The Improvement of Measurement in Education and Psychology

Contributions of Latent Trait Theories

Edited by
Donald Spearritt
Professor of Education, University of Sydney
Vice-President, The Australian Council for Educational Research

The Australian Council for Educational Research
Golden Jubilee Year Invitational Seminar
22–23 May 1980

Australian Council for Educational Research

Published by
The Australian Council for Educational Research Limited
Radford House, Frederick Street, Hawthorn, Victoria 3122

Typeset, printed and bound by Brown Prior Anderson
5 Evans Street, Burwood, Victoria 3125

National Library of Australia Cataloguing-in-Publication data

The Improvement of measurement in education and psychology.

Bibliography.
Includes index.
ISBN 0 85563 226 7.

1. Educational tests and measurements.
2. Grading and marking (Students).
3. Psychological tests. 4. Personality tests.
I. Spearritt, Donald, 1921- II. Australian
Council for Educational Research. III. Title.

371.2'6

Contents

Preface

The Australian Council for Educational Research was established in 1930, under a grant from the Carnegie Corporation of New York. Three functions were held in 1980 to mark the fiftieth anniversary. Two were invitational seminars—one on the improvement of measurement in education and psychology and another on societal change and its impact on education. The third function was the presentation