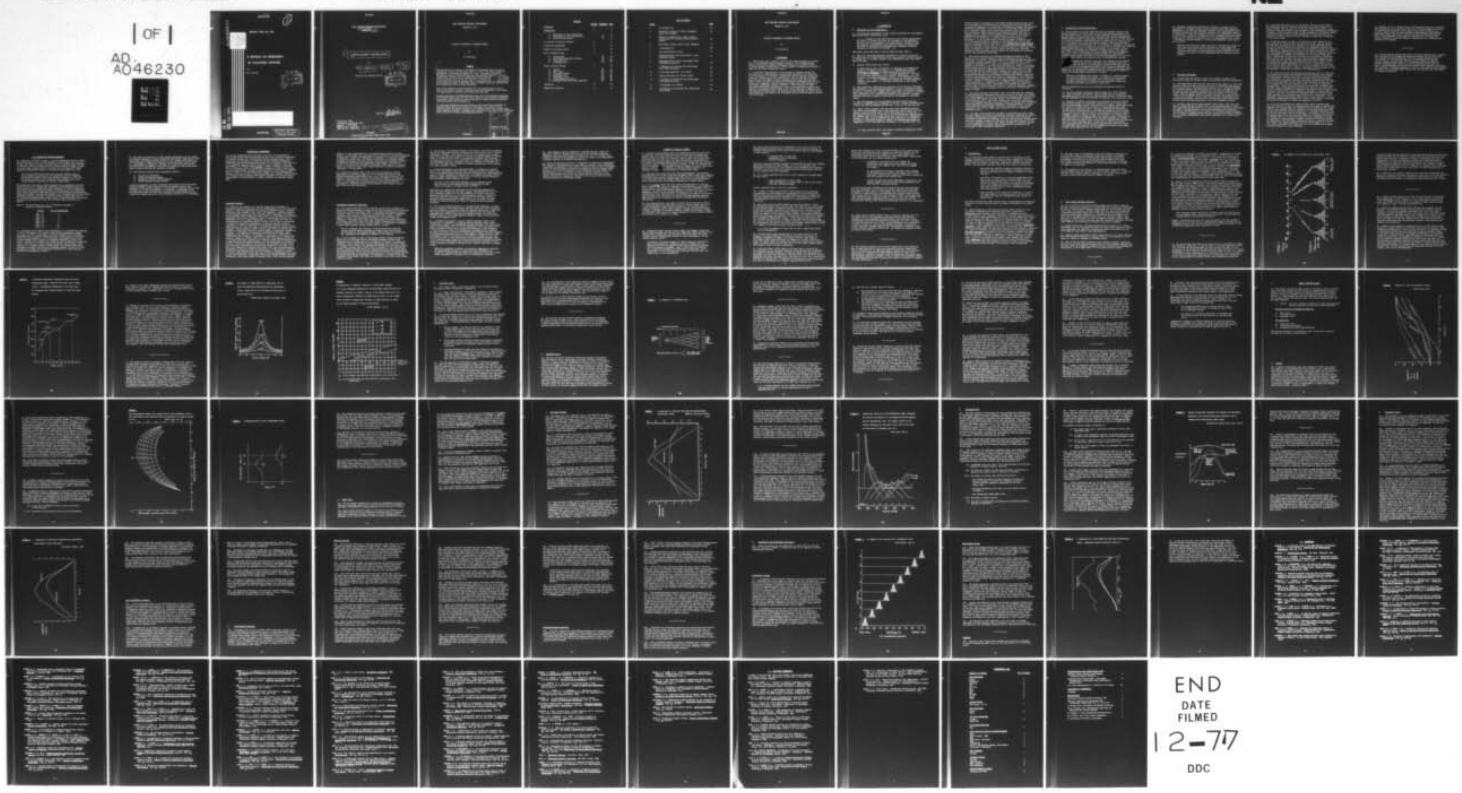
AD-A046 230 ARMY PERSONNEL RESEARCH ESTABLISHMENT FARNBOROUGH (EN--ETC F/G 5/10
A REVIEW OF RESEARCH IN TAILORED TESTING, (U)
1976 M C KILLCROSS

UNCLASSIFIED APRE-9/76 NL

| OF |
AD
AD46230



END
DATE FILMED
12-77
DDC

UNLIMITED

REPORT APRE No. 9/76

A REVIEW OF RESEARCH IN TAILORED TESTING

by

M. C. Killcross

MINISTRY OF DEFENCE
ARMY PERSONNEL RESEARCH ESTABLISHMENT

UNLIMITED

UNLIMITED

ARMY PERSONNEL RESEARCH ESTABLISHMENT

(14) APRE-
REPORT No. 9/76

(6) A REVIEW OF RESEARCH IN TAILORED TESTING,

by

(10) M.C. Killcross

(11) 1976

(12) 76p.

Received for printing 16.6.76.

Approved _____
Director

Project No: P560
(continuing as Study No. R43)
Sponsor: DAR (APSG)
APRE File No.: 502/1/30
MOD File No: 86/Res/2124

403 510

UNLIMITED

ARMY PERSONNEL RESEARCH ESTABLISHMENT

REPORT No. 9/76

A REVIEW OF RESEARCH IN TAILORED TESTING

by

M C Killcross

SUMMARY

Tailored testing is a method of psychological measurement that sets out in some degree to match the questions asked in a test to the person taking it. Matching may be achieved, for example, by using performance on earlier questions to select a subsequent question. The necessary flexibility is most conveniently available using a computer-assisted presentation in which a testee receives questions from an on-line terminal. A tailored test is likely to yield more information from each question than a conventional test. The uniformity of presentation which made conventional tests so widely viable and effective also set a restrictive limit to their efficiency.

This review considers research reported up to the second half of 1975. It looks at the general concept of tailored testing and especially at its potential application to selection and allocation in the Army.

Tailored testing has developed gradually using statistical methods originating in the middle 1940's. The last five years, however, have equalled the previous twenty-five in the volume of research reported as computer-assisted possibilities have become more realisable.

As a necessary foundation the review looks at the statistical antecedents of tailored testing and at the effects of a question's test context on its characteristics. The Report then goes on to consider research on tailored tests using sequential analysis, two-stage tests, branching tests, short tests, flexilevel tests, and the more recent item-finding procedures.

CONTENTS

LIST OF FIGURES

ARMY PERSONNEL RESEARCH ESTABLISHMENT

REPORT No. 9/76

A REVIEW OF RESEARCH IN TAILORED TESTING

by

M C Killcross

I. BACKGROUND

1. APRE Project 560 is concerned with the wide-ranging subject of alternatives to conventional pencil-and-paper testing. It is a forward-looking project which does not address an immediate problem nor necessarily expect to provide immediate applications: it hopes rather to break new ground wherein might lie the seeds of productive ideas.

2. This Report looks generally at the concept of tailored testing and especially at its application to selection and allocation in the Army. The aim of the Report is to present a constructive view of tailored testing and a review of research up to the second half of 1975. In this way the Report gives the basis for APRE's research programme in this area. Killcross and Cassie (1973), and Killcross (1974) have introduced the general aims for this research and suggested a possible approach it might take. The substance of this Report corresponds largely to chapters 1 and 2 of Killcross (1975) - with some modifications and additions - and is presented here to help accessibility. In retrospect McGill (1968) can be seen as putting the APRE toe into this particular research pool; the decision to enter the pool was taken in 1972 when Project 560 was being set up; this Report and the previous papers represent the first steps.

2. INTRODUCTION

A. Uniformity in test presentation.

3. Describing the thoroughness of the revision procedures for the Stanford-Binet scale Terman (1942) writes (p.8),

> "..... strange as it may seem there are still clinical psychologists who prefer 'a flexible test that can be adapted to the individual', one that 'will be custom-made to fit each subject'. Needless to say, the progress of psychometrics has consisted largely in escape from the chaos of subjectivity resulting from the impromptu procedures advocated by the author just quoted". (1)

(The author quoted and taken to task by Terman was Kent (1937)).

4. While not advocating impromptu procedures or a return to the chaos of subjectivity it is explicitly the aim of the research to develop a method of testing that is flexible, adapted to the individual testee, and custom-made to fit him.

5. Anastasi (1954) writes (pp. 22-24),

> "A psychological test is essentially an objective and standardised measure of a sample of behaviour. ..... Standardisation implies uniformity of procedure in administering and scoring the test. ..... Such a requirement is only a special application of the need for controlled conditions in all scientific observations. In a test situation, the single independent variable is the individual being tested. ..... Such standardisation extends to the exact materials employed, time limits, ... and every other detail of the testing situation".

6. Both authorities speak without reservation of the virtue of uniformity. The achievement of an unvarying presentation is seen as providing an armour by which a test might more strongly resist the thrust of extraneous influence. These views could be put so emphatically because they were given in a period when psychologists were comforted and perhaps in part beguiled by demonstrations that their youthful science had its own examples of scientific rigour.

7. What had happened was that uniformity had been imposed under the umbrella of standardisation. Uniformity is one way to achieve standardisation - and at the time was probably the best way - but it is not the only way.

8. DuBois (1970) considers the test item a very remarkable invention and traces its development from a crude subjective form in 1902 to an item-analysis supported entity that by the US Army Alpha test of World War 1 had already acquired some of what was to be its considerable sophistication. The emphasis came to be on group testing and on the test as a homogeneous assembly of items. That the assembly was also uniform for all testees was not merely a fashionable display of rigour: the undoubted philosophical congeniality of the requirement was secondary to its technical merit. The

(1) This quotation came to the author's attention through Wood (1973).

forceful advance of psychological measurement during the period when called on by the needs of two world wars was critically assisted by the benefits of cultivated uniformity of presentation. Nonetheless, although such uniformity contained a psychometric truth this was not a whole truth. That uniformity was a constraint caused little concern and attracted little attention or even realisation. Indeed variability was generally equated with loss of adequate control - hence Terman's chiding. Yet the general philosophy of psychological measurement is to sample behaviour with a view to extrapolating to the wider area of behaviour about which the sample is informative - so an accurate choice of behaviour sample is called for. If it is now stipulated that an estimate of intelligence, say, must be made based on identical behaviour sampling for all irrespective of their ability then this uniformity can be seen to be a handicap. Why ask a lengthy fixed series of questions of which perhaps a half are likely to be dead wood, being either too hard or too easy to do any useful work? Would not questions adjusted to ability provide a better sample? Yet the handicap of uniformity was not only accepted but advocated.

9. Why was the constraint of uniformity accepted? One answer is that removal of the constraint appears to depend on knowing the result of the test before it is given - on knowing the outcome before picking the questions. This answer, however, assumes that a test is an indivisible unit, whereas testing can alternatively be seen as an ongoing process in which it is possible to intervene. A test from this view-point is a series of encounters between testee and test item, each encounter providing a little more information and improving an accumulating estimate of the characteristic being measured. Why then was testing mostly assumed to be a unit and not viewed as a process? One reason was perhaps an incomplete emancipation from pre-Darwinian physicalism; while individual differences and dynamic change were central to the Darwinian theme and directly contributory to later interest in psychological measurement, the traditional methodologies of the physical sciences - suited to the manipulation of single passive variables and to total repeatability - continued to be influential. Among the physicalist concepts were the ideas that a measurement was something extracted in its entirety from a relatively defined situation, and that a measuring instrument was an enduring piece of equipment used for a particular variable over a range of values.

10. The early individual intelligence tests did acknowledge the redundancy of administering test questions which were too easy or too hard, and, within the overall sequence of the test, starting and stopping places were matched to the individual testee: even some choice in the ordering and selection of sub-tests was left to the tester. Perhaps the earlier test constructors were less blinkered than their successors became, or perhaps their flexibility simply represented an unresolved vagueness in conceptualisation. In either case later events moved strongly towards the elimination of any residual variability, not only for large scale testing but also for subsequent individual tests.

11. For large scale testing the constraint of uniformity also had a sound practical reason, the lack of an alternative flexible technology. The pencil-and-paper testing medium developed to facilitate large scale testing was a solution dictated by the technology of the day and made group testing synonymous with a uniform treatment of the group. For such group testing the move to strict uniformity to ensure truly common treatment is necessary, and one of the consequences of this has been an unproductive spill-over of the philosophy to other measurement approaches.

B. Variability in test presentation.

12. From as early as the late 1940s the idea of a variable test has been mooted. Variable in the sense that the test is deliberately varied to suit the individual testee, and moreover varied dynamically during testing using previous answers to help select later questions. One form of assessment has indeed always followed an individually varied course, and this is the interview: in a sense the previous sentence describes an interview. In a sense too the aim of a variable test can be construed as trying to keep the reliability of a conventional pencil-and-paper test while giving it something of the individuality (and humanity) of the interview. It was hardly an accident that Terman was remonstrating above with a clinical psychologist - a specialism in which the needs for assessment and recognition of individuality come together. Hutt (1947), also writing from a clinical setting, used the term adaptive testing to refer to an individually adjusted method of Stanford-Binet administration. However, the slant of our research approach is not clinical, it is concerned rather with objective forms of individualised testing.

13. The development of the idea of testing as an ongoing process adjusted to the individual testee will be presented in the following review, but in general terms the aim of such adjustment is greater efficiency and more recently this aim has perhaps been joined by one of greater individual consideration. In a very forward looking paper Hick (1950), considering intelligence tests in the light of information theory, puts forward in embryo an outline of an individually adjusted test (p.161),

> "Hence an intelligence test should, in theory, be a 'branch process'; ie. the first question should have a 0.5 chance of being answered by anyone from the general population. If the subject answers it, the next should have a 0.5 chance of being answered by anyone who has been successful with the first; and so on."

Hick is considering here only the information transmission aspects of efficiency.

14. The term tailored testing was coined by Lord (at a 1968 conference in a paper subsequently published as Lord (1970b)).This is the generic name adopted here for all forms of individualised testing. Traditional forms of pencil-and-paper tests will generally be referred to as conventional tests.

15. Progress in computer technology has now provided a practical medium for tailored testing. An individually tailored test can be conveniently given by maintaining a pool of test questions in computer storage and sitting the testee at a linked terminal which has a visual display unit (VDU) - of television screen type - and a keyboard. Questions are presented on the VDU and the testee answers on the keyboard. Depending on his performance on earlier questions the testee's next question is chosen by computer programme to match a running estimate (of his ability for example) that is being up-dated as testing proceeds. The attempt is made to optimise the questions chosen for presentation. Much the same technology has enjoyed wide exploitation by many workers in the field of Computer-Assisted Instruction. It is this technology that is in mind for an Army selection and allocation application. Tailored testing approaches have also been tried by researchers using computer teletype terminals and various non-computer methods; these approaches are also reviewed.

16. Although a convenient technological solution is necessary for any application of tailored testing, the main difficulty has been rather the development of an effective conceptual framework. In an individualised test different testees will take different questions. This immediately removes a cornerstone of classical psychometric theory and creates numerous problems of how to place testees on a common scale. The conceptual hiatus for early researchers in tailored testing should not be underestimated. Lord (1971e) - introducing tailored testing to a statistical readership - felt able to write (p.707),

> "However, the statistical reader need not be familiar with [even] the basic ideas of classical mental test theory - in particular, the notions of 'true score' and 'reliability'. The in-consequence of the classical theory here is surprising. Perhaps this indicates that the approach to be used is no less fundamental than the classical theory itself".

C. *Army entry procedures*

17. To make clear the context at which the research is aimed a brief description follows of the main points of the present Army entry procedures.

18. Non-commissioned entry into the Army for men and juniors is a two stage procedure. There are some variations in detail as between men and juniors, and as between men in Scotland and men from the rest of the United Kingdom: in principle, however, the broad approach is the same, and WRAC entry also has moved towards increased conformity with this general framework. The largest applicant group following one particular variant of the general scheme is men in England, Wales and Northern Ireland. This group provides - as will be described - the most fertile ground for the introduction of individualised testing, and the details and discussion which follow refer to this context and not necessarily to the other procedural variants.

19. The first stage of the entry procedure takes place at the local Army Careers Information Office (ACIO). This is the screening stage at which both applicants and Army take most of their decisions about broad suitability. The Army's screening decision is based on biographical and educational information from interview, on the results of a 30 minute pencil-and-paper test of general reasoning ability and basic arithmetic and verbal attainment, and on a medical examination. About 50% of applicants are screened out at the ACIOs.

20. Successful applicants go on to the Recruit Selection Centre (RSC) at Sutton Coldfield for the 2-2½ days which make up the second stage of the entry procedure. The men who go on to RSC are technically recruits, having enlisted at the ACIO; however, this enlistment is not binding and an honourable discharge is freely available whilst at RSC. Depending on recruitment between 12,000 and 20,000 recruits may pass through RSC in a year and only about 10% drop out from all causes. Tailored testing is visualised as being used at this second stage.

21. At RSC the concern is mainly with allocation, with finding the best match between the interests and abilities of the recruit and the needs of the Army. This matching process is helped by a two-way information flow: the recruit learns about the Army and about the employments available and the Army learns more about the recruit. For the recruit a formal and extensive job briefing is supplemented by a question and answer session and by interview and informal discussion opportunities; for the Army cognitive abilities and attainments are measured by a standard set of five pencil-and-paper tests, occupational interests and motivation are assessed by a pencil-and-paper inventory and at interview, and a detailed medical examination may be carried out. At interview a Personnel Selection Officer (PSO) continues the information exchange and ultimately helps the recruit decide on his first three allocation preferences. One of these is almost always offered to the recruit, and in 75% of cases it is his first choice. Now the recruit must accept the allocation or claim his discharge. With the minimum of delay the allocated recruit will be posted to his training depot.

22. The Army's selection and allocation procedures are seen to follow the traditional pattern based on standardised pencil-and-paper tests designed for uniform group administration. Up to five years ago this pattern was more appropriate, but then the present two-stage centralised procedures were introduced. Previously selection and initial allocation had both taken place at the many ACIOs, and the fuller pencil-and-paper testing with the five standard tests had been carried out only at a recruit's allocated training depot - where his final allocation was generally confirmed within the small range of employments available there. This was the traditional context calling for uniform test treatment. With many testers in many testing locations a system with a heavy emphasis on uniformity is exactly what is needed to combat the wide variety of experience that an allocated recruit would encounter. The traditional group test could be relied upon to give meaningful results when given by almost anyone almost anywhere - or, more moderately, administered according to its rules it was robust in resisting the potential influence of rather wide background variations.

23. It can be speculated that computer-linked remote terminals might have been possible at ACIOs or training depots even within an uncentralised framework, but clearly many more terminals would be needed and their usage would be less intensive than if installed at an all-Army Selection Centre. But now that there has been the change-over to centralised selection, and now that an alternative flexible technology is available, we are left with the paradox that the earlier necessary emphasis on uniformity has become an overkill. What was a protective armour is now an impediment, what was so successful in safe-guarding a minimum standard now limits the maximum. In evolutionary terms the earlier adaptation based on a standard selection test battery has become maladaptive for the new environment. It is as if all recruits continued to be issued with one size of uniform even though tailoring capacity has become available.

24. Benefits are to be expected from an individualised approach to psychological measurement: and in particular the view is held that cognitive testing for selection and allocation in the Army at high volume centres is well placed to benefit from the improved behaviour sampling that individualised measurement allows. The idea of tailored testing is deliberately limited here to the measurement of one cognitive characteristic at a time. The logical extension of tailored testing to varying the set of characteristics assessed is not pursued.

---

25. Tailored testing involves some kind of question selection from a question pool. The question is becoming the working unit rather than the whole test. It is perhaps not entirely fanciful to see an analogy between the move from the test to the individual question, and the progress of physical science through successive layers of increasingly microscopic levels of inquiry. Similarly from a wider view one can see parallels between this move to individualisation and the reaction against mass production, conformity, and increasing individual anonymity. It may be in the end that individualised procedures will be adopted simply because they are individual and not because of technical superiority.

3. AN OVERVIEW OF PREVIOUS RESEARCH

26. Research on tailored testing is a post-war phenomenon. Even so, despite having a history of 25 to 30 years behind it, such research has not grown to any great volume. Tailored testing has been a persisting idea, cherished in turn by a series of researchers, but laid down almost as often as taken up. It has held a promise which it has been slow to fulfil. Wood (1973) speaking of educational interest in tailored testing comments (p. 529),

> "For the past twenty-five years, this idea has exerted a more or less continuous fascination on the educational research community, and there has probably always been somewone working on it. Yet with the greatest respect to all concerned, these enquiries have never really amounted to anything of practical significance."

