Which formula describes this relation?

A $b = d^2$

B $b = 2d$

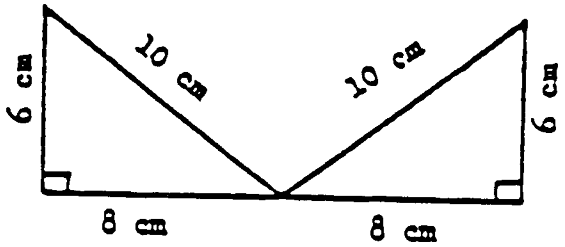
• C $b = \frac{d}{2}$

D $b = d + 25$

E $b = d - 25$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM GG: QUESTION 33 (180)



The total area of the two triangles is

[✓ A] $6 \times 8 \text{ cm}^2$

B $\frac{6 \times 8}{2} \text{ cm}^2$

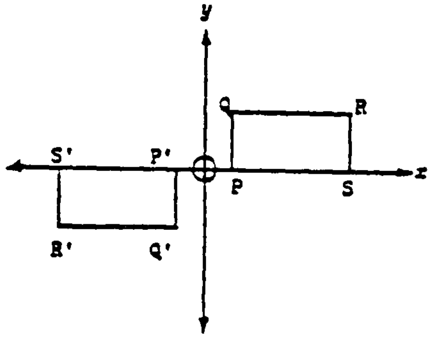
C $\frac{10 \times 6}{2} \text{ cm}^2$

D $\frac{16 \times 12}{2} \text{ cm}^2$

E $\frac{20 \times 12}{2} \text{ cm}^2$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM HH: QUESTION 34



PQRS is a rectangle. Its image after a transformation is the rectangle P'Q'R'S', as shown above. The transformation used could have been

● A a rotation about the origin.

B a reflection in the *y*-axis

C a translation parallel to the *x*-axis

D a reflection in the *x*-axis

E a translation parallel to the *y*-axis.

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM II: QUESTION 35

One of the following points can be joined to the point (-3,4) by a line segment which cuts NEITHER the *x* NOR the *y* axis. Which one?

● A (-2,3)

B (2,-3)

C (2,3)

D (-2,-3)

E (4,-3)

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM JJ: QUESTION 36

In a quadrilateral, two of the angles each have measure of 110°, and the measure of a third angle is 90°. What is the measure of the remaining angle?

● A 50°

B 90°

C 130°

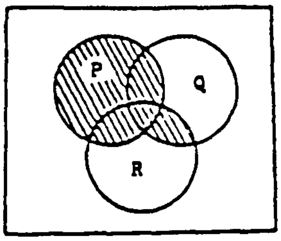
D 140°

E None of the above

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM KK: QUESTION 37 (184)

The symbol $P \cap Q$ represents the intersection of sets P and Q and the symbol $P \cup Q$ represents the union of sets P and Q. Which of the following represents the shaded portion of the diagram below?



A $(P \cap Q) \cup R$

● B $P \cup (Q \cap R)$

C $P \cap (Q \cup R)$

D $(P \cap Q) \cap R$

E $(P \cup Q) \cap R$

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM LL: QUESTION 38

There are 7,000,000 girls under the age of 21 in a country with a total population of 36,000,000. If a circle graph were drawn showing the distribution of the population, the angle in the sector representing girls under the age of 21 would have measure

A 7°

B 20°

C 21°

● D 70°

E 72°

PLEASE DO NOT TURN BACK TO THIS PAGE.

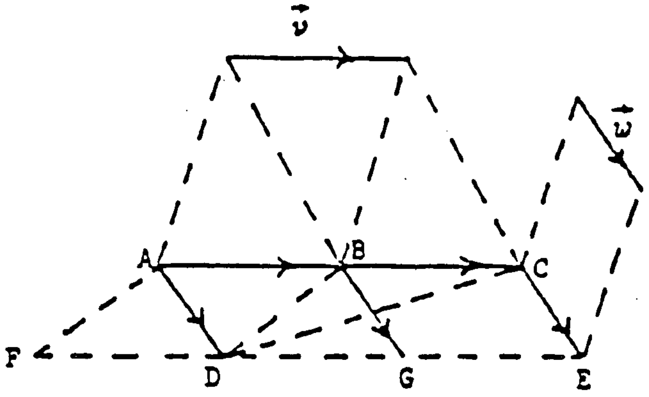
ITEM MM: QUESTION 39

If $5x + 4 = 4x - 31$, then x is equal to

- ● A. -35
- B. -27
- C. 3
- D. 27
- E. 35

PLEASE DO NOT TURN BACK TO THIS PAGE.

ITEM NN: QUESTION 40 (187)



A $\vec{v} + \vec{w}$

B $\vec{v} - \vec{w}$

C $\vec{w} - \vec{v}$

D $-\vec{w} - \vec{v}$

E $\vec{v} + 2\vec{w}$

Given $\vec{v}$ and $\vec{w}$ as shown in the figure above, what is $\overrightarrow{DB}$, the vector from D to B

PLEASE DO NOT TURN BACK TO THIS PAGE.

END OF TEST

Close your test booklet.

Do not go back to any of the questions.

Make sure you have filled in your answer sheet with your name, sex, grade, and birth date.

Make sure that the Identification No. Box A has a 3 in it.

Turn in your test booklet and answer sheet.

THANK YOU

AUTHOR INDEX

Author Index