The attraction of the idea is almost tangible but an operational real-life application has yet to emerge. The traditional pencil-and-paper group test is, of course, one of psychology's major successes. It tends to monopolise educational and psychological assessment as IBM and Hoover have monopolised other fields. The major traditional test users have developed smooth-running, effective procedures that they are unlikely to be persuaded from by merely modest temptation. However, research on tailored testing is now growing: most of the literature references made in this review to work specifically on tailored testing were published after 1970. The following table illustrates the growth.

Table 1. The distribution over time of published literature on research in tailored testing.

Period	No. of publications
1944 - 47	2
1948 - 51	2
1952 - 55	2
1956 - 59	2
1960 - 63	3
1964 - 67	6
1968 - 71	21
1972 - 75	22

27. The growth can be linked to the availability of a facilitating technology - that of time-shared, fast computers. The earliest experimental studies - up to the late '60's in some cases - perforce attempted pencil-and-paper implementations of tailored testing. This was a stony road and a testimony to the drive and ingenuity of the researchers; such administrative inconvenience weighed heavily against any subsequent application. Much of the earlier literature also tended to be conceptual and concerned with theoretical results. The use of an on-line terminal was quickly seen as almost a pre-condition of a workable tailored testing system, and research gives every appearance of marking time while the technology ripened. It may be too that the upturn of concern for the individual characteric of recent years has provided a more supportive climate.

28. The nature of research on a topic develops and matures, but not evenly in different countries or across professional specialisms. Ideas persist with some groups, are dropped quickly by others or are perhaps never taken up at all. The analogy of changes in fashions of dress would have several points of correspondence. Consequently, although this review will parcel up the literature into four bundles, little more is claimed for the classification than a certain structural and conceptual usefulness.

29. The research literature may be grouped as follows:-

A. Statistical antecedents.
B. Effects of question context.
C. Background tailored testing research.
D. Recent tailored testing research.

A and C are important for their conceptual contributions rather than their detailed findings. C refers mainly to early pre-computer work or work peripheral to the research planned. B is necessary to establish the case that questions can in some circumstances be considered as independent units. D contains much detail of value and constitutes the main substance of the review. The research areas A to D are reviewed in Sections 4 to 7 following.

4. STATISTICAL ANTECEDENTS

30. In this background Section the aim is to trace the development of statistical methods that have provided the basis for a variety of approaches to tailored testing. These methods have usually originated with a view to applications in assessment or estimation problems outside psychological measurement. The main distinguishing feature of the methods is that they call for a sequential approach. Such an approach does not specify a one-piece experiment to be carried through in toto to permit estimation of the parameter of interest, rather it proceeds by a sequence of trials. These trials are not pre-determined, instead the specification for each trial is dependent upon the results of the preceding trial sequence. A second distinguishing feature is the type of data to which the methods may be applied. These methods are concerned with dichotomous experimental responses (in our case wrong or right answers to test questions) - usually referred to as quantal response data.

Sequential analysis

31. There are two threads to be followed through the development of sequential methods of estimation. The one which emerges as of less persisting interest is that associated with the Statistical Research Group of Columbia University (1945) and Wald (1947, 1950). This approach, sequential analysis, has been adopted extensively in the quality control procedures of manufacturing industries and is applicable where there is a large number of ostensibly equivalent items (rivets, resistors, spools of thread and the like). The problem here is how to sample effectively so as to estimate the level of a characteristic in a particular batch of output. (In our case we are wanting to estimate the ability or attainment of a person (the batch) from his responses to a sample of questions.) The sequential analysis solution is to take items one at a time and check if each in turn meets the required quality standard - thus providing a stream of yes/no data. After an item has been examined the additional evidence is used to update an appropriate cumulative statistic - for example, the Sequential Probability Ratio (Wald, 1947). Depending on the new value of the statistic a decision is taken either to classify the batch finally as acceptable or unacceptable or to increase the sample by taking in a further item, in which case the sequential procedure is repeated. This final classification decision is made with prescribed risks of false-rejection and false-acceptance. The tailored element of sequential analysis is thus the length of the sequence. The procedure concerns itself with successive decisions about whether there should be a next item or not, there is no question of tailoring the nature of the item. In psychological measurement it would not usually be appropriate or often possible to present a series of test questions that could be regarded as identical in nature. If it were sufficiently certain that such a series was appropriate this would in many cases mean that a sufficient estimate was already available. However, researchers, from Cowden (1946) to Ferguson (1971), have used this approach

either as a first approximation or in an educational setting for mastery testing. In the latter case, and especially in relation to criterion-referenced testing where a specific accomplishment is involved, it can well be a matter of repetitive testing (say, of division of fractions) to establish whether acceptable proficiency has been achieved. Research using sequential analysis in individual educational and psychological measurement is reviewed in Section 6. A statistical development by Armitage (1950) supports a multiple final classification rather than a simple split and is used by one group of researches to be described.

32. Sequential analysis, then, forms the basis of useful but limited applications of tailored testing. It is not a method applicable to the general measurement problem where there is no fixed value in mind for the parameter being estimated. However, the explicit formulation of decision risks is a characteristics relevant to the tailored testing application in view. It will emerge that the second thread to be picked up from the development of sequential methods, while generally more helpful, does not have a decision risk orientation.

Up-and-down sequential estimation

33. The second thread leads to the research of most direct relevance. It began with methods originally devised for testing the sensitivity of explosives by Anderson, McCarthy and Tukey (1946) of the Statistical Research Group at Princeton. These are the "staircase" or "up-and-down" methods of sequential estimation. Dixon and Mood (1948) suggested that these methods could be applied in other fields and proposed estimators that were taken up in bioassay or toxicology. These ideas were subsequently developed extensively in bioassay, and much of this work offers useful comparison with its psychometric equivalent. Lord (1970 b) in what amounts to a foundation contribution to much recent tailored testing research writes, (p.140),

> "It is a fortunate fact that most of the problems dealt with here closely parallel similar problems in bioassay. Much fruitful work has been done on the bioassay problems. This provides the inspiration, the background, and indeed the backbone of this chapter."

34. It may be helpful to look at the analogy between bioassay and psychometrics in a little detail before following developments further. The bioassayist has an insecticide, say, for which he is trying to estimate the lethality. He has control over the dose administered and can observe death or survival in his insects. The confrontation between dose and insect results in a quantal outcome, life or death. This is analogous to the confrontation between a person's ability and a test question and the outcome fail or pass. The analogy does not hold for what is controlled. Whereas we have information about the difficulty of out test questions and try to infer an unknown ability by manipulating question difficulty, the bioassayist is unable to vary the resistance of his insects and infers the lethality of his insecticide through varying its dose.

35. The essence of up-and-down methods is that after a trial with an observed outcome the independent variable is altered for the next trial so as to favour the opposite outcome - after survival the dose is increased, after a wrong answer an easier question is asked. In this way an overall balance of outcomes tends to be achieved. The details of up-and-down procedures are concerned with how the next trial is to be specified (for example, in what steps should the independent variable be changed), with how a decision to terminate the trials is to be made, and with how the observed responses are to be converted to a final estimate or score.

36. Further developments of up-and-down methods in bioassay are outlined below. A few additional details will also be given in Section 7, where they can be more appropriately mentioned after the introduction of background theory which it would be unhelpful to present here in a general statistical context. Brownlee, Hodges and Rosenblatt (1953) wrote of the slow initial take-up of Dixon and Mood's (1948) proposals (p. 262),

> "In spite of this efficiency advantage, the up-and-down method does not seem to have been given much consideration in such fields as bioassay or fatigue testing of metals."

They went on to confirm the superiority of the sequential approach over previous probit methods even for small samples, and proposed the use of a more convenient estimator of the main parameter of interest. They also proposed the possible use of two (or more) parallel series of trials. In their case they were concerned to make good use of the delay sometimes necessary between consecutive bioassay trials, but this idea is of some interest in tailored testing as a means of checking possible anomalous responses.

37. The block up-and-down method is a straightforward extension, of convenience in bioassay, that treats several insects in one trial. This convenience does not translate to tailored testing but administering blocks of questions does permit more complicated rules for choosing the next block and such approaches have attracted some tailored testing research. In bioassay the blocking method has been investigated by Wetherill (1963), Cochran and Davis (1964), and Tsutakawa (1967).

38. A different development is that by Robbins and Monro (1951). Their proposal may be regarded as a shrinking-step up-and-down method. Larger alterations are made to the independent variable initially, with ever smaller steps as the procedure zeroes in on the appropriate level. Such methods were found by Wetherill (1963) to be extremely satisfactory in some instances. It will be seen later that full Robbins-Monro approaches are not possible in tailored testing, but modified shrinking-step procedures have been proposed. The Robbins-Monro proposals were for large samples. Cochran and Davis (1965) investigated a number of Robbins-Monro procedures for samples of fifty and less and were able to offer useful gains over non-sequential designs. Davis (1971) compared several sequential bioassay methods and concluded that delayed variants of both Robbins-Monro and up-and-down procedures gave good results in all situations. Writing so recently he was, however, still able to comment (p.80),

> "While the asymptotic properties of sequential experiments, especially the Robbins-Monro process, are relatively well established, the accuracy of estimates and the guiding principles for the design of small sample experiments in bioassay are as yet incompletely explored."

39. A new approach, and one paralleled in tailored testing at about the same time, is that of Freeman (1970) who introduces Bayesian sequential estimation. This will be developed in Section 7 but it is of interest that research in tailored testing appears at about this stage to be coming abreast of general advances in stochastic approximation.

40. Before leaving this section it is appropriate to mention for completeness that research in psychophysics has also taken an interest in the developments in sequential estimation in other fields. Cornsweet (1962), Taylor and Creelman (1967) with their Parametric Estimation by Sequential Testing, Kappauf (1969) and Rose et al (1970) are examples. As the psycho-physicist (in common with the bioassayist) is in a position to vary the physical intensity of his stimuli (the analogue here of ability) rather than the sensitivity of his subjects there has not apparently, perhaps for this reason, been any direct cross-fertilisation with tailored testing.

5.EFFECTS OF QUESTION CONTEXT

41. Tailored testing methods select for presentation questions drawn from a larger pool. Two testees may well receive no questions in common; when they do receive the same question it will in most cases follow different preceding items and occur at a different stage in their test session. Does such variation in context affect an item's psychometric characteristics? The tailored testing procedures proposed have all assumed that item characteristics will remain stable irrespective of context. It will be necessary for real-life applications of tailored testing to examine the size of any context effects. To maintain the viability of tailored testing research it is necessary to establish the possibility of context-free items.

42. "Question" and "item" have been used interchangeably in the previous paragraph and will be used in this way throughout. While "item" and "response" are strictly more accurate -because many tests are not in interrogative form - the terms "question" and "answer" often allow less stilted descriptions and are used here with a more general meaning than literally theirs.

43. It is a priori likely that the content and difficulty of a question series could be made such as to influence the performance of some constituent items. However, the appropriate question is not, "Are substantial effects possible?" but rather, "Are effects likely?" A tailored test tries to present homogeneous items of about the same difficulty. It is in the nature of tailored tests that item difficulty is concentrated in a more or less narrow band appropriate for the testee. In this way the individualised approach might avoid the worst situations for effects stemming from frustration or demotivation.

44. The investigations so far made of context effects have been confined to pencil-and-paper tests apart from one or two recent studies. The use of a Visual Display Unit (VDU) computer terminal as the testing medium is a change that demands caution when looking to findings from pencil-and-paper settings. Accordingly the research findings reviewed below cannot be taken as definitive: their function is simply supportive. All the studies reviewed were carried out in schools or colleges.

---

45. Mollenkopf (1950), Sax and Cromack (1966), and Flaugher, Melton and Myers (1968) establish the basic general finding that under essentially non-speeded (power) conditions item statistics and correlations with other variables are not significantly affected by item rearrangement. Sax and Cromack conclude (p.311),

> "In general, the results support the thesis that test constructors have a responsibility of arranging items in ascending order of difficulty if tests are lengthy or time limits restricted. Evidently little is gained in arranging items if time limits are generous. Nor is there any advantage in constructing 'motivational' tests, consisting of a few easy items mixed with more difficult ones, over random forms of item arrangements".

46. Marso (1970) carried out two experiments. In the first a pool of two hundred 4-option multiple-choice vocabulary items was used to assemble a 139-item test displaying a wide range of difficulty. This test was arranged in three formats,

- ascending order of difficulty
- descending order of difficulty
- randomly arranged.

The three forms were randomly assigned to one hundred and twenty two students, previously classified as high, average, or low on test anxiety, and administered as power tests. The different item arrangements were found not to relate to score achieved. Test anxiety did affect achievement score but did not interact with item arrangement.

47. In a second experiment Marso used a course examination arranged again in three forms,

- topic presentation in course order
- topic presentation in reverse order to that of the course
- questions randomly arranged.

Results confirmed those of the first experiment.

48. A number of studies have looked at item context in the course of investigations of item sampling for estimating test norms (Lord (1962), (1965)). Here subsets of items are administered and used to estimate the mean and standard deviation of the whole test. Any systematic effects on item performance would evidence themselves in systematic errors of estimation.

49. Owens and Stufflebeam (1960) comparing contrasting samples of about two thousand 4th grade school children used item subsets of 3, 6 and 9 items from 50 multiple-choice vocabulary questions. 17, 8 and 4 different subsets of these three lengths respectively were administered to fractions of each sample. The population mean and standard deviation were as well estimated from item samples as from equivalent pupil samples. Both sampling techniques showed less precision in estimates of the mean for the higher ability pupils from advantaged neighbourhoods. Pupils, having taken an item subset, went on to attempt the rest of the 50 items, so that Owens and Stufflebeam were also able to look specifically at whether variations of item sequence affected test performance. The results from varied and standard sequences were so close as to suggest, the authors conclude cautiously (p.82),

> "..... that the sequence of items need not have a significant effect on test performance".

50. Sirotnik (1970) looked specifically at the context effect in item sampling. He investigated mean and variance estimates from subsets of vocabulary (synonym), arithmetic, and teacher attitude items taken by 180 students under power conditions. No support for a context effect was found, the author giving his opinion that the mean estimates were relatively immune to context effect for all three types of item, while further studies were needed to look at variance estimation.

51. Feldt and Forsyth (1974) looked at the same topic as Sirotnik for school grades 9 to 12. All pupils took one of two special tests in addition to a regular attainment battery. For about 130 pupils from each grade the additional test was of the ability to identify correct and effective written expression. For about 350 pupils from each grade the test was of quantitative thinking and involved some interpretation of graphical and tabular material. Both sets of experimental test material comprised subsets from

parent tests parallel to a test in the main battery. No net context effect of any size was evidenced by the language test material. However, for the quantitative questions the mean estimates from the item samples were consistently larger than for the whole test. Feldt and Forsyth, speculating on the difference, gave as possible explanations,

- a decrement in motivation with test length, the quantitative item sample was only a quarter the length of the full test compared with a half for the language item sample.

- or, and possibly more likely, the greater mental demands of the quantitative test led in the longer test to clear experience of failure with negative motivational consequences that were avoided in the shorter test.

- or that the time factor had inadvertently favoured the item samples (against this was the fact that noncompletion was less than 1% in the main battery).

If either motivational explanation were correct this would not necessarily mean that performance on item subsets had lower predictive validity than performance on these items in a full-length test: the reverse could even be argued. It would mean that item-sampling norm estimates would initially require some form of corroboration. For tailored testing it would mean that item calibration from a long test might be suspect. Once item standards had been validated against longitudinal criteria in the usual way then for selection and allocation the calibration difference would be of no consequence.

---

52. Apart from an overall context effect a carry-over influence from the difficulty level of the immediately preceding item has been claimed - especially an error-proneness following failure. Huck and Bowers (1972) reviewed such claims and investigated the possibility of bias in estimates of item difficulty from such a cause. Course examinations were prepared in a variety of orders for 120 and 160 psychology students. An analysis of variance procedure designed expressly for testing whether treatments (items in this case) have carry-over effects (Williams (1949)) was employed but did not detect such effects.

---

53. The research reviewed in this Section clearly allows the possibility that in some situations at least item characteristics are context-free. This is sufficient for the immediate purpose. Transferable item characteristics are best obtained from untimed administrations of short tests. Multiple-choice vocabulary (synonym) items are among those which have shown (for students) immunity to context. It may be that in the psychometric theory that will evolve for tailored testing a place should be reserved for indices of context-reliability.

6. EARLY TAILORED TESTING

A. Introduction.

54. In this section three kinds of research will be discussed. All are directly concerned with tailored testing but have in common that they are somewhat distant from the main line of Project 560. Sometimes the distance results from a difference in approach, sometimes it reflects the vastly greater computing power available today. The three kinds of research are,

- work using the sequential analysis procedures outlined in Section 4. These are the procedures associated with Wald and were described above as leading to research of lesser relevance

- work before 1970 based on a fixed step up-and-down method that steers a testee through a pre-determined lattice or network of paths between items. Tests of this kind and period were usually referred to as branching or programmed tests. This nomenclature reflects a general influence or push from the then topical field of programmed learning and teaching machines.

- work, possibly rich in ideas, but limited in its scope by the limitations of the technological facilities used or available. Such work (as reviewed here) was carried out before 1970.

The above classification simply defines what is being regarded as background research: it is not a division which can always be followed in the review below.

55. Before moving on to the earliest tailored testing research it will be useful to distinguish the following ways used to collect data. Theoretical studies attempt within the limits of mathematical tractability to model a test situation. Mathematical functions which might show or have shown working approximation are used to explore tendencies, relationships and limits. The range of theoretical studies possible has been considerably extended by the availability of computers capable of executing solutions by numerical methods for the less tractable situations. Within the limits of their assumptions such methods are very powerful. Monte Carlo simulation studies generate test data from a theoretical base. This data will be a planned sample from the given area and will help explore a situation too complex or difficult to explore more exhaustively by theoretical means. Real-data simulations are based on data from encounters between real people and real questions. Such data is used as if it had occurred in a tailored test. In this way a sequence of individually selected items may be taken from a testee's test record with no regard to the original test order of items. Empirical studies are real-life tailored tests, presenting real people with real items and tailoring the choice of items to the individual person during testing according to the procedure being investigated.

56. The above order of presentation of the four methods is generally one of decreasing research accessibility. Empirical research with computer assistance is an expensive undertaking and usually follows only after preliminary research by one or more of the other methods. On the other hand attempts at pencil-and-paper implementations of tailored testing are not expensive and have been embarked on without preliminaries in a number of studies to be described.

57. The remainder of the Section is a study-by-study review with interspersed summary views and comments. For these earlier studies it is often the case that other details become as important as the results proper. These instances provide particular pegs for comments to hang on.

B. Tests using sequential analysis.

58. Let us take first for review research making use of Wald's sequential analysis. The earliest application to educational and psychological measurement was that of Cowden (1946). Perhaps unsurprisingly his was an empirical study with a class of statistics students. Grades for the course were assigned by a sequential procedure using a pool of 200 items from which subtests of 20 items were administered separately as conventional pencil-and-paper tests. Each subtest was marked before students went on to the next. Students only went on to a further subtest if - in the sequential analysis method - their performance so far had not classified them with sufficient confidence. He found three subtests were sufficient to classify a majority of students.

59. Moonan (1950) used a real-data simulation from responses to a 75-item achievement test. He investigated how well an item by item sequential analysis could approximate the pass/fail classification based on the whole test. On the average 40 items showed a good approximation.

60. Anastasi (1953) and Burgess (1955) reversed the roles of testee and item. They used sequential analysis to classify items for test suitability on the basis of a series of responses by different people.

61. None of the four early researches above apparently offered a persuasive utility, for the next tailored testing studies to use sequential analysis were not until 1968. So although the approach had been demonstrated in psychometric applications - and to some effect - it was not perceived as useful.

62. A group of researchers - Cleary, Linn and Rock - experimented in a number of studies with variations of an elementary form of tailored testing. This form is two-stage testing. Here testing is in two parts. The first stage is common to all testees and is aptly termed a routing test as its function is to steer or allocate testees to the most appropriate of several tests making-up the second stage of the procedure. The tests in the second part are relatively specialised, say by ability level, and are referred to as measurement tests. Figure 1 illustrates a two-stage procedure in which a 10-item routing test directs a testee to one of five 20-item measurement tests. Two-stage testing could be used with pencil-and-paper tests, especially if there were a little time between stages. A screening test governing admission to a full test battery could be viewed as a special application of a two-stage strategy, but more typically both routing and measurement tests are short by conventional test standards and all testees proceed to the second stage.

63. Cleary et al used sequential analysis for some of their routing tests and with some success. (Other forms of two-stage testing are discussed later). Their technique was that of Armitage (1950) in which allocation to measurement test depended on the cumulative value of a probability ratio statistic. Real-data simulation from responses to items in scholastic tests taken by large samples of 11th-grade pupils and college students allowed an item-by-item consideration of performance on the subset of items selected to make a routing test. Of course, contrary to the requirement of Wald's approach, the items in the routing subset were not equivalent and this was recognised by the researchers; the items differed in both difficulty and discrimination so that the theoretical assumption regarding responses as random trials of a random variable held only to an approximation. However, this approximate sequential analysis strategy was among the more successful of the routing possibilities explored. In the first study (Cleary et al, 1968 a) "sequential item sampling" was one of four routing methods tried. (Their other methods are referred to later). Of these methods (p.357),

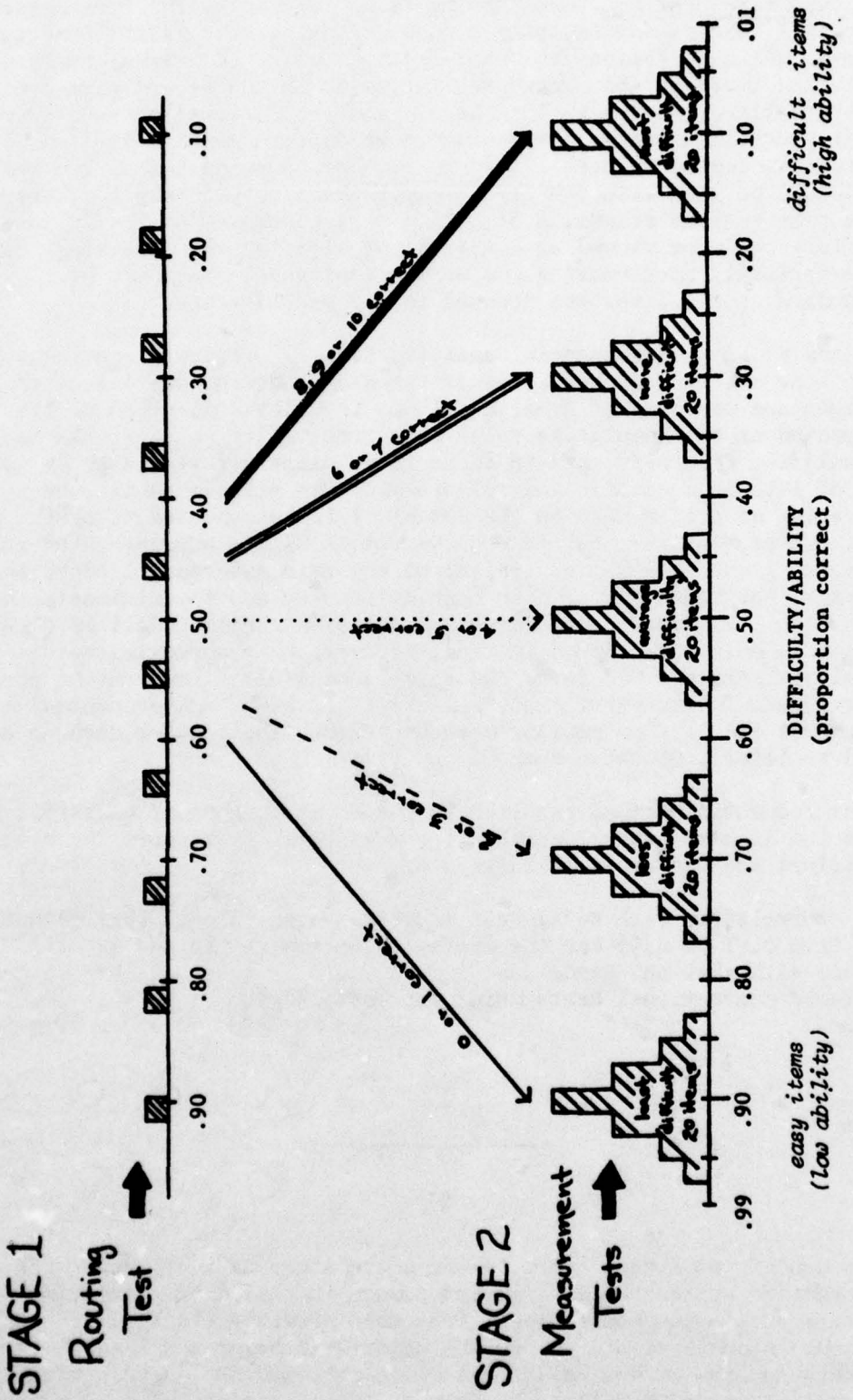
> "The sequential method resulted in the fewest errors of classification and the highest overall correlation with total test score for both the original and the cross-validation samples".

However, correlation with total test score was high for all four methods - ranging from 0.91 to 0.96 for the cross-validation sample and generally only comparable with what the study also showed could be achieved through the use of shortened conventional tests using the best items.

---

64. The use of total test score in the above study as a criterion for comparing alternative approaches is a device common in real-data simulation studies. It is the score on the conventional test that provides the basis for the simulation. As one estimate of the characteristic being assessed it is clearly appropriate to look at how well total score corresponds in turn with estimates by alternative means. Nonetheless, as a criterion, total score on a conventional test has limitations. Reproduction of conventional test estimates is not the prime purpose of tailored testing. Both conventional testing and

FIGURE 1 An example of a two-stage test (after Weiss, 1974).



tailored testing have the common aim of assessing psychological characteristics, and both no doubt can be expected to achieve this less than perfectly. Consequently while useful as a screening criterion a conventional test score is inappropriate for finer evaluations. By definition an improved method of assessment (or an equal but different method) will have a high but significantly imperfect correlation with existing methods. The research of Feldt & Forsyth (1974) mentioned in Section 5 suggests, for example, that conventional and tailored testing might in some cases differ in their susceptibility to motivational influence.

65. In Cleary et al's study it should also be noted that correlations with total score carried a part/whole inflation (40 items out of 190). The simulated shortened conventional test also carried the same inflation so that comparability was not lost. However, an independent total score could well be used.

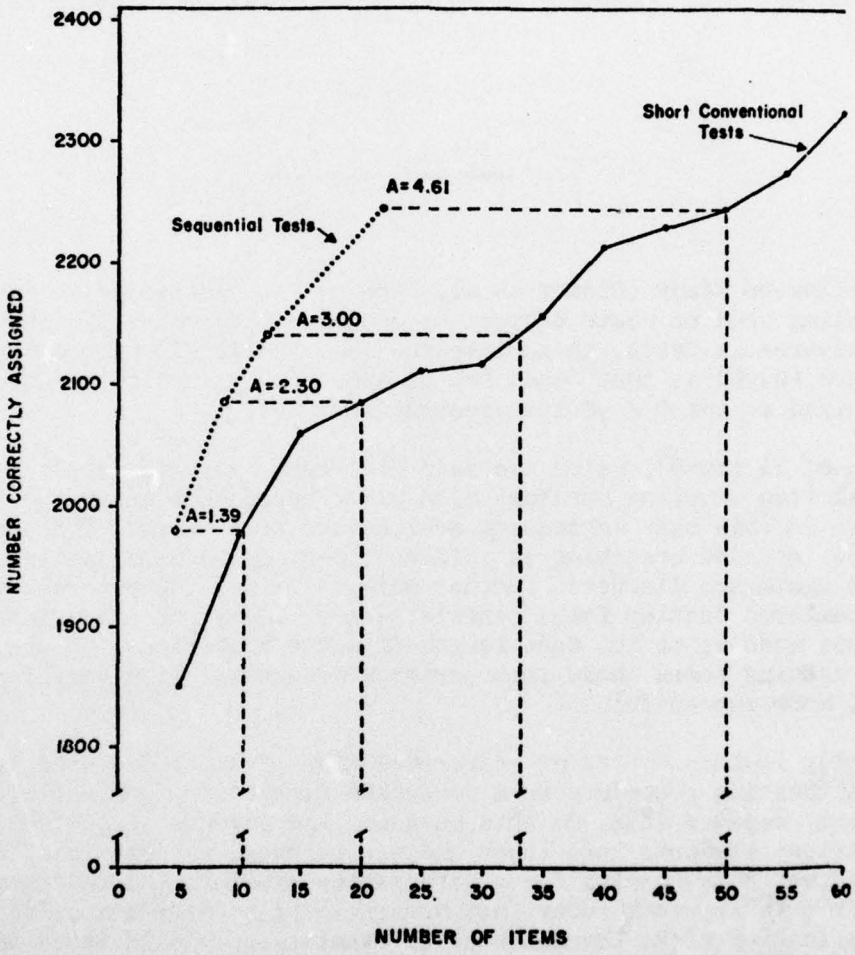
---

66. A follow-on study (Cleary et al, 1968 b) was restricted to sequential item sampling used to route testees to one of either three or four second-stage measurement tests. To achieve the same (inflated) correlation with total score (0.96) as that found for an average 37 items in the two-stage test required at least a 50-item conventional test.

67. Linn et al (1969), using the same real-data base, compared the same sequential item sampling strategy with other approaches against external criteria - in this case subsequent achievement test scores. The other approaches included branching as well as two-stage forms of tailored testing and these again are discussed further below. Against the external criteria all the tailored testing forms correlated more highly than conventional short tests made up to the same length from the best items. Of the several tailored testing forms those incorporating sequential item sampling were among the more successful.

68. Finally in this series of researches Linn et al (1972) used the same sequential testing procedure in a real-data simulation from college student examination response data. On this occasion the success of sequential testing in classifying students into lower and upper groups was examined. Figure 2 illustrates their results for a mathematics examination. The increasing values for A in Figure 2 refer (not numerically) to decreasing levels of misclassification risk. The mathematics examination was 75 items in length. In this study the sequential testing took items in the same order as in the examination. Generally sequential testing required about half the items needed by conventional tests for the same number of correct classifications. For sequential testing it is average number of items required that is plotted; students away from the cutting point would generally need fewer than this average, those closer would need more.

FIGURE 2 A simulated comparison of sequential tests and short conventional tests. (Results from Linn, Rock & Cleary (1972): a mathematics examination is the basis here for assigning 2420 college students to lower and upper groups.)



69. Results in two other examination subjects were similar. These results agree closely with those of a theoretical study by Green (1970) which are illustrated in Figure 3 for a sequential test used to classify Ability Level as less or greater than a standard score of zero.

---

70. Working in a context better suited to the requirement of equivalent items Ferguson (1960, 1971 a & b) also employed a sequential analysis approach. The context was individually prescribed instruction and he was working on the assessment of proficiency in learning objectives. Especially at the elementary level, and perhaps especially in mathematics, it becomes possible to formulate item generation procedures (for example, to produce items calling for the addition of two 2-digit positive numbers less than 50). Where this is possible computer assistance can be used to generate further equivalent items to a built-in specification as they are required for testing. Figure 4 is the classic quality control chart as applied by Ferguson to proficiency testing. Ferguson (1971 a) describes an application of the sequential approach using item generators for testing various levels of addition-subtraction proficiency in an empirical study with pupils in grades 1 to 6. Questions were presented on a computer teletype terminal and responses made on a partially covered keyboard. No practical difficulties were reported. Branching rules for moving to the next objective were written so as to allow skipping up the objective hierarchy when high proficiency was established. Branching reduced the testing time required (although from an educational standpoint a more important finding was that more items were generally found necessary for proficiency decisions than the conventional test procedures had allowed). Assessments from the sequential procedure were judged as valid and reliable as those from conventional tests.

---

71. The sequential analysis procedures reviewed above have shown benefits in applications calling for assessments to divide people into two (and possibly three and four) subgroups on either side of a pre-determined cutting level. This situation is likely to arise in educational or training programmes in relation to mastery of units of instruction: on the other hand the method would not cope comfortably with the provision of diagnostic information in the case of non-mastery. A possible use in selection would be for the initial screening of job applicants where minimum qualifications on critical abilities and attainments could be tested in this way. (A further APRE research project, P582, "The use of visual display units to improve testing procedures at Army Careers Information Offices", is looking at this possibility). However, sequential analysis is not an appropriate method for helping the general allocation of personnel, although its explicit formulation of misclassification risks is a desirable feature.

FIGURE 3 The number of items needed by a sequential test to match the operating characteristics of conventional tests n items long for the decision Ability above or below scale zero.

(Theoretical results from Green, 1970)

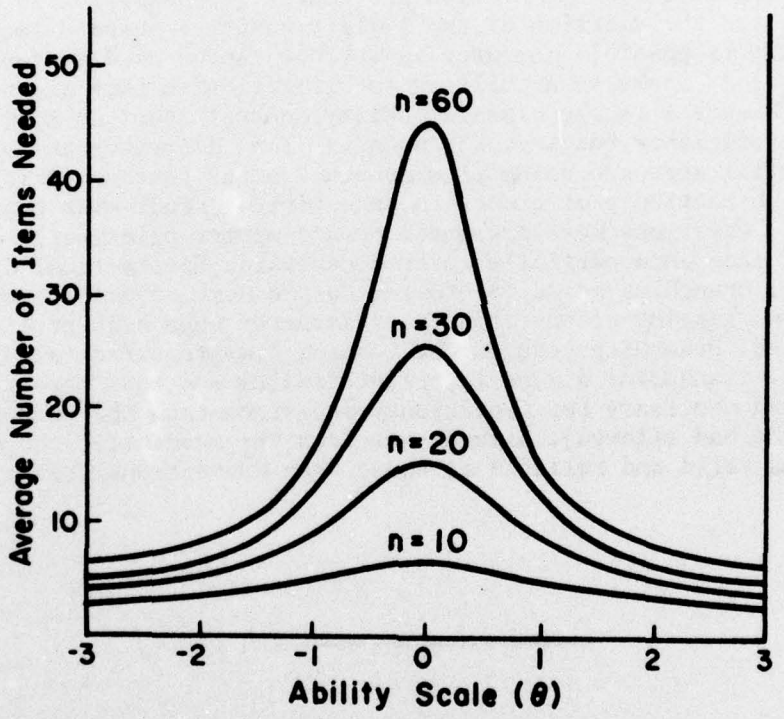
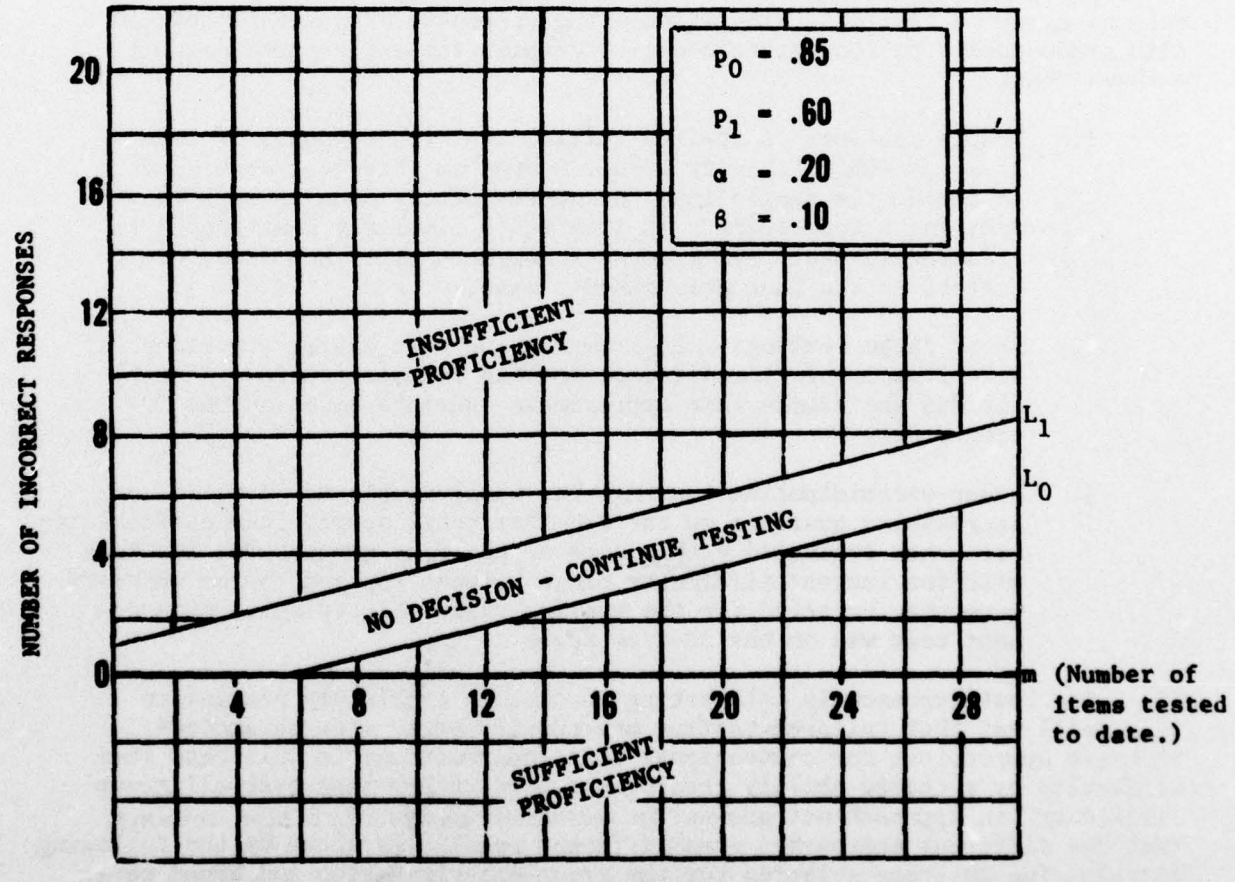


FIGURE 4

An application of sequential analysis to proficiency testing.

If p is the (unknown) proportion of all equivalent items that would be answered correctly by a pupil, then p_0 is the highest level for p for which a wrong-reject decision is judged serious, and p_1 is the lowest level for which a wrong-accept decision is judged serious. α and β are the risks accepted for these two decisions.

(after Ferguson, 1971 a)



H_0: p=.85 (Student has sufficient proficiency, omit instruction)

H_1: p=.60 (Student does not have sufficient proficiency, give instruction)

C. Two-stage tests.

72. Now we will return to look more closely at the two-stage testing procedures already introduced at paragraph 62.

73. In a large scale empirical study Angoff & Huddleston (1958) compared two-stage college entrance tests in verbal and mathematical aptitude to conventional tests. In both subjects a routing test "directed" pupils to one of two measurement tests. In fact, using a sample of 6,000 pupils all possible combinations of measurement and routing test were administered so that a subsample necessarily took the appropriate measurement tests as if routed. The measurement tests were more reliable than the conventional tests, and showed slightly higher predictive validity against grade point average. The routing procedure made some 20% of routing misclassifications. The technical superiority of the two-stage procedures was not considered sufficient to offset the administrative difficulties that would arise.

74. In research already referred to for its use of sequential item sampling for a routing test Cleary et al (1968 a) also experimented with three other routing tests. All routing was to one of four 20-item measurement tests. This study was a real-data simulation using responses of several thousand 11th grade pupils to 190 multiple-choice verbal items. The three routing methods were:-

1. Double routing: A 10-item initial test was composed of items of about 50% difficulty level. Scores on this test were used to divide the sample into two approximately equal groups who went on to two separate 10-item tests similarly constructed in relation to their own groups. A further split then directed testees to the four measurement tests.

2. Broad range routing: A 20-item routing test having a rectangular distribution of item difficulties (as illustrated in Figure 1) divided the sample into approximate quarters based on the 20-item score.

3. Group-discrimination routing: The total sample was divided into approximate quarters on the 190-item total score. Item difficulties were then evaluated within each of the four groups. The 20 items with the largest difficulty range between top and bottom quarters were then selected for the routing test. Allocation to measurement test was on the 20-item score.

75. The last approach is interesting in that it explicitly recognises in a small way that tailored testing may require other item parameters than are appropriate for conventional test construction: in this case item difficulty by a coarse ability grading was used rather than over-all group difficulty (an approach not unknown in conventional work but less common). That the different approaches give different results is shown by the following details. The 20 items selected for the group-discrimination and broad range routing tests had only six items in common. The sequential item sampling routing test (described earlier), made up of the 23 items having the highest point-biserial correlations with total test score, had only 10 items in common with the group-discrimination test.

76. For classifying the total sample into quarters compared with the "true" 190-item classification, group-discrimination routing (29% misclassifications) was clearly superior to broad range (39%) or double (41%) routing. Sequential routing (27%) did slightly better. In terms of reproducibility of the 190-item score similar relativities obtained between correlations of this score and the four two-stage approaches. However, only the sequential approach was as effective as a 40-item conventional test.

77. In Linn et al's (1969) follow-on research using external criteria (paragraph 67 above) the group-discrimination approach was superior to the other approaches in allowing prediction of these criteria, and much superior to a conventional test of the same length. A conventional test three times the length would be needed to give comparable results.

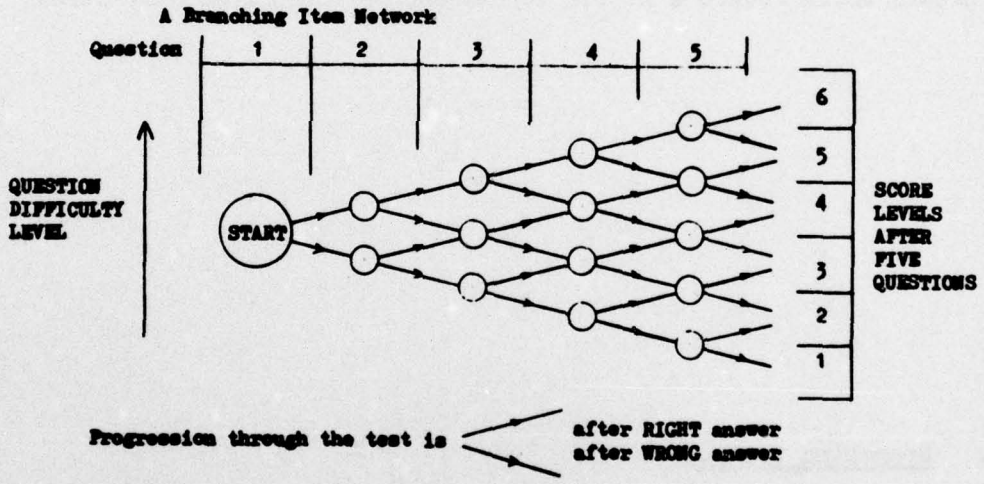
---

78. The two-stage testing research reviewed is generally encouraging. One would like to see Linn et al's (1969) favourable results confirmed in empirical studies before accepting the absolute size of the advantage. It may be significant that the most favourable result was achieved by the approach which looked a little beyond conventional item statistics.

D. Branching tests.

79. Next are the earlier fixed-step up-and-down sequential procedures. These procedures form an evolutionary line which continues through into Section 7 of this review. The archetypal procedure is based on a branching network of pathways through a fixed lattice of questions. An example is illustrated in Figure 5, but there are many variations. All testees begin with the same START question, usually of middle difficulty, and move through the network along routes which depend on their performance on successive items. Any testee will be steered through only five of the fifteen questions in the network. Referring to Figure 5, the more able testee will tend to get his initial questions right but after branching upwards to questions of greater difficulty he will find a better match. "Fixed-step" refers to the constant difference between neighbouring difficulty levels; "up-and-down" refers to the method of steering. In a more extensive network than the 5-stage plan of Figure 5 it will only be extreme testees who by the end of their test have not been encountering items approximately matched to their ability. In the later stages of such tests the answers of most

FIGURE 5 An example of a branching test.



testees can be expected to show a rough balance between wrong and right. In this way the test taken is tailored to suit the individual testee. Such tests are variously referred to by later researchers as branching, programmed, or pyramidal tests: an earlier term was sequential item tests, but to avoid confusion with sequential analysis methods this term is not used below.

---

80. Krathwohl & Huyser (1956) were the first to employ a branching test of this kind. Interestingly they had been looking at a sequential analysis approach (which we have already seen achieved earlier adoption) for switching testees from one block of questions to another. However they came round to the automatic-routing design of the branching test. Their thinking, of course, had a pencil-and-paper context in mind where sequential analysis sets administrative problems. First they used a real-data simulation from a 60-item college-level ability test. The test had 5-option multiple-choice items and Krathwohl & Huyser distinguished not only right and wrong answers but also better and poorer wrong answers. Their branching test had three exit paths from each item rather than the two of Figure 5. Because guessing in the conventional test data base seemed to be raising branching test scores unduly a new design was tried for further simulation. This design had two items at each node in the branching network - a block design in bioassay terms. Again there were three exits from each node, depending this time on whether two, one, or none of the items there were correctly answered. Correlations of about 0.77 with total score were obtained by a three-stage branching test of this kind which considered only six of a student's 60 responses.

81. Subsequently Krathwohl & Huyser tried an empirical pencil-and-paper implementation of their scheme; the most important outcome of this trial being that they ran into considerable practical difficulties in test administration.

---

82. The United States Army took up research on branching tests primarily with the aim of finding shorter tests. This is reported in a number of studies from 1960 onwards. Bayroff et al (1960) and Seeley et al (1962) constructed four 6-stage branching tests to a modified Krathwohl and Huyser design. Then in an empirical study they tried out pencil-and-paper implementations of two tests - verbal and arithmetic reasoning. The branching tests were administered to 327 enlisted men. Despite finding that the tests were too easy (no suggestion of motivational causes was made for the high scoring) correlations of 0.68 and 0.74 respectively were found for the 6-item branching tests with independent parallel 50 and 40 item conventional tests. On the other hand Seeley et al (1962) also concluded (p. 7),

> ".... it became apparent that the SIT/the branching test/ possessed some characteristics not entirely advantageous in terms of intended Army use".

83. They went on to detail these as follows:-

1. The branching test was more costly and time consuming to construct.
2. Administration of the 6-item branching tests was lengthy. For the two tests 10 to 15 minutes of initial instruction were required as well as the 15 minutes allowed for test completion.
3. Scoring presented problems as a testee's self-routing through the branching test had to be checked.
4. The instructions for the branching test were not understood by substantial proportions of men. Overall 9% of the verbal and 21% of the arithmetic reasoning test records were not scorable (note that the arithmetic reasoning test was attempted second within the single 15 minute time limit).

84. As might be expected the proportion of not-scorable records was related to performance on the Armed Forces Qualification Test. Men in Mental Category IV (10th - 30th percentile) had the highest proportion of not-scorable records.

85. The researchers suggested that further experimentation with branching tests in this form was not worthwhile, but that the basic concept may have considerable utility for presentation using a testing machine. Bayroff (1964) reported a feasibility study for a programmed testing machine but this was not then built although reported as within the state of the art. (However, it will be seen in Section 7 that Bayroff et al (1974) do develop a programmed testing system).

---

86. Leaving the US Army studies temporarily, Paterson (1962) had explored widely at a more abstract level. He used a computer-assisted Monte Carlo simulation applied to 6-item conventional and branching tests. The limitation to such short tests was imposed by his computing facilities. Within his branching test he placed the most discriminating items first within their difficulty level. He departed from a fixed difficulty step between items by allowing higher item discriminations to call for a larger step in difficulty level in the choice of the subsequent item. He also studied the influence of item discrimination, and of the shape of the ability distribution assumed. He found that his branching method gave more precise ability estimates for more extreme levels of ability, but that overall there was little to choose in precision against the conventional test. The branching test results reflected non-normal ability distributions more sensitively. Errors in estimating the item statistics were found not to be critical.

---

87. For the US Army Waters (1964) carried out a theoretical study comparing 5-item branching and conventional tests. She assumed a normal distribution of underlying ability and normal ogive item characteristic curves. For both open-ended and multiple-choice questions she showed that branching test scores correlated more highly with underlying ability than the best of various conventional tests. The difference was small - of the order of 0.03 (open-ended) and 0.01 (multiple-choice) on coefficients around 0.8. Whether this advantage would increase with more extended branching tests and at what test length (if any) such advantage would dissipate were unanswered questions.

88. Bayroff and Seeley (1967) in a further empirical study of branching tests - but now with computer assistance - administered 8-stage branching tests of verbal and arithmetic reasoning abilities to 102 enlisted men. (The most able testees also went on to a 9th item). This was possibly the first example of a computer-assisted tailored test. Test items were presented to individual men using on-line teletype computer terminals. Responses were made on the keyboard - the items were multiple-choice so that only option identification was called for. Correlations of branching test scores with independent 50-item verbal and 40-item arithmetic reasoning conventional tests were 0.78 and 0.74. Short conventional tests would need to be twice the length of the 8-item branching test to achieve comparable results.

---

89. In the British Army McGill (1968) reports the construction of a 10-stage branching test under the supervision of K D Duncan. The test was constructed from the multiple-choice items of a predominantly mechanical aptitude conventional test. The most and least able testees were provided with further stages beyond the tenth. A real-data simulation from recruit response data showed close agreement with corresponding 60-item parent test scores. Further work then followed to produce a manageable pencil-and-paper format for empirical study. A technique that seemed to offer promising simplicity was one using an answer sheet over an embossed card so that embossed numbers would appear on shading a chosen answer space with a soft pencil (following Duncan (1964)). The number which appeared directed the testee to his next question. Small scale partial trials were reported to be successful.

90. Hansen (1968) carried out two empirical studies of branching tests presented by online teletype. His subjects were university freshmen taking a physics course examination. In his first study 56 freshmen took five topic-centred 3- and 4- stage branching tests, 17 items were attempted in all. Hansen also explored a variety of scoring methods. So far in this review only the straightforward scoring scheme illustrated in Figure 5 has been introduced for branching tests. There are other possibilities and these are discussed in Section 7 in relation to more recent work. Generally the scoring methods intercorrelate highly - the four methods used by Hansen had intercorrelations from 0.84 to 0.94. The validity of the four scoring methods for predicting final course grade ranged from 0.38 to 0.49 - not high values but all higher than achieved by a 20-item

conventional classroom test also taken by all students. A second study, also small (30 freshmen), is of interest because after the teletype test sessions the students completed an attitudinal scale about computer-based testing as they had experienced it. Generally their ratings were favourable. Guessing was reported as happening very seldom. Disappointingly perhaps for putative motivational benefit students reported that they were relatively unaware of the efforts to individualise the test material, but comparative results against conventional testing on material considered less suitable are not available.

91. In the real-data simulation study by Linn et al (1969) referred to previously, two branching tests were included in the methods tried. One branching test was a normal 10-stage network, but with a weighted scoring system in which more difficult questions had higher scoring weights. The second test was to a block design. A block of five verbal items occupied each node in a 5-stage network. Testees thus attempted 25 items. The five items at a node were closely similar in difficulty. Branching from a node depended on whether two-or-less or three-or-more of the five items were answered correctly. Again a weighted scoring system was used. For equal success in predicting an external test criterion it was found that conventional tests would need to be 1.65 and 1.76 times as long as the two branching forms respectively.

92. Finally, in the pre-1970 branching test studies, Wood (1969) made up three tests of four, five and six stages on CSE mathematics topics. These were administered in an empirical study to 91 CSE candidates. The method of presentation was an improvised pencil-and-paper technique using self-adhesive labels. Wood experienced about 5% of spoiled papers. The correlations between the summed branching test scores (15 items in all) and subsequent CSE grade was 0.51 - compared with an almost identical value, 0.52, found for a short conventional test composed of the 15 best items.

---

93. So far the research on branching tests has shown persistent glimpses of possible benefits among a variety of cautionary results. Branching tests have in some instances, and often by small margins, nudged in front of equal length conventional tests in their relationship to underlying ability, their validity, their precision of estimate for non-average levels of ability, and in their reproduction of independent conventional test scores. Such encouragement proved at least sufficient to sustain the converted researchers.

94. A number of empirical studies have been reviewed, all on smallish samples for obvious reasons. The later pencil-and-paper formats go some way towards relieving the despair of the first proponents and appear usable in some applications. The online use of computer terminals resolves the administrative problems most satisfactorily. All the terminals used have been teletypes, the relative slowness and noise of which have not attracted any adverse comment.

95. Research so far has been concerned mostly with short branching tests, 10 items or less. This restriction arises partly because this Section is dealing with early research, but it also partly reflects a deliberate search for short test forms. A short-test viewpoint may well emerge as a rather blinkered perception of the possibilities. Tailored testing can be more than an abbreviated 'substitute,' and a wider regard is more appropriate.

96. A heavy reliance on correlational methods of evaluating tailored testing approaches (and evident in this Section) was criticised earlier from the stand-point that the mere reproducibility of other estimates is an insufficient criterion. Correlation evaluations have also been criticised (Lord (1970 b), Wood (1969)) on the grounds that

- the correlation coefficient is a group statistic while for an individualised method of testing the focus should be on individual accuracy
- the value of a correlation coefficient is dependent upon the distribution of the characteristic in the particular group.

Consequently although it is entirely appropriate to look for predictive validity in a tailored test estimate the force of this criticism is that validation necessitates looking beyond a group correlation. The matter of proper evaluation is important and is taken up again in Section 7.

7. RECENT TAILORED TESTING

97. The research reviewed in this section is that which, with a few exceptions, has been reported from 1970 onwards. As compared with the work in Section 6 the more recent research is characterised by greater sophistication of theory and equipment, and by growing coherence: there is also more of it - there having been more work reported in this period than in all the years before. This Section draws the research together under the following headings:-

A. General - this part introduces a number of concepts and approaches which are generally helpful and are used thereafter.

Further research on procedures already met

B. Short tests
C. Two-stage testing
D. Branching tests

New procedures

E. Flexilevel tests
F. Item-finding procedures
G. Stradaptive and broad range approaches

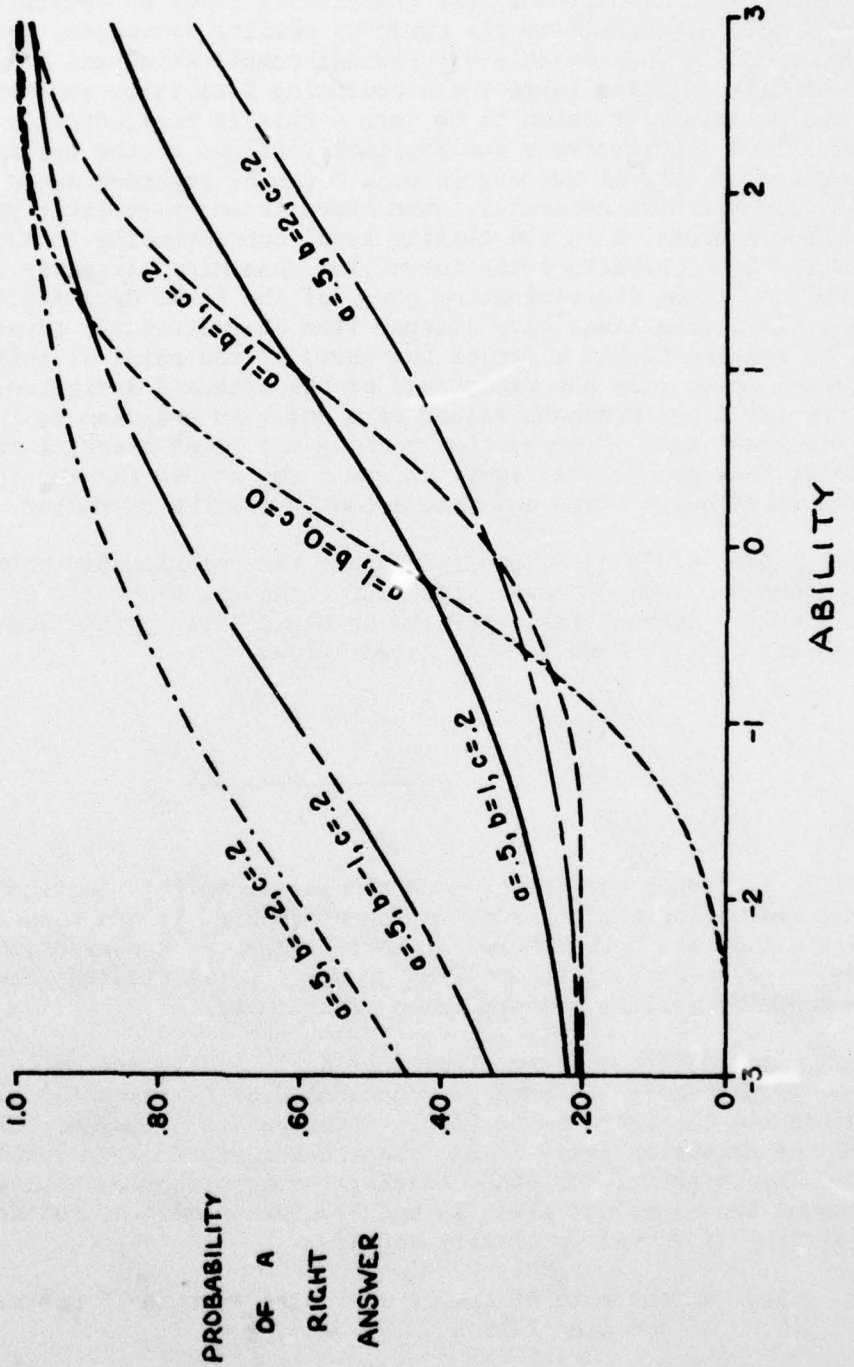
The testing strategies to be reviewed in parts E and G are of greatest relevance to the Project research plans.

A. General

98. A number of researchers base their approach on latent trait mental test theory, or item characteristic curve theory as it is perhaps more descriptively also known. Figure 6 illustrates a number of item characteristic curves. Each curve represents the probability of success on a particular test question in relation to ability level. The basic theory assumes the curves to be normal ogive or alternatively logistic functions and is given in Lord (1952), Birnbaum (1968), and Lord & Novick (1968). In terms of numerical outcomes the choice between the alternative functions is of little consequence. The fit of the models to test item data has been evaluated in a number (but not a large number) of studies (for example, Lord, (1970 a)) with positive results.

FIGURE 6 Examples of item characteristic curves.

(after Lord, 1969)



99. Each item characteristic curve is specified by three parameters, a, b, and c together with the function assumed. Parameter c is the probability of chance success on a question: in multiple-choice questions c is often taken to be the reciprocal of the number of options, although this is a questionable assumption; more empirically c may be estimated from the asymptote approached by the curve as ability decreases. The parameter c represents a considerable theoretical complication and its estimation for real-data requires large scale computing facilities and even so difficulties remain. When c is taken to be zero - this is realistically so for open-ended questions - parameter b can be simply defined as the ability level for which the probability of success is 0.5. b can be regarded as an index of item difficulty. More generally, when there is an appreciable probability of chance success, b is the ability level corresponding to the point of inflexion on the item characteristic curve. The remaining parameter a, can be taken to represent the discriminating power of the item. Graphically the more discriminating items have steeper item characteristic curves. Parameter a is related to the slope of the curve at the point of inflexion. (For the normal ogive a is the reciprocal of the standard deviation.) Figure 7, from Urry (1971 b), presents values of a and b in relation to the conventional item statistics of proportion passing and point-biserial correlation with total test score: the figure is for c set at 0.2 (as may apply for a 5-option multiple-choice item) and this accounts for the asymmetry.

100. Lord (1974 a) helpfully reviews the relationship between tailored testing and item characteristic curve theory. Generally the theory offers a useful framework for real-data or Monte Carlo simulations and several studies of this kind are described below.

---

101. A further concept of general utility in this Section has to do with the evaluation of tailored testing procedure. It has been seen already that evaluation may be furthered by correlations with conventional measures, and, in the case of theoretical or Monte Carlo studies, correlations with underlying ability and precision of estimate.

102. An additional form of evaluation is by the use of information functions, and in particular a function recommended by Birnbaum (1968) and Lord (1952). Referring to Figure 8 for illustration we are concerned there with the ability of the measuring scale to distinguish the two ability levels A1 and A2. The Figure shows, for these levels, the distribution of measurement errors around the expected values X1 and X2. The success of the scale in distinguishing A1 from A2 is clearly dependent

(i) on the rate of change of X with A; that is the slope of the line P1P2

& (ii) inversely on the dispersion of the error distributions.

FIGURE 7

The relationship between item characteristic curve parameters a and b and conventional item statistics when the probability of chance success is 0·2. (after Urry, 1971 b)

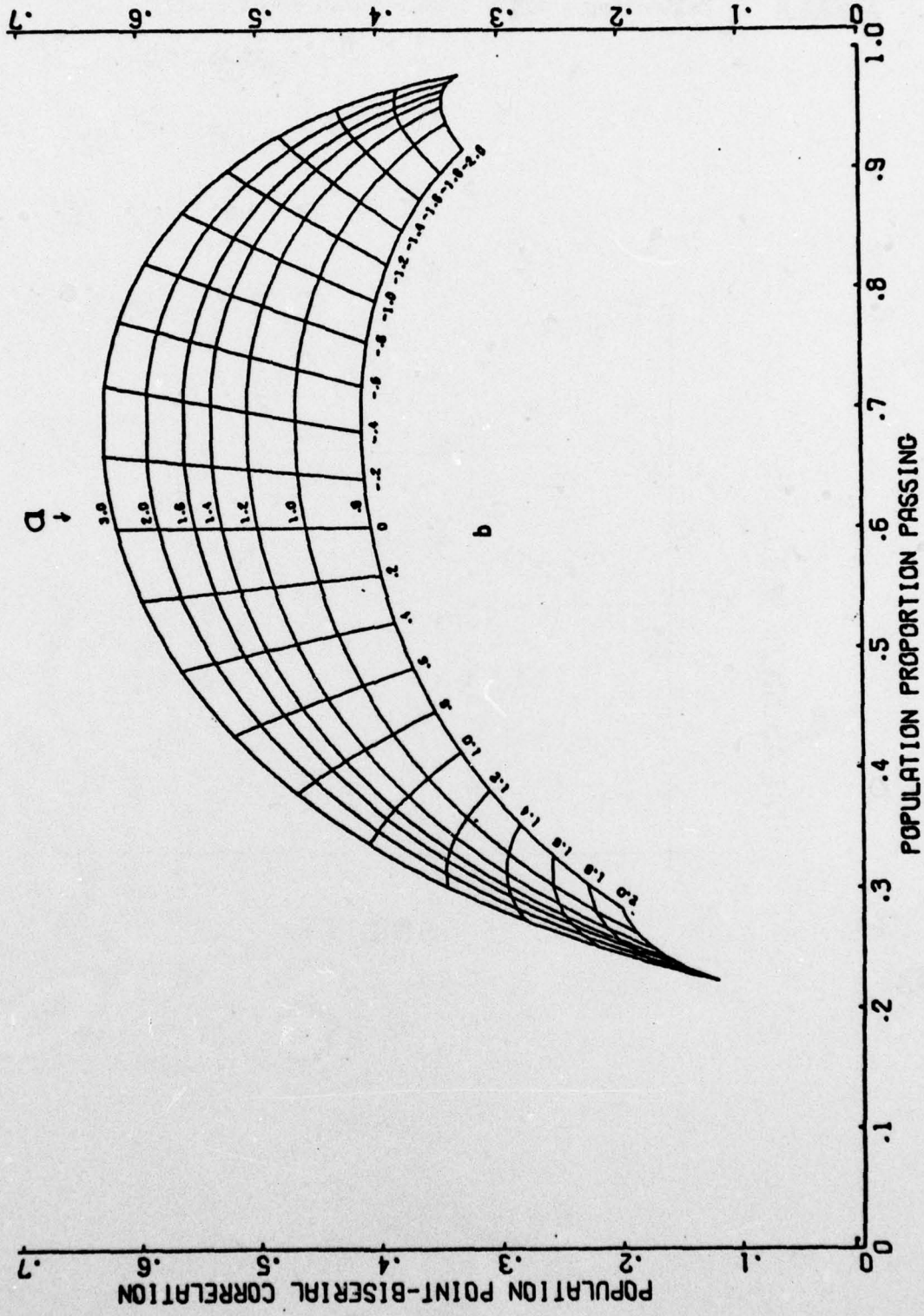
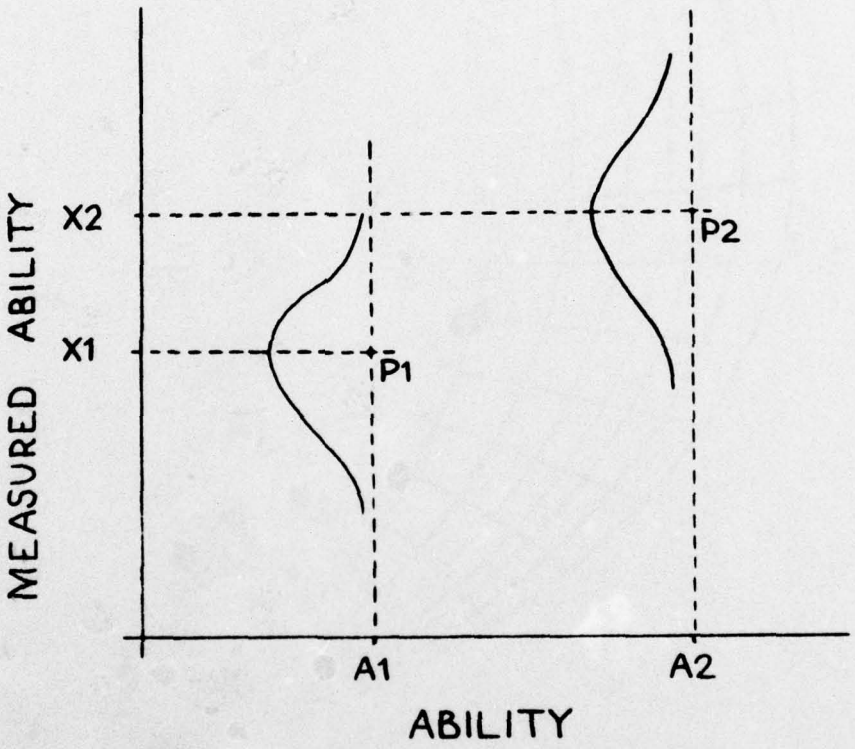


FIGURE 8 Estimating ability from a measurement scale.



103. The information function of Birnbaum and Lord is defined as the square of the ratio i/ii. Two other related interpretations of this information function have also been made which are helpful in appreciating its characteristics. An increase in the information function achieved by a modified test design is the equivalent for a conventional test of a proportionate increase in test length. Also the information function, or more precisely its square root, is inversely proportional to the confidence interval for estimating ability level from test score. Where an information function is subsequently referred to it is this function.

104. Generally the information function is most usefully employed in comparing two tests by looking at the ratio of their information functions for different levels of ability. This ratio is termed the relative efficiency of the two methods and has the advantage, in this ratio form, of being invariant in relation to the idiosyncracies of the ability scale incorporated in the information functions. It follows that an absolute interpretation of an information function may be misleading. This danger is underlined by Lord (1975 a) who points to certain deficiences in the ability scale normally used in item characteristic curve theory.

---

105. In these general preparatory remarks on recent research the writer would also like to point to valuable reviews by Weiss & Betz (1973 a) and Wood (1973). Their preferred terms for tailored testing are adaptive testing and response-contingent testing. Weiss (1974) also presents a useful comparative commentary on the various approaches tried for tailored testing.

B. Short tests

106. The few studies reviewed here continue the predominantly military concern of producing shorter tests. This abbreviation may be attempted by any means; individualised testing procedures are but one line of attack.

107. Bryson (1971 and 1972) looked at four methods of producing 5- or 6-item tests. Initially she used real-data simulation based on a response bank from 10,000 men in recruit training at a Naval Training Centre. Responses to two tests were used, the Navy General Classification Test and

the Navy Mechanical Aptitude Test. One of the four methods was a shrinking-step individualised branching procedure referred to as BRANCH. In BRANCH the question with the highest internal validity is first used to split the total group. Internal validities for the remaining questions are then re-computed separately for the two groups. For each group the question with the highest validity for that group is then used to make a further split. Thus the procedure routes testees through a series of forks so that after five questions there are 32 exit points.

108. Thus a major characteristic of BRANCH is that question selection is based on a question's local (rather than total group) characteristics for a sub-group of a narrower range of ability. This seems vital to tailored testing. Procedures based on total group item statistics only make sense in so far as these statistics offer approximations to the performance of the items for people of more homogeneous ability. Essentially tailored testing treats people differentially in relation to ability. It will be preferable to avoid the approximation from total group statistics (and the assumptions inherent therein) and to work directly with item indices related to ability levels.

109. A critical disadvantage of BRANCH is that it offers no recovery route after an incorrect forking decision.

110. Compared with the other methods BRANCH was most successful in reproducing total test score. For the general classification test (the more internally consistent of the two test used) correlations with total score for the four short test methods ranged from 0.86 to 0.94 for 5-item tests. For the mechanical aptitude test the range was 0.69 to 0.82.

111. Bryson (1971) went on to give empirical trials to the four short test methods. In these trials BRANCH tests were administered by online VDU terminal to 263 recruits. Each question was given with a separate (55 second) time limit. Under these conditions BRANCH was no better than the best of the other methods in reproducing total score. Bryson points to the original choice of BRANCH questions being based on item characteristics which for later items would be influenced by time pressures not present in the BRANCH VDU presentation. This is a likely factor and emphasises the importance of realism in any response base used for simulation.

112. Only in the context of short tests is correlation with total test score more than a start in the evaluation of alternative procedures.

c. Two-stage testing.

113. Two-stage tests are a marginal form of individualised procedures. Generally a two-stage test will offer perhaps only three to six alternative diets of questions. The saving virtue of a two-stage test may be that because pencil-and-paper implementation is possible it does offer a realisable prospect of large scale testing. Whereas large employers, such as the Army, maintain continuous recruitment so that online testing can be achieved with a relatively small number of computer terminals, other settings - notably educational examinations - may demand the capacity to test large numbers of people simultaneously.

114. At a theoretical level Lord (1971 c) used item characteristic curve theory to investigate nearly 200 two-stage designs. He assumed a normal ogive characteristic curve and also equal discriminating power (constant value of parameter a) for his items. Largely he worked with an overall limit of 60 items for routing and measurement tests combined. He considered both no-guessing (c=0) and with guessing (c=0.2) conditions.

115. His basis for comparison was a 60-item peaked conventional test. By peaked he means a test in which all items are of identical difficulty. A test peaked at ability level A would be such that the probability of someone of ability A answering any one question correctly would be 0.5 (excluding chance success). Hence a peaked test differs from most conventional tests in regular use: such tests albeit geared to specified populations typically have a spread of item difficulty. The routing and measurement tests of his two-stage designs were also taken to be peaked at appropriate ability levels. A "best" up-and-down branching test of equal length provided another basis for comparison.

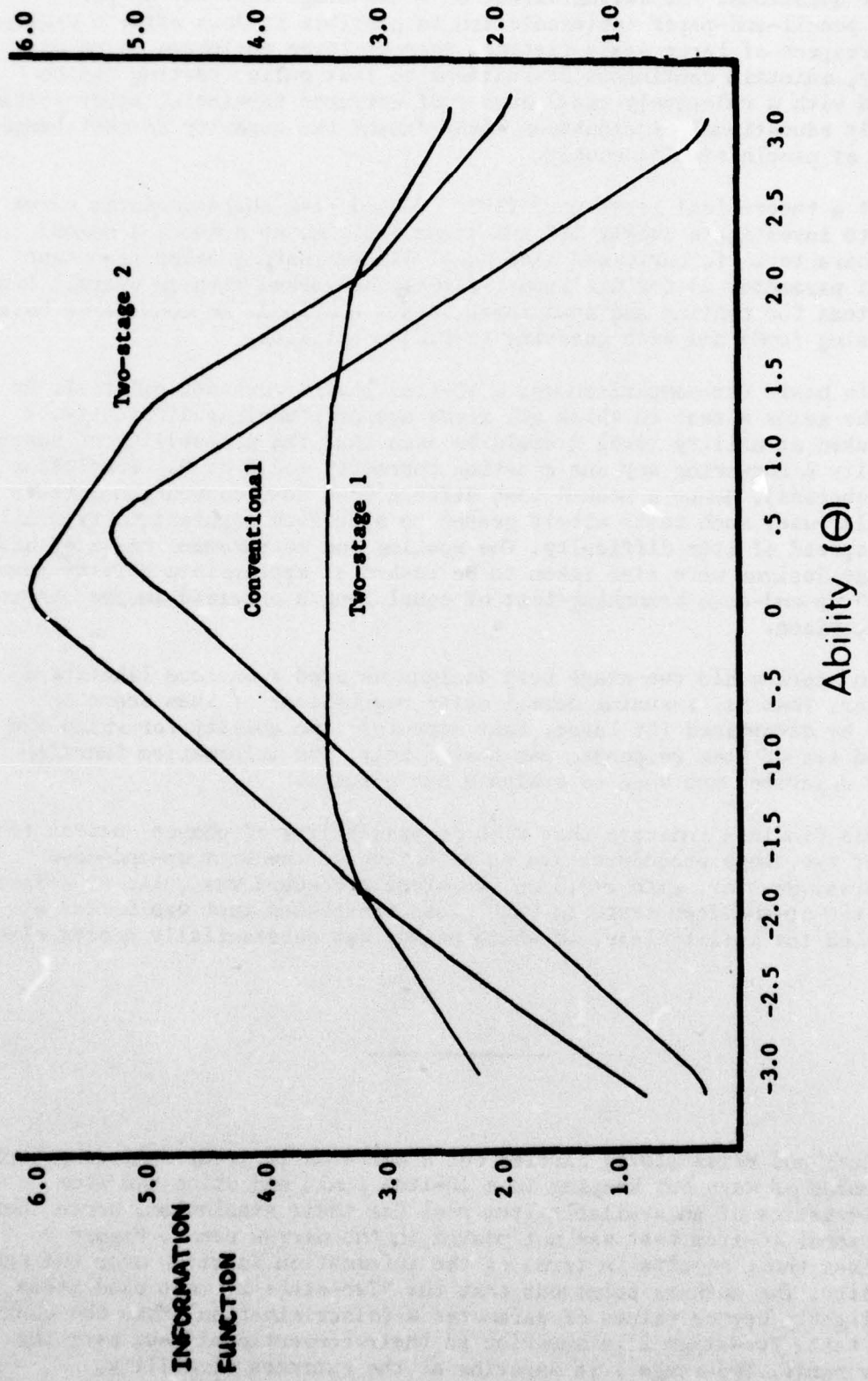
116. In scoring his two-stage test designs he used a maximum likelihood estimator. That is, assuming normal ogive regressions of item score on ability he determined (by large, fast computer) the ability for which the observed set of item responses was most likely. The information function already described was used to evaluate his results.

117. His findings indicate that with no possibility of chance success (c=0) the best two-stage procedures are as effective as the best up-and-down procedures. However, with c=0.2 no two-stage procedure was quite as effective as the up-and-down test. In both cases the peaked test was better at and around the ability level at which peaked but substantially poorer elsewhere.

---

118. Betz and Weiss (1974) carried out a Monte Carlo study following Lord's in a number of ways but keeping to a 40-item limit and using the item characteristics of an available item pool for their simulation. Hence their conventional 40-item test was not peaked in the narrow sense. Figure 9 summarises their results in terms of the information function over the range of ability. The authors point out that the "Two-stage 2" test used items with slightly better values of parameter a (discrimination) than the conventional test. Two-stage 2 is superior to their conventional test over the ability range. Two-stage 1 is superior at the extremes of ability.

FIGURE 9 A comparison of simulated two-stage and pseudo-peaked conventional tests. (Results from Betz & Weiss, 1974)



119. In an earlier study using the Two-stage 1 design Betz and Weiss (1973) had carried out the first computer-administered empirical study of two-stage testing. 214 psychology students were tested using an online VDU terminal. Difficulties had been encountered with both the measurement tests and the cutting scores on the routing test that determined allocation to measurement test. These difficulties can be attributed to their use of total group item statistics. Two-stage 2 resulted from modifications to this first design.

120. From the early '70s a research group led by D J Weiss has worked on adaptive testing - to use their term - at the University of Minnesota and further references will be made to their work. In particular the group has started on a programme of empirical trials of various approaches to tailored testing using VDU online terminals. Empirical work in tailored testing remains rare and the experience of the Minnesota group, albeit confined to psychology students, often provides pioneering information on the topics covered.

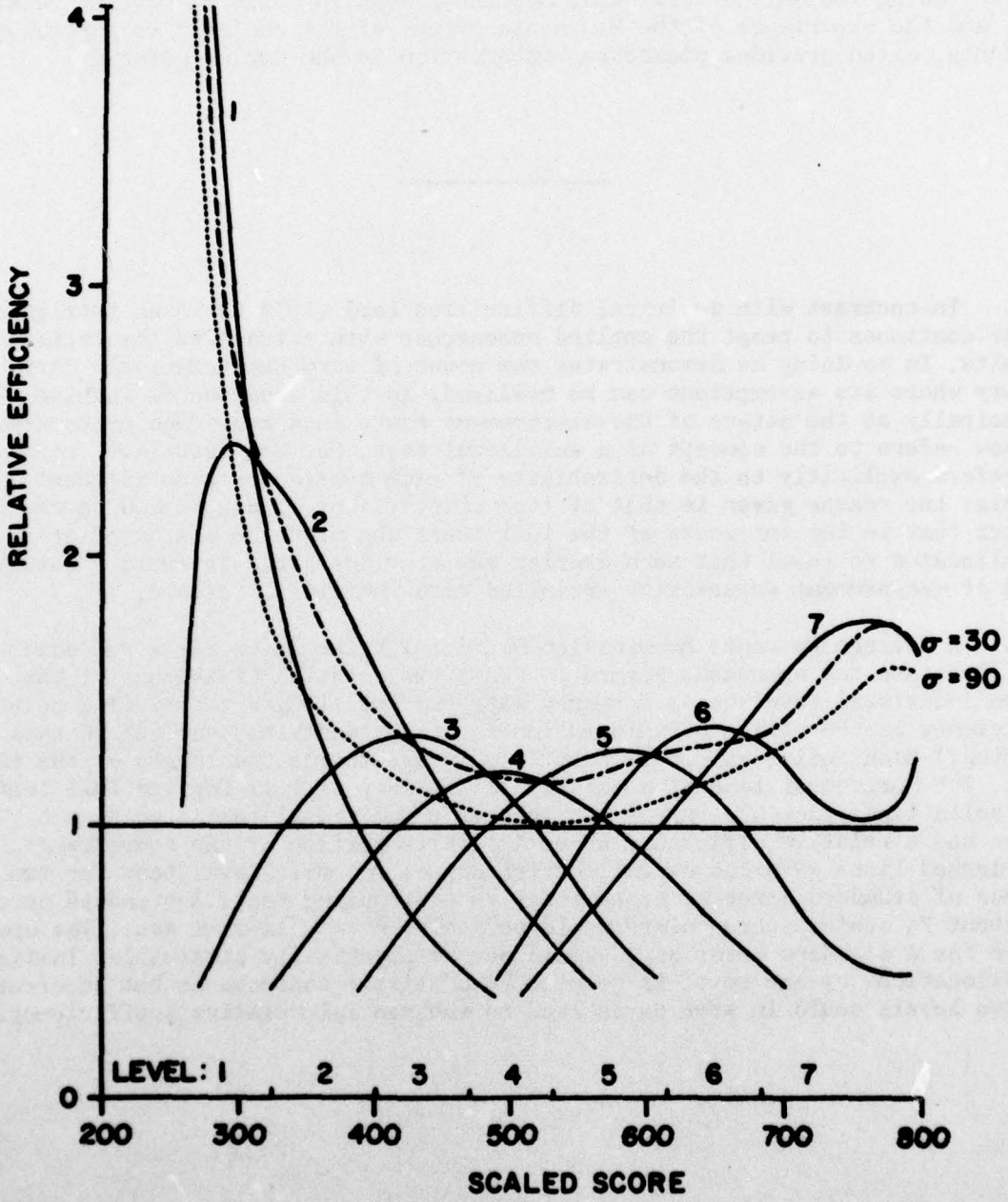
---

121. In contrast with empirical difficulties Lord (1974 b) in an intriguing paper continues to tempt the applied researcher with attractive theoretical results. In so doing he demonstrates the power of item characteristic curve theory where its assumptions can be realised. In this study he is looking principally at the nature of the measurement tests in a two-stage procedure. He now refers to the concept as a multilevel test. For his multilevel test he refers explicitly to the desirability of item overlap between adjacent levels: the reason given is that of item economy, but it has seemed to the writer that in the interests of the individual who might be misrouted or misallocated to level that such overlap was very desirable to avoid a patchy kind of measurement superiority sprinkled with individual failure.

122. A College Entrance Examination Board SAT Mathematics paper was used to illustrate the approach. Figure 10 shows the relative efficiency of the seven individual level tests compared with the full length tests. (The relative efficiency is the ratio of information functions described earlier in this Section.) Each individual level test is only two-thirds the length of the full test. The horizontal line at a relative efficiency of 1 is for the full test. The solid lines plot the curves for the seven individual tests: each such curve has a relative efficiency above 1 for its portion of the score range. The dashed lines give the overall efficiency of the multilevel test for two values of standard error of measurement in the routing test. A standard error of about 75 scaled score points would be achieved by a 12-item test. The upper curve for a standard error of 30 would not be practically attainable. Initial misallocations by one level is seen to be of little consequence but an error of two levels could in some cases lead to substantial relative inefficiency.

FIGURE 10 Theoretical results on an SAT Mathematics paper comparing the relative efficiency of a two-stage multilevel approach and the conventional test. The dashed curves are the overall efficiency of the seven (solid line) local tests for two levels of standard error (σ).

(from Lord, 1974 b)



D. Branching tests

123. The basic concepts of branching tests have already been introduced and an illustrative scheme was given at Figure 5. Probably more work has been done on such schemes than on any other individualised approach. This approach has the critical disadvantage of requiring a very prescribed item pool. Items are needed to fit the nodal points where routes meet and diverge. A branching test cannot be used until every node has an item of approximate fit. Neither are the item requirements negligible: $n/n + 1)/2$ items are required for an n-item branching test - 120 items for a 15-item test. The test constructor would be considerably dismayed at the thought of how many items would need to be written to obtain the 120 to match the specification. Additionally the item specifications are made in terms of conventional total group statistics which at best can only be an approximate indication of performance for the relatively narrow ability band of testees encountering any one item.

124. Several theoretical studies have been carried out based on item characteristic curve theory. Often for simplicity fixed values are assumed for discriminating power and probability of chance success (parameters a and c). For multiple-choice items the questionable assumption of random guessing is usually made.

125. Lord (1970 b) is an influential foundation paper, the outcome of which is somewhat pessimistic for tailored testing. The pessimism is attributable to the low value of parameter a which he largely assumes (a value of 0.5 - corresponding to a point biserial less than 0.3, see Figure 7) and the constraint of a test of fixed length which he works within. (Green (1970) provides a healthy counterblast keeping variable test length in mind as illustrated in Figure 3). The variables Lord investigated include,

(i) up-and-down step size, that is the fixed difference in difficulty (parameter b) between adjacent questions.

(ii) the value of a smaller up than down step where a probability of chance success exists. This is referred to as offset.

(iii) the method of scoring. Some possibilities would be

- the average difficulty of items attempted, excluding the first (as common to everyone) but including a notional (n+1)th item that depends on performance on the nth, final, item

- the final difficulty level, that is of the (n+1)th item as in Figure 5

- the conventional number-right score.

(iv) the effect of chance success.

(v) the value of Robbins-Monro shrinking step procedures (introduced earlier, in Section 4).

126. Figure 11 illustrates some typical results. The tailored tests (solid curves) are more effective at the ability extremes, the peaked tests (dashed curves) more at the central ability at which they are aimed. The probability of chance success depresses the information function and leads to asymmetry in both curves. Peaked tests are idealised fictions; the semi- or pseudo-peaked test in applied existence would have a curve of intermediate shape which might or might not top the tailored test curve for central ability.

A paraphrase of Lord's further conclusions is,

(i) the number-right score is perfectly correlated with the final difficulty score.

(ii) in terms of the information function the average difficulty score provides better measurement (and this score is subsequently used).

(iii) for 60 items a step size of 0.4 (in the difficulty level parameter b) seems best, and for 10 items a step size of 1.0.

(iv) offset step sizes improve accuracy of measurement when there is chance success.

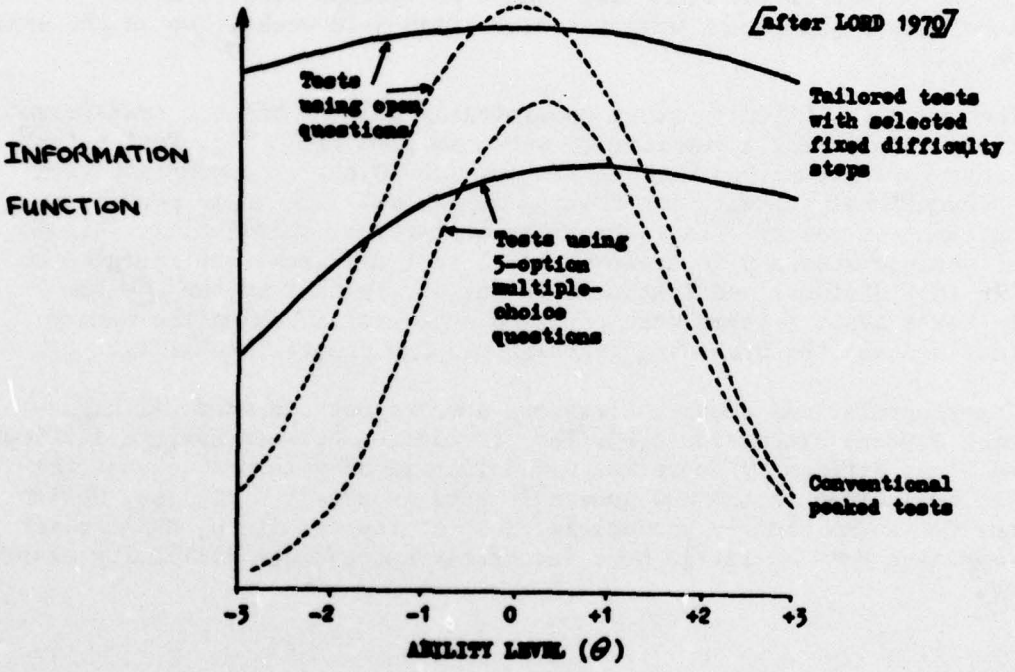
127. Stocking (1969) essentially followed Lord's study but for a 15-item branching test. Her conclusions also followed Lord's but she was also able to study Robbins-Monro shrinking step procedures more extensively. These were found to be marginally superior to the best fixed step procedures. However, for an n-item test the Robbins-Monro procedure calls for 2^{n-1} items - over 32,000 items for a 15-item test. Hybrid procedures were studied which attempted to capture some shrinking step advantages using a change in fixed step size, but the procedures tried failed to do so. Lord (1971 a) reaches the same conclusions for Robbins-Monro and hybrid procedures.

128. Mussio (1972) aimed to cut down the item requirements for a branching test by curtailing the item network at lower and upper difficulty limits. For example, a 60-item test restricted to 11 difficulty levels requires 605 items compared with 1830 for a full network. The penalty is some loss of precision at extreme abilities, but results remain superior to those for a conventional test.

129. Further theoretical work for the US Army by Waters & Bayroff (1971) had the particular merit of looking at the effect of varying item discrimination. They compared various 5-, 10-, and 15-item branching tests with various conventional tests of the same length. Scores were evaluated by their correlation with underlying ability (after Lord (1952)). For item discrimination at 0.6 or above (as assessed by biserial correlations with underlying ability, not with a fallible total score for which the equivalent values would be lower) the highest correlation was always for a branching test. For lower item discriminations a conventional test achieved equivalent results, while for the lowest biserial assumed, 0.3, a conventional test was superior. This latter result can perhaps be regarded as an indicator of a conventional test's robustness under conditions of misuse. All the observed differences were small, perhaps expectedly so for a global measure like the correlation coefficient. To round off the series of US Army studies it is appropriate to mention here the work reported by Bayroff, Ross and Fischl (1974). Here they describe an advanced online individualised test set-up - far more

FIGURE 11 Typical up-and-down branching test results for open-ended questions (c=0) and multiple-choice questions (c=0·2) compared with conventional peaked tests.

(Theoretical results after Lord, 1970 b)



sophisticated than the equipment visualised by Bayroff (1964). Cynically this appears a case of the electronic technology overtaking psychometric technique for they report no plans or decisions for the forms of individualised testing to be tried. Essentially this might be better taken as a comment that it is not yet clear that any form of tailored testing has established a convincing case.

---

130. *Finally on branching test research an all too rare empirical study* is reported by Larkin & Weiss (1974). They worked with multiple-choice vocabulary items. Three 15-item branching tests were used and a variety of scoring methods. Both the branching tests and a 40-item pseudo-peaked conventional test were administered by online VDU. Three groups of over 100 students each took one of two of the branching tests. Two groups also took the conventional test. All groups were retested after 5-10 weeks, two on the same branching test.

131. The average difficulty score consistently had the highest test-retest correlations (confirming a superiority shown in Lord (1970 b)). Test-retest coefficients for this method were of the order of 0.86. In comparison with 15-item conventional subtests the testing design was such as to permit the disentanglement of memory effects from the test-retest stabilities. This is important because whereas, in a conventional test all items are repeated on retest, in an individualised test this is not so. In fact in the 15-item branching tests about 8 items were repeated on *average. Taking the memory* effect into account the branching tests showed the greater stability.

132. Intercorrelations among the various scoring methods were all high - always over 0.9 and often over 0.95. The correlation between average difficulty score and final difficulty score was 0.91. This is of interest because the latter scoring method is the one generally used in pre-1970 studies, while the former now seems clearly preferable. Some of the results of the earlier studies may have been a little more favourable had average difficulty scoring been used.

---

133. The theoretical branching test studies consistently demonstrate superiority over a peaked test outside the central ability range; possibly for a pseudo-peaked test this could be so across the whole range. The relationship with underlying ability also tends to be a little closer. In an empirical study test-retest stability was a little higher. The tendency for the tailored approaches to nudge ahead is showing more consistently here. Notwithstanding this the writer anticipates that research on branching tests will tend to decline in favour of the newer methods to be described at F. and G. below.

E. Flexilevel tests

134. The flexilevel test is an ingenious attempt by Lord (1971 b) to produce a practicable pencil-and-paper procedure with the capacity for a limited degree of tailoring. As such it is peripheral to the planned research which assumes a more flexible technology. However, two empirical VDU administrations are reported in the literature and these will be described - in the view of the writer these attempts are misguided.

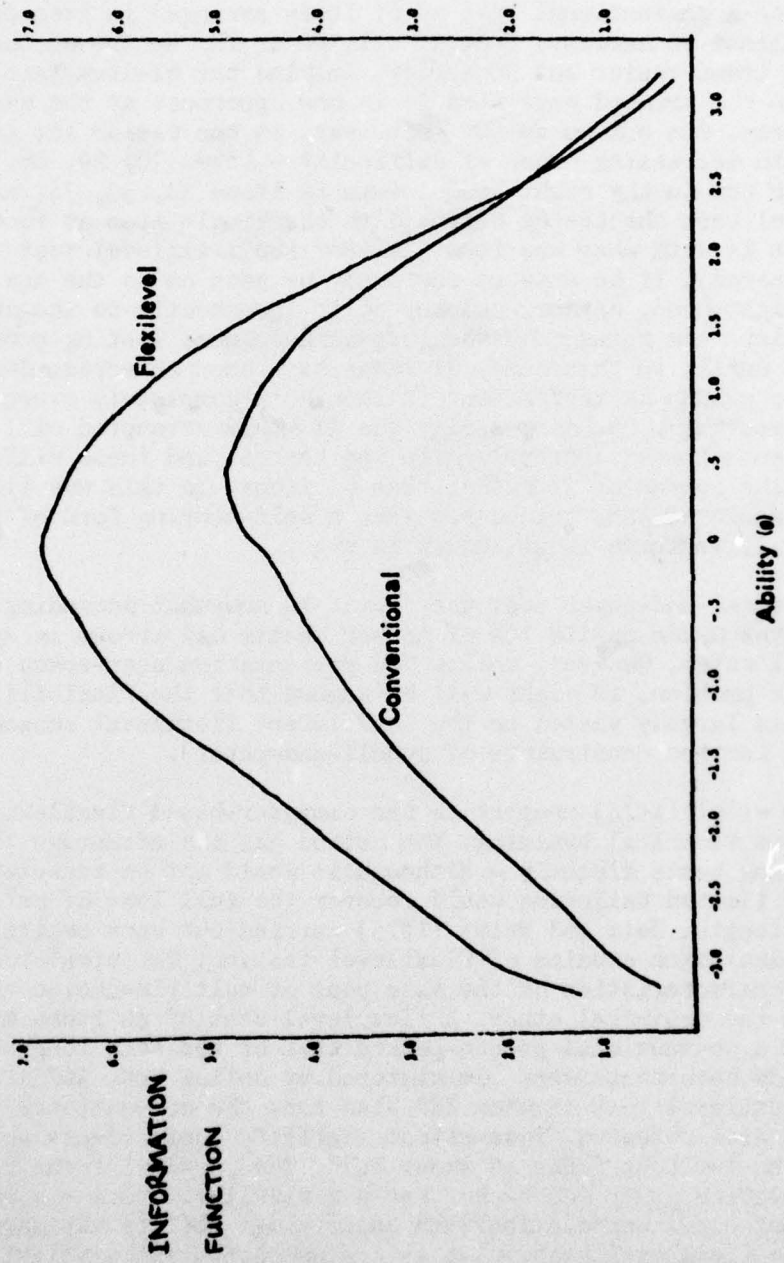
135. Consider a conventional test of 61 items arranged in item difficulty order from easiest to hardest. Item 31 will be at the centre of the difficulty order with 30 items easier and 30 harder. Imagine the 61-item test bent in two so that on the printed page item 31 is now uppermost at the head of two columns of items. One column on the left, say, is the easier set and will now be found in decreasing order of difficulty - items 30, 29, 28, and so on. The harder set in the right-hand column is items 32, 33, 34, and so on. In a flexilevel test the testee begins with the single item at the head of the page, that is with what was item 31. (For the flexilevel test the items will be renumbered). If he answers correctly he goes on to the next unattempted item in the right-hand, harder, column, or if incorrectly to the next available easier item going down the left-hand column. Testing proceeds following this rule until, in this case, 31 items have been answered. Switching from column to column is inefficient (although it completely overcomes the danger of misrouting), but necessarily the 31 items attempted will tend to include that subset most appropriate to the testee, and these will have been attempted in the course of 31 rather than 61 items. In this way limited tailoring is achieved. The method requires a self-scoring form of answer sheet which will indicate if an answer is right.

136. As a pencil-and-paper test the format is somewhat demanding. In a study with Eighth Grade pupils 10% of answer sheets had errors in applying the procedural rules. However, online VDU presentation over-comes the administrative problem. It might well be argued that the flexibility of online presentation is largely wasted on the inefficient flexilevel scheme (drawn up explicitly for the constraints of pencil-and-paper).

137. Hansen et al (1974) propose to use computer-based flexilevel testing in US Air Force technical training. The method has the advantage that it can use existing tests directly - although it would not be expected in this case that the limited tailoring would recover the full loss of reliability from reduced length. Betz and Weiss (1975) carried out both empirical and Monte Carlo simulation studies of flexilevel testing. The simulation was based on the characteristics of the same pool of multiple-choice vocabulary items used in the empirical study. A flexilevel test of 40 items was given together with a conventional pseudo-peaked test of the same length. In the empirical study both tests were administered by online VDU, 367 students taking the flexilevel test of whom 227 also took the conventional test. Some students were also retested. Test-retest stability coefficients were comparable for the two test forms at about 0.89. The parallel forms reliability from the simulatiom study was higher for the flexilevel test - a mean of 0.84 as against 0.80. Correlation with underlying ability was marginally higher for the flexilevel test, 0.91 as against 0.89. The simulation study was also able to look at the information function of the two test forms in relation to ability. Figure 12 summarizes some of the results. Being based on real item pools these curves are in substantial contrast to the peaked and flat crossing curves typical of conventional and tailored tests in theoretical studies (compare Figure 11). Some features of Figure 12 can be explained by some differences in discrimination in the items used for the two tests. Establishing exact comparability in empirical studies is very difficult.

FIGURE 12 A comparison of simulated flexilevel and conventional tests based on real item pools.

(from Betz & Weiss, 1975)



138. The simulation study here included a probability of chance success which was set at 0.2 because the multiple-choice questions had five options. The artificiality of this assumption is perhaps suggested by the correlation between flexilevel and conventional tests under empirical as compared with simulation conditions; this is 0.89 for the former (N=103) and 0.82 for the latter (N=10,000): however, there are also other factors making for consistency which affect real testees but not their simulations.

139. The stability coefficients from the empirical study (0.89) are the same as for the 40-item two-stage test of Betz and Weiss (1973) and only a little higher than for the much shorter 15-item branching test (0.86) using average difficulty scoring (Larkin & Weiss(1974)). The branching test can provide closer tailoring than the other two methods and the stability coefficients tend to confirm its greater efficiency.

Some intermediate comments

140. In the recent approaches to test individualisation a fairly consistent pattern of advantage (often small) over conventional testing has become apparent. The details of this advantage are confused by the difficulties of comparative empirical studies and by the simplifying assumptions necessary in theoretical work. Benefits in measurement precision at extreme abilities and a closer relationship to underlying ability would be conservative claims. However, all the methods are limited in the degree of tailoring they provide. In the cases of two-stage and flexilevel tests the limitation is their physical structure which limits the item pool and the number of possible routes. The limitation for a branching test is partly inherent in its item requirements, and the fact that in practice they will be imprecisely met, rather than in its physical form.

141. A concept seen as of especial relevance to tailored testing (although not to conventional testing) is what might be called resistance to anomaly, or maintenance of equilibrium. The tailoring process can be viewed as a control or steering mechanism. The target is questions of appropriate difficulty - matching testee ability. The adjustments that have to be made to a testee's route must be sensitive to his current performance, but not over-sensitive, or else the occurrence of anomalous responses will result in excessive reorientation with possible loss of direction and consequent need for recovery. In engineering a servo-mechanism to correct the mismatch between course and direction is subject to damping so that wild movement and oscillation are avoided. The tailoring process needs to have similar

damping to give it the required control characteristics. When a test is nicely on target an even balance of right and wrong answers will be produced with small variations in question difficulty.

142. Essentially in two-stage testing there is no damping nor recovery mechanism. There is one steering opportunity only. The routing test aims the testee by dead reckoning and the course, once laid, is beyond further control. Lord has demonstrated that an appropriate multilevel test can absorb a certain amount of target error.

143. A flexilevel test has only two directions but it has a steering choice between fixed alternatives after each question: it has damping for the current direction but not for the alternative. A flexilevel test passes through the target questions and continues, it has no recovery after overshooting.

144. The small change in difficulty between the successive items of a branching test gives damping in both directions (easier and harder). Again there is a steering decision between fixed alternatives. The ongoing test can consequently hover in the target zone.

145. The amount of damping is important. The more damping there is then the more items are needed and the slower the test is to reach the target zone. The Robbins-Monro shrinking-step procedures have increasing damping as the test proceeds, but this also impairs their recovery after anomalous responses.

146. The approaches discussed so far have very limited steering or tailoring capacity. The procedures to be looked at in Part F below differ in having much more flexible control over steering.

F. Item-finding procedures

147. Tailoring a test to suit a testee would be done most closely if each item were individually chosen rather than one of many predetermined item networks being followed. Ideally a procedure is wanted which at any stage in testing, after taking stock of the information to hand, will select the next item best to achieve the purpose of the assessment. Two such item-finding procedures have been proposed - based on Bayesian and on maximum-likelihood methods. The two procedures differ in some points of approach and in the method of selecting the next item; they have in common a theoretical base in item characteristic curve theory.

Bayesian approach

148. Most work has been done on a Bayesian approach. Owen (1969) put forward a theoretical model which included the possibility of chance success. He assumes, and similar assumptions are common to most of the research discussed here, normal ogive item characteristic curves with known item parameters, and a normal prior distribution of testee ability. He derives an expression for the posterior distribution of testee ability that will obtain after answering a given question. He goes on to indicate a criterion by which that next question can be selected from the available pool so as to give the smallest variance in the resulting estimate of ability.

149. Owen's procedure includes two approximations. The posterior distribution always depends on a normal approximation for the actual distribution of ability prior to the current item. And in choosing the next item this is only optimum in the 1-step sense. How well a series of locally optimum single steps produces a globally optimum sequence is an open question.

150. Urry (1971) and Wood (1971) both used Owen's model. Urry carried out Monte Carlo simulations using three item banks, two of which were idealised, while the third took the parameters of an existing test. For his idealised item banks he took high values of item discrimination (a=1.6) with a probability of chance success of 0.2 - so that Figure 7 applies. 50 testees were simulated for each of these banks, and 100 testees for the existing-test simulation.

151. A distinctive advantage of item-finding procedures is that the individualised test does not have a fixed length. As few or as many items may be selected in turn as necessary to achieve a specified degree of precision. Urry specified standard errors of measurement of 0.32 and 0.25 as termination values (the assumed distribution of ability being taken as having a standard deviation of unity). In the case of the existing-test simulation (with items of lower discrimination) an alternative termination criterion of 30 items was additionally employed.

152. The less precise termination criterion was achieved by the idealised high discrimination item banks in about 11 or 12 items on the average, the more precise criterion in about 17 or 18. The existing-test simulation used an average of 27.5 items before reaching either the 0.25 precision criterion or the 30-item limit. Correlations with underlying ability were of the order of 0.94/0.95 - which is to be expected being only an alternative way of defining precision, although Urry presents this confusingly as a validity rather than a reliability relationship.

153. Even for the existing-test simulation these are good results. The reliability achieved in 27.5 items was comparable with that for the simulated test total score based on 80 items.

154. Wood's research included a Monte Carlo simulation based on a real pool of vocabulary items. Applying Owen's model he found that about 40 items were able to match a 60-item conventional test. Better reduction in measurement error was achieved in some parts of the ability range than in others and this could be attributed to the skewed nature of the item pool. A 60-item two-stage procedure was better than the Bayesian item-finding approach at the poorer end of the item pool. In the Bayesian approach rapidly diminishing returns were experienced after about item 20.

156. Jensema (1974) also developed and tried out a Bayesian approach, again with minimisation of posterior variance as the criterion for successive item selections. A real data simulation was carried out using a response bank obtained from the administration of four quantitative tests to high school students. A sample of 5,000 pupils was used to estimate approximately the characteristic curve parameters of the items. From the 110 initial items sixteen were dropped as being too often unattempted, and a further 35 items were dropped as their discrimination (parameter a) was below 0.6. A further sample of 1,000 pupils was then used to obtain more exact maximum-likelihood estimates for the characteristic curve parameters of the remaining 59 items. At this stage one further item was deleted and 6% of pupils eliminated as repeatedly not converging during maximum-likelihood estimation.

157. The termination criteria were those of Urry's - a standard error of measurement of 0.25 or 30 items. The average number of items used was about 27. The ability estimates correlated 0.85 with the conventional 110-item combined test score, but this is inflated by a part/whole relationship and hence is surprisingly low.

158. Jensema also carried out Monte Carlo simulations using idealised item banks with item discrimination, parameter a, set at 0.8, 1.6, and 2.4. Estimates correlated 0.95 with underlying ability. For the least discriminating item bank no test sequence reached the required precision in 30 items - 35 items was a subjective estimate of the average number of items required. The two item banks with high discrimination required an average of about 18 and 10 items respectively. Owen (1975) has produced a further theoretical Bayesian model which has the considerable advantage that it does not require an exact choice of item parameters. In practice item parameters will not be known exactly so that some tolerance is necessary. In other assumptions and approximations the model is the same as his earlier one.

159. Some of the Bayesian approaches have included a choice of starting point where there has been prior information to base this on. The capacity for a tailored start has generally been seen as desirable and likely to improve test effectiveness. However, Jensema (1974) also studied the value of such prior information. Where prior information correlating 0.6 with the ability being assessed was available this only gave an average saving of about one item for the a=1.6 item bank. The saving would be greater for less discriminating items or for a less precise termination criterion. While any saving is worth having if readily available the value of an appropriate start is perhaps better viewed as largely motivational.

---

160. The limited evidence available on Bayesian item-finding procedures suggests an appreciable advantage over previous individualised testing approaches. Control over error of measurement is also a useful benefit. The procedures make a number of assumptions which will need further real-data simulation and also empirical studies to bring to light any resulting deficiencies.

161. Despite the writer having argued that global item parameters can only be a first approximation to their usefulness in tailored testing, it is clear in the Bayesian studies that higher values of global item discrimination mean fewer items needed to termination. This is because we are dealing here with assumed normal ogive characteristic curves. Given this theoretical basis three item parameters completely specify the item. However, real items can be expected to show some deviations from the assumed distribution. This will degrade the effectiveness of item selection, and the sensitivity of the procedures here to variations in item discrimination confirms the likelihood of this. Procedures which are aware of actual characteristic curves should show to advantage. Jensema (1974) makes a related point (p.44),

> "A more basic question, which directly challenges the assumptions of the Bayesian item-finding model, is whether the guessing parameter is constant over all levels of ability. The model assumes that the Cg [guessing parameter] value is the same for any θ [ability] value. This seems questionable because an incorrect choice which appears reasonable at one level of knowledge may appear absurd at another".

162. It is characteristic of the Bayesian methods that before each item selection they scan all of the unused item pool. This requires much greater computing capacity than methods previously looked at. It also acts to limit the size of the item pool used, and this is a considerable snag as efficiency of testing would be expected to be related to the quality, the coverage and the depth of cover of the item pool.

Maximum-likelihood approach.

163. The other proposal for item-finding procedures uses maximum-likelihood methods. That is, after any sequence of item responses it is possible, given known item parameters and assuming some form for the characteristic curves, to determine the ability at which the observed sequence is most likely. The item next selected is then the one with difficulty level closest to the current ability estimate.

164. Urry's (1970) is the only general research of this kind although Reckase (1974 a & b) also uses a maximum-likelihood approach but working with the Rasch 1-parameter item model which considers only differences in item difficulty.

165. Maximum-likelihood estimation can only sensibly begin once a testee has made both right and wrong answers. Consequently initial item paths have to be available to route the testee until he satisfies this precondition. Urry chooses to proceed immediately to the appropriate extreme of difficulty after a first question of median difficulty, while Reckase progresses by halving or doubling difficulty as appropriate until a contrary answer has been obtained.

166. Urry's Monte Carlo simulation study also had a basis in the Rasch model but went on to include a two-parameter variation which took the probability of chance success into account, and - more importantly - he systematically varied item discrimination. It is relevant to note that his approach, because of his initial routing tactics, necessarily included items at one extreme or other of the difficulty range, for his results indicated that an item bank with a rectangular distribution of difficulty was better than one with a peaked distribution. This can be seen to be a direct consequence of his approach. He found for his method that item discrimination needed to be high, with parameter a at 0.8 or higher, to show advantage over conventional testing. When these conditions were satisfied considerable reductions in test length were achieved for the same standard error of measurement as compared with a conventional test.

167. The maximum-likelihood approach also requires an assumption for the form of the item characteristic curve, but does not require an assumption for the distribution of ability. In scanning the remaining item pool to select the next question the specification is simpler than for the Bayesian approach being only a match on difficulty. On the other hand to update the ability estimate after successive items becomes increasingly onerous as all previous answers and their item characteristic curves must be appraised afresh. Again, then, the method requires substantial computing resources. In this case the computing requirements would act to limit test length rather than to limit the item pool. Urry's initial strategy in particular seems at risk to an anomalous response to the first item, but this is not central to the maximum-likelihood procedure and can be readily overcome.

---

168. Altogether the item-finding approaches show potentially high benefits, although the best results are achieved by unrealistically high levels of item discrimination. The requirement for high levels of global item discrimination also seems partly self-defeating. An apologist for conventional testing could justifiably argue that the conventional methods have evolved to work with the items that are available and that the availability of super-items is by no means guaranteed.

169. With continuing progress in computer technology there is perhaps little point in emphasising the possible restrictions from computer requirements. Even so there will presumably continue to be a cost advantage to methods which can function with slower smaller machines.

G. Stradaptive and Broad-range approaches.

170. These two methods are due respectively to Weiss (1973 b) and Lord (1975 b); they have a number of similarities and can be regarded as simpler item-finding strategies

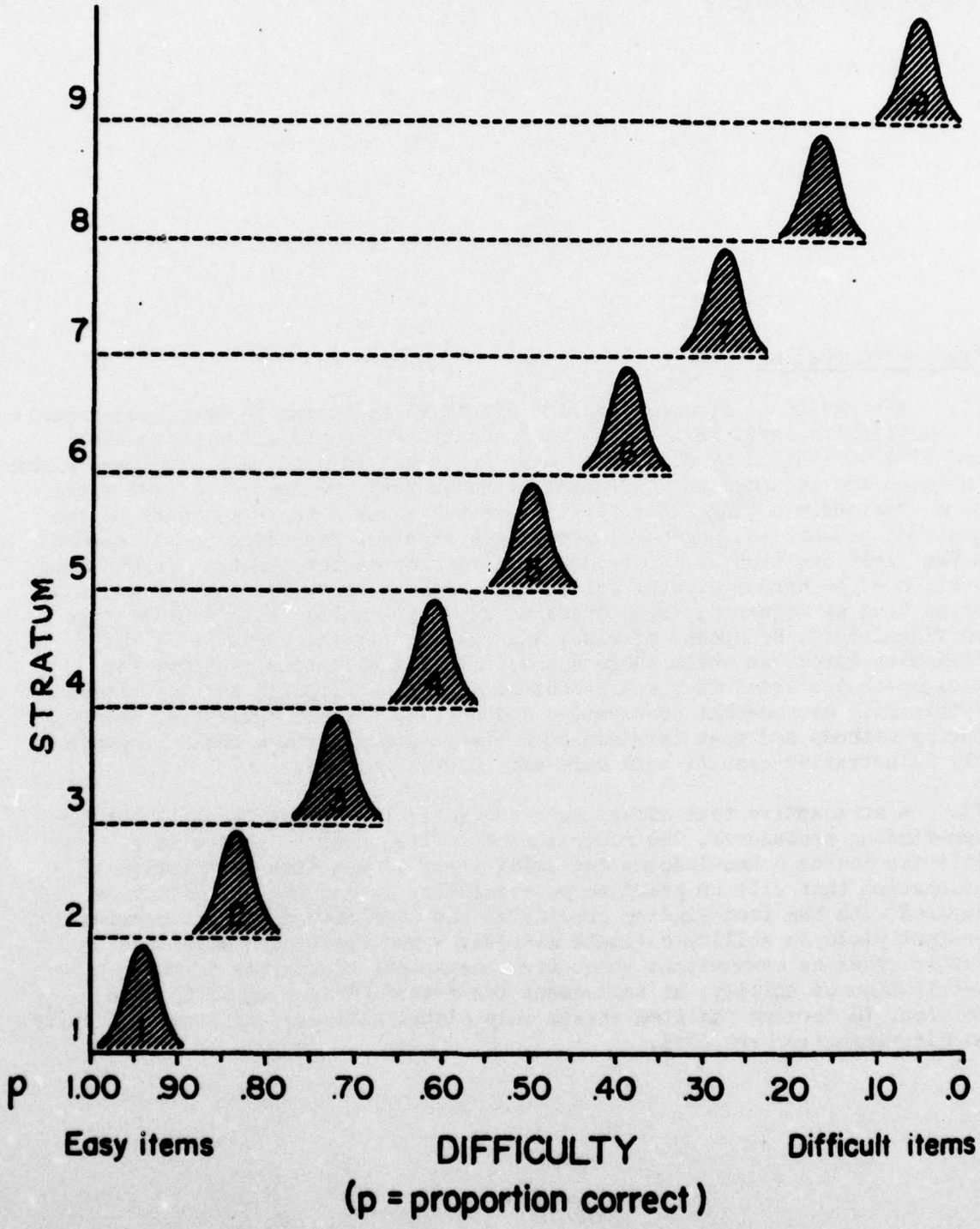
Stradaptive testing

171. A stradaptive (from stratified adaptive) test uses an item pool organised by difficulty level into a number of strata. Figure 13 illustrates the kind of distribution by difficulty that Weiss has in mind. All the items within a stratum are regarded as equivalent although they are queued for use with the most discriminating items first. A testee makes a tailored entry to the item pool at what is judged an appropriate stratum. Depending on his answer to the first question he is moved to a harder or easier stratum for his next question - the harder stratum following a right answer. Testing may continue for as long as required. Weiss draws an explicit analogy with a Binet-type individual test. He speaks of basal and ceiling strata, these being the difficulty levels at which success and failure are certain. Failure for multiple-choice items is taken as chance success - although the definition of this will be somewhat problematic and necessarily probabalistic. Several scoring methods and test termination criteria are possible. Weiss presents only illustrative results both here and in Weiss (1974).

172. A stradaptive test offers more controllable testing than all but the item-finding procedures. The reduction of an item pool to strata is a realistic device acknowledging the fallibility of the item descriptive information that will in practice be available. It has the deficiency as compared with the item-finding procedures that the method of test scoring does not yield an ability estimate directly - the associated advantage is that it makes no assumptions about item characteristic curves or the distribution of ability; at the moment its method of scoring is an open question. In forming the item strata only global estimates of item difficulty and discrimination are used.

FIGURE 13 An example of an item pool for a stradaptive test.

(after Weiss, 1973 b)



Broad-range testing

173. Lord's broad-range tailored test is so called because its aim is to provide effective assessment from Fifth Grade pupils upwards. The test described is one of verbal ability: it draws on a wide range of existing tests to give an item pool of 182 items of five types. Items were chosen for type and difficulty level not for discriminating power.

174. The items are grouped in ten difficulty levels. Essentially items are again queued within difficulty level in decreasing order of their discrimination. In fact, because there are five item types certain adjustments of detail are made here and elsewhere to ensure a generally uniform mix of types. A testee makes a tailored start at an appropriate entry level. He is then routed to easier or harder levels depending on his answer. This routing continues only until at least one wrong and one right answer are available. At this point maximum-likelihood procedures are introduced in conjunction with item characteristic curve theory to find the ability at which the observed answers are most likely. Now the next item is selected, from all items of the appropriate type, that gives the most information at the estimated ability level. The procedure continues for a fixed test length of 25 items.

175. The design presented is reported by Lord as one chosen from about thirty following simulations based on 1,000 or so simulated testees. An item pool of double the size gave results that were twice as good, the gain being mainly attributed to the availability of more items and only partly to their arrangement in more and closer difficulty levels. Figure 14 shows the information function for the broad-range test (for entry at ability level 0.75) compared with that for three conventional tests adjusted down to the same 25-item length - the conventional tests are three forms of the Preliminary Scholastic Aptitude Test.

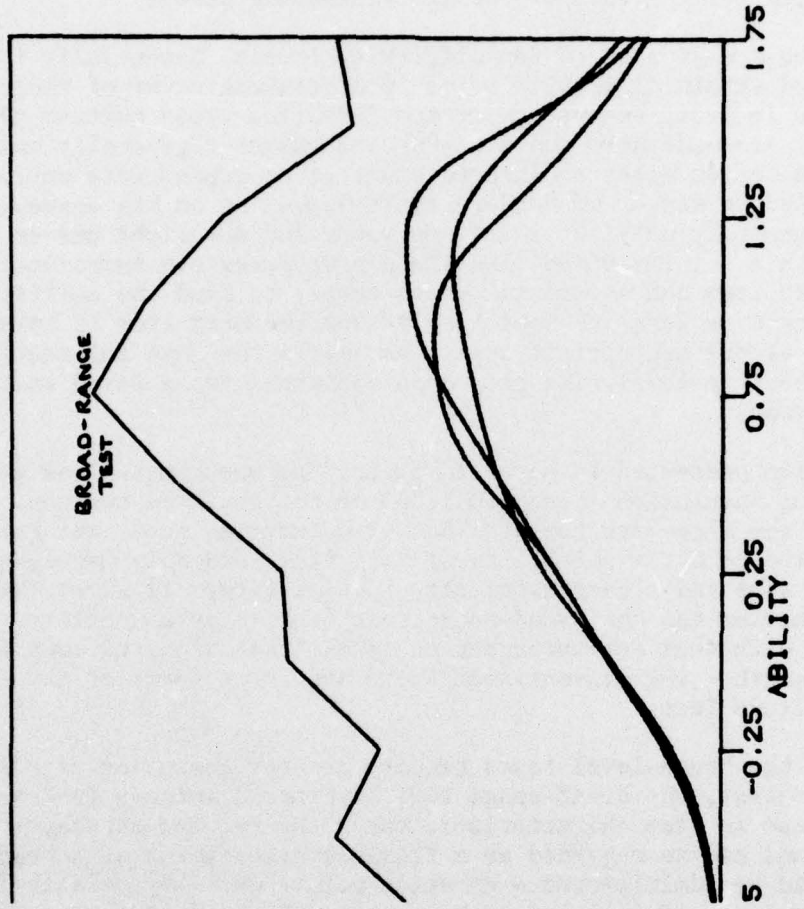
176. Clearly the broad-level tests require greater computing capacity than the stradaptive test. The broad-range test has the advantages (and disadvantages) of a base in item characteristic curve theory. The stradaptive test on the other hand can be regarded as a flexible development of a branching test which could be administered - as Weiss points out - by relatively simple equipment. Lord does not give information about the kind of route through his 10-levels that is taken after his maximum-likelihood procedures come into play - granted that it becomes irrelevant to his method it would nevertheless be of interest if such routes approximated to some stepping rule.

177. The limitation of the broad-level test to 25 items is somewhat arbitrary, but some practical limit is imposed by the item pool available and possibly by the increasing computing load for maximum-likelihood estimation .

OVERVIEW

178. Sections 4 and 5 respectively reviewed the statistical antecedents of tailored testing and made a case for the context-free use of individual items.

FIGURE 14 A comparison of a broad-range test and three conventional tests. (Simulation results from Lord (1975 b).)



179. It has been Sections 6 and 7 which have traced the development of tailored testing in educational and psychological measurement. Recent developments have shown considerable promise and there seems now little doubt that operationally useful instances of individualised testing will be with us shortly. However, there have been many difficulties - not least of which has been the translation from theoretical to empirical modes of research. A number of concepts and points have been picked up in the course of this review and these are indicative of the method of tailored testing that is proposed for initial development in study R43 (the longer term research study within which the work begun in project P560 is now continuing). At the time of writing a first series of real-data simulation studies has been completed with promising results. Details will be published in a subsequent Report.

8. REFERENCES

ANASTASI, A. An empirical study of the applicability of sequential analysis to item selection. Educational and Psychological Measurement, 1953, 13, 3-13.

ANASTASI, A. Psychological Testing. New York: Macmillan, 1954.

ANDERSON, T. W., McCARTHY, P. J., & TUKEY, J. W., "Staircase" methods of sensitivity testing. Navord Report 65-46. Statistical Research Group, Princeton University, 1946, 1-134.

ANGOFF, W. H. & HUDDLESTON, E. M., The multi-level experiment. A study of a two-level test system for the College Board Scholastic Aptitude Test. Statistical Report 58-21, Princeton, New Jersey: Educational Testing Service, 1958.

ARMITAGE, P. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. Journal of the Royal Statistical Society, 1950, 12, 137-144.

ATKINSON, R. C. & WILSON, H. A. (Eds.) Computer Assisted Instruction. New York: Academic Press, 1969.

BAYROFF, A. G. THOMAS, J. J., & ANDERSON, A. A. Construction of an experimental sequential item test. Research Memorandum 60-1 Personnel Research Branch, U.S. Dept. of Army, 1960.

BAYROFF, A. G. Feasibility of a programed testing machine. Report 64-3, U.S. Army Personnel Research Office, 1964.

BAYROFF, A. G. & SEELEY, L. C. An exploratory study of branching tests. Technical Research Note 188, Washington D.C. U.S. Army BESRL, 1967.

BAYROFF, A. G., ROSS, R. M., & FISCHL, M. A. Development of a programed testing system. Technical Paper 259, U.S. Army, RIBSS, 1974.

BETZ, N. E. & WEISS, D. J. An empirical study of computer-administered two-stage ability testing. Research Report 73-4, Psychometric Methods Program, University of Minnesota, 1973.

BETZ, N. E. & WEISS, D. J. Simulation studies of two-stage ability testing. Research Report 74-4, Psychometric Methods Program, University of Minnesota, Minneapolis, 1974.

BETZ, N. E. & WEISS, D. J. Empirical and simulation studies of flexilevel ability testing. Research Report 75-3, Psychometric Methods Program, University of Minnesota, 1975.

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. In Lord & Novick (Eds.), 1968, Chapters 17-20.

BROWNLEE, K. A., HODGES, J. L. & ROSENBLATT, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 262-277.

BRYSON, Rebecca. A comparison of four methods of selecting items for computer assisted testing. Technical Bulletin STB 72-8, San Diego, California, Psychological Sciences Division, ONR, 1971.

BRYSON, Rebecca. Shortening Tests: effects of method used, length, and interval consistency on correlation with total score. Proceedings, 80th Annual Convention of the American Psychological Association, Honolulu, 1972, 7, 7-8.

BURGESS, G. C. Use of sequential analysis for determining test item difficulty level. Educational and Psychological Measurement, 1955, 15, 80-86.

CLEARY, T. A., LINN, R. L. & ROCK, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345360, a.

CLEARY, T. A., LINN, R. L. & ROCK, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187, b.

COCHRAN, W. G. & DAVIS, M. Stochastic approximation to the median effective dose in bioassay. In J. Gurland (Ed.), Stochastic models in Medicine and biology, Madison: University of Wisconsin Press, 1964, pp. 281-300.

COCHRAN, W. G, & DAVIS, M. The Robbins-Monro method for estimating the median lethal dose. Journal of the Royal Statistical Society, Series B, 1965, 27, 28-44.

CORNSWEET, T. N. The staircase method in psychophysics. American Journal of Psychology, 1962, 75, 485-491.

COWDEN, D. J. An application of sequential sampling to testing students. Journal of American Statistical Association, 1946, 41, 547-556.

CRONBACH, L. J. & GLESER, G. C. Psychological tests and Personnel decisions. Urbana, University of Illinois Press, 1957 (1st. ed.) 1965 (2nd. ed.).

DAVIS, M. Comparison of sequential bioassays in small samples. Journal of the Royal Statistical Society, Series B, 1971, 33, 78-87.

DIXON, W. J. & MOOD, A. M. A method for obtaining and analyzing sensitivity data. Journal of American Statistical Association, 1948, 43, 109-126.

DuBOIS, P. H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25, 532-536.

DUNCAN, K. D. Experiments with an inexpensive device for programmed instruction in the multiple choice branching style. Programmed Learning, 1, 145-154, 1964.

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

FERGUSON, R. L. Computer-assisted criterion-referenced testing. Working paper No. 49, University of Pittsburgh, Learning Research & Development Center, 1969.

FERGUSON, R. L. Computer assistance for individualising measurement. Report 1971/8, Pittsburgh, Pa., University of Pittsburgh, Learning Research & Development Center, 1971, a.

FERGUSON, R. L. & HSU, T. The application of item generators for individualising measurement. Report 1971/14, University of Pittsburgh, Learning Research & Development Center, 1971, b.

FLAUGHER, R. L., MELTON, R. S. & MYERS, C. T. Item re-arrangement under typical test conditions. Educational and Psychological Measurement, 1968, 28, 813-824.

FREEMAN, P. R. Optimal Bayesian sequential estimation of the median effective dose. Biometrika, 1970, 57, 79-89.

GREEN, B. F. Comments on tailored testing. In W. J. Holtzman (Ed.), 1970.

GREENWOOD, D. I. & TAYLOR, C. Adaptive testing in an older population. Journal of Psychology, 1965, 60, 193-198.

HANSEN, D. N. An investigation of computer-based science testing. In Atkinson, R. C. & Wilson, H. A. (Eds.), 1969.

HANSEN, D. N., JOHNSON, B. F., FAGAN, R. L., TAM, P. & DICK, W. Computer-based adaptive testing models for the air force technical training environment Phase 1: Development of a computerized measurement system for air force technical training. Report AFHRL-TR-74-48, Air Force Human Resources Laboratory, Technical Training Division, Lowry Air Force Base, Colorado, 1974.

HICK, W. E. Information theory and intelligence tests. British Journal of Psychology, Statistics Section, 1951, 4, 157-164.

HOLTZMAN, W. H. (Ed.). Computer-assisted instruction, testing and guidance, New York: Harper and Row, 1970.

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

HUTT, M. L. A clinical study of "consecutive" and "adaptive" testing with the revised Stanford-Binet. Journal of Consulting Psychology, 1947, 11, 93-103.

BROWNLEE, K. A., HODGES, J. L. & ROSENBLATT, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 262-277.

BRYSON, Rebecca. A comparison of four methods of selecting items for computer assisted testing. Technical Bulletin STB 72-8, San Diego, California, Psychological Sciences Division, ONR, 1971.

BRYSON, Rebecca. Shortening Tests: effects of method used, length, and interval consistency on correlation with total score. Proceedings, 80th Annual Convention of the American Psychological Association, Honolulu, 1972, 7, 7-8.

BURGESS, G. C. Use of sequential analysis for determining test item difficulty level. Educational and Psychological Measurement, 1955, 15, 80-86.

CLEARY, T. A., LINN, R. L. & ROCK, D. A. An exploratory study of programmed tests. Educational and Psychological Measurement, 1968, 28, 345360, a.

CLEARY, T. A., LINN, R. L. & ROCK, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187, b.

COCHRAN, W. G. & DAVIS, M. Stochastic approximation to the median effective dose in bioassay. In J. Gurland (Ed.), Stochastic models in Medicine and biology, Madison: University of Wisconsin Press, 1964, pp. 281-300.

COCHRAN, W. G, & DAVIS, M. The Robbins-Monro method for estimating the median lethal dose. Journal of the Royal Statistical Society, Series B, 1965, 27, 28-44.

CORNSWEET, T. N. The staircase method in psychophysics. American Journal of Psychology, 1962, 75, 485-491.

COWDEN, D. J. An application of sequential sampling to testing students. Journal of American Statistical Association, 1946, 41, 547-556.

CRONBACH, L. J. & GLESER, G. C. Psychological tests and Personnel decisions. Urbana, University of Illinois Press, 1957 (1st. ed.) 1965 (2nd. ed.).

DAVIS, M. Comparison of sequential bioassays in small samples. Journal of the Royal Statistical Society, Series B, 1971, 33, 78-87.

DIXON, W. J. & MOOD, A. M. A method for obtaining and analyzing sensitivity data. Journal of American Statistical Association, 1948, 43, 109-126.

DuBOIS, P. H. Varieties of psychological test homogeneity. American Psychologist, 1970, 25, 532-536.

JENSEMA, C. J. An application of latent trait mental test theory. British Journal of Mathematical and Statistical Psychology, 1974, 27, 29-48.

KAPPAUF, W. E. Use of an on-line computer for psychophysical testing with the up-and-down method. American Psychologist, 1969, 24, 207-211.

KENT, G. H. Suggestions for the next revision of the Binet-Simon scale. Psychological Record, 1937, 409-432.

KESTON, H. Accelerated stochastic approximation. Annals of Mathematical Statistics, 1958, 29, 41-59.

KILLCROSS, M. C. & CASSIE, A. The potential use of tailored testing for allocation to Army employments. Occasional Note APRE 41/73, Farnborough, Hants: Army Personnel Research Establishment, 1973, also in Singleton, W. T. & Spurgeon, P. (Eds.) Measurement of human resources, London: Taylor & Francis, 1975, pp. 117-122.

KILLCROSS, M. C. A tailored testing system for selection and allocation in the British Army. Montreal: paper presented at the 18th International Congress of Applied Psychology, 1974.

KILLCROSS, M. C. Tailored testing for selection and allocation. Doctoral dissertation, University of Edinburgh, 1975.

KILLCROSS, M. C., HAMMOND, D. R. F. & PRESTON, L. R. Revision of the selection centre test battery: I. A multiple-choice verbal test. APRE Report 41/75, Farnborough, Hants: Army Personnel Research Establishment, 1976.

KRATHWOHL, D. R. & HUYSER, R. J. The sequential item test. American Psychologist, 1956, 11, 419 (Abstract).

LARKIN, K. C. & WEISS, D. J. An empirical investigation of computer-administered pyramidal ability testing. Research Report 74-3, Psychometric Methods Program, University of Minnesota, 1974.

LARKIN, K. C. & WEISS, D. J. An empirical comparison of two-stage and pyramidal adaptive ability testing. Research Report 75-1 Psychometric Methods Program, University of Minnesota, 1975.

LAZARSFELD, P. F. Latent structure analysis. In S. Koch (Ed.) Psychology: a study of a science. Vol. 3, New York: McGraw-Hill, 1959, 476-542.

LINN, R. L., ROCK, D. A. & CLEARY, T. A. The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 1969, 29, 129-146.

LINN, R. L., ROCK, D. A. & CLEARY, T. A. Sequential Testing for dichotomous decisions. Educational and Psychological Measurement, 1972, 32, 85-95.

LORD, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

LORD, F. M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

LORD, F. M. Item sampling in test theory and in research design. Research Bulletin 65-22, Princeton, New Jersey: Educational Testing Service, 1965. (Now superseded by Chapter 11, Lord & Novick (1968)).

LORD, F. M. Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. Psychometrika, 1970, 35, 43-50, a.

LORD, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), 1970, b.

LORD, F. M. Robbins-Monro procedures for tailored testing. Educational and Psychological Measurement, 1971, 31, 3-31, a.

LORD, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151, b.

LORD, F. M. A theoretical study of two-stage testing. Psychometrika, 1971, 36, 227-242, c.

LORD, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 31, 805-813, d.

LORD, F. M. Tailored testing, an application of stochastic approximation. Journal of the American Statistical Association, 1971, 66, 707-711, e.

LORD, F. M. Individualised testing and item characteristic curve theory. In Atkinson R. C. et al (Eds.), Contemporary developments in mathematical psychology, Vol. 2, San Francisco: W. H. Freeman & Co., 1974, a.

LORD, F. M. Practical methods for redesigning a homogeneous test, also for designing a multilevel test. Research Bulletin 74-30, Educational Testing Service, Princeton, New Jersey, 1974, b.

LORD, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-217, a.

LORD, F. M. A broad-range tailored test of verbal ability. Research Bulletin 75-5, Princeton, New Jersey, Educational Testing Service, 1975, b. (Published in Proceedings of the First Conference on Computerized Adaptive Testing. US Civil Service Commission, 1976 - see Section 9 below).

LORD, F. M. & NOVICK, M. R. (Eds.) Statistical Theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

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

McBRIDE, J. R. & WEISS, D. J. Recent and projected developments in ability testing by computer. Paper presented at "Occupational Research and the Navy: Prospectus 1980," a symposium sponsored by the Navy Personnel Research and Development Center, San Diego, California, 1973.

McBRIDE, J. R. & WEISS, D. J. A word knowledge item pool for adaptive ability measurement. Research Report 74-2, Psychometric Methods Program, University of Minnesota, 1974.

McCARTHY, P. J. A class of methods for estimating reaction to stimuli of varying severity. Journal of Educational Psychology, 1949, 40, 143-156.

McGILL, P. A. The concept of a programmed, branching, or sequential item test. Paper presented at the Defence Psychologists Symposium, Shrivenham, 1968. Army Personnel Research Establishment, 1968.

McNEMAR, Q. The revision of the Stanford-Binet Scale. Boston: Houghton Mifflin Co., 1942.

MOLLENKOPF, W. G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. Psychometrika, 1950, 15, 297-315.

MOONAN, W. J. Some empirical aspects of the sequential analysis technique as applied to an achievement examination. Journal of Experimental Education, 1950, 18, 195-207.

MUSSIO, J. J. A modification to Lord's model for tailored tests. Unpublished Doctoral dissertation, University of Toronto, 1972.

OWEN, R. J. A Bayesian approach to tailored testing. Research Bulletin 69-92, Princeton, New Jersey, Educational Testing Service, 1969.

OWEN, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

OWENS, T. R. & STUFFLEBEAM, D. L. An experimental comparison of item sampling and examinee sampling for estimating test norms. Journal of Educational Measurement, 1969, 6, 75-83.

PATERSON, J. J. An evaluation of the sequential method of psychological testing. Doctoral dissertation, Michigan State University, 1962, also Ann Arbor, Michigan: University Microfilms, 1962, No. 63-1748.

RECKASE, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behaviour Research Methods & Instrumentation, 1974, 6, 208-212, a.

RECKASE, M. D. An application of the Rasch simple logistic model to tailored testing. Paper presented at the Annual Meeting of the American Educational Research Association, 1974, b.

ROBBINS, H. & MONRO, S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 22, 400-407.

ROSE, R. M., TELLER, D. Y. & REMDLEMAN, P. Statistical properties of staircase estimates. Perception and Psychophysics, 1970, 8, 199-204.

SAX, G. & CROMACK, T. R. The effects of various forms of item arrangement on test performance. Journal of Educational Measurement, 1966, 3, 309-311.

SEELEY, L. C., MORTON, M. A. & ANDERSON, A. A. Exploratory study of a sequential item test. Technical Research Note 129, Washington D.C. U.S. Army Personnel Research Office, 1962.

SIROTNIK, K. An investigation of the context effect in matrix sampling. Journal of Educational Measurement, 1970, 7, 199-207.

STATISTICAL RESEARCH GROUP, COLUMBIA UNIVERSITY. Sequential analysis of statistical data: applications. New York: Columbia University Press, 1945.

STOCKING, M. Short tailored tests. Research Bulletin 69-73, Princeton, New Jersey, Educational Testing Service, 1969.

TAYLOR, M. M. & CREELMAN, C. B. PEST: efficiency estimates on probability functions. Journal of the Acoustical Society of America, 1967, 41, 782-787.

TERMAN, L. M. in McNEMAR, Q., 1942, Chapter 1.

TSUTAKAWA, R. K. Asymptotic properties of the block up-and-down method in bioassay. The Annals of Mathematical Statistics, 1967, 38, 1822-1828.

URRY, V. W. A Monte Carlo investigation of logistic mental test models. Unpublished Doctoral dissertation. Purdue University, 1970.

URRY, V. W. Individualised testing by Bayesian estimation. Bureau of Testing, University of Washington, April 1971, a.

URRY, V. W. Approximation methods for the item parameters of mental test models. Bureau of Testing, University of Washington, December 1971, b. (Also Educational and Psychological Measurement, 1974, 34, 253-269).

WALD, A. Sequential analysis. New York: Wiley, 1947.

WALD, A. Statistical decision functions. New York: Wiley, 1950.

WATERS, C. J. Preliminary evaluation of simulated branching tests. Technical Research Note 140, Washington, D.C. U.S. Army Personnel Research Office, 1964.

WATERS, C. W. & BAYROFF, A. G. A comparison of computer-simulated conventional and branching tests. Educational and Psychological Measurement, 1971, 31, 125-136.

WEISS, D. J. & BETZ, N. E. Ability measurement: conventional or adaptive? Research Report 73-1, Psychometric Methods Program, University of Minnesota, 1973, a.

WEISS, D. J. The stratified adaptive computerized ability test. Research Report 73-3, Psychometric Methods Program, University of Minnesota, 1973, b.

WEISS, D. J. Strategies of adaptive ability measurement. Research Report 74-5, Minnesota, MN. Psychometric Methods Program, University of Minnesota, 1974.

WETHERILL, G. B. Sequential estimation of quantal response curves. Journal of the Royal Statistical Society, Series B, 1963, 25, 1-38.

WILLIAMS, E. J. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research, 1949, 3A, 351-363.

WOOD, R. The efficacy of tailored testing. Educational Research, 1969, 11, 219-222.

WOOD, R. Computerized adaptive sequential testing. Unpublished doctoral dissertation, University of Chicago, 1971.

WOOD, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

9. ADDITIONAL REFERENCES

A number of articles have come to the author's notice since completing the original review. The opportunity is taken here of including these later entries without comment.

BETZ, N. E.& WEISS, D. J. Effects of immediate knowledge of results and adaptive testing on ability test performance. Research Report 76-3, Psychometric Methods Program, University of Minnesota, 1976.

BETZ, N. E. & WEISS, D. J. Psychological effects of immediate knowledge of results and adaptive ability testing. Research Report 76-4, Psychometric Methods Program, University of Minnesota, 1976.

CLIFF, N. Complete orders from incomplete data: interactive ordering and tailored testing. Psychological Bulletin, 1975, 82, 289-302.

CLIFF, N. A basic test theory generalisable to tailored testing. Technical Report No 1, Los Angeles: Department of Psychology, University of Southern California, 1975.

McBRIDE, J. R. & WEISS, D. J. Some properties of a Bayesian adaptive ability testing strategy. Research Report 76-1, Psychometric Methods Program, University of Minnesota, 1976.

MILLER, T. W. & WEISS, D. J. Effects of time limits on test-taking behaviour. Research Report 76-2, Psychometric Methods Program, University of Minnesota, 1976.

URRY, V. W. Computer assisted testing: the calibration and evaluation of the verbal ability bank. Technical Study 74-3, Washington: Personnel Research and Development Center, U.S. Civil Service Commission, 1974.

URRY, V. W. Computer-assisted testing with live examinees: a rendezvous with reality. Technical Note 75-3, Washington: Personnel Research and Development Center, U.S. Civil Service Commission, 1976.

URRY, V. W. Ancillary estimators for the item parameters of mental test models. Washington: Personnel Research and Development Center, U.S. Civil Service Commission, (in press).

U.S. CIVIL SERVICE COMMISSION. Proceedings of the First Conference on Computerized Adaptive Testing (PS-75-6). Washington: Personnel Research and Development Center, 1976.

VALE, C. D. & WEISS, D. J. A study of computer-administered stradaptive ability testing. Research Report 75-4, Psychometric Methods Program, University of Minnesota, 1975.

VALE, C. D. & WEISS, D. J. A simulation study of stradaptive ability testing. Research Report 75-6, Psychometric Methods Program, University of Minnesota, 1975.

WATERS, B. K. Empirical investigation of the stradaptive testing model for the measurement of human ability. Report AFHRL-TR-75-27, Williams Air Force Base, Arizona: Air Force Human Resources Laboratory, Flying Training Division, 1975.

WEISS, D. J. (Ed). Computerized adaptive trait measurement: problems and perspectives. Research Report 75-5 (Proceedings of a symposium at the 1975 APA Convention), Psychometric Methods Program, University of Minnesota, 1975.

WEISS, D. J. Final report: computerized ability testing, 1972-1975. Psychometric Methods Program, University of Minnesota, 1976.

DISTRIBUTION LIST

Ministry of Defence	No. of Copies
Army Department	
DCS(A)	1
AD/SAG(A)3	1
SAG(A)3a	6
VAG	1
DA Psych	1
DAT	1
DM(A)	1
DAR	1
HQ APSG	5
DA Ed	1
DWRAC	1
Central Staffs	
MOD Library (C & A)	2
Navy Department	
DSC(N)	1
SP(N)	1
Air Force Department	
CS(RAF)	1
ACS(T) RAF	1
Procurement Executive	
DRIC	2
Arms Directors Schools and Establishments	
RMAS	1
Staff College - TDRC	1
RNCS	1
WRAC Centre, Guildford	1
ASIT	1
APU/MTR/ARL	1
C Res O, HQ Training Command, RAF Brampton	1
RAF School of Education	1
Home Commands	
HQ UKLF	1
Overseas Liaison	
HQ BAOR	1
BDLS (Ottawa)	1
BDLS (Canberra)	1
BDS Washington	3
Medical Research Council	
Assistant Sec APRC	8

Standardisation and Liaison Staff in UK

Canadian Defence Research Board Member, 1 Grosvenor Square, London W1	4
Australian Army Liaison Staff, (for onward transmission to Australian Army Pyschological Research Unit)	6
U.S. Army Standardization - attention Lt Col Corby	12

Consultants to APRE/DCS(A)

D. McMahon	1

Miscellaneous

Behavioral Sciences Research Division, CSD	2
Director, Institute for Perception TNO Soesterberg, Netherlands (Dr P. L. Walraven)	1
Dr Norman Cliff, University of Southern California	1
Dr M. A. Fischl, U.S. Army Research Institute for Behavioral and Social Sciences	1
Dr Frederic M. Lord, Educational Testing Service	1
Dr V.W. Urry, U.S. Civil Service Commission	1
Dr D.J. Weiss, University of Minnesota	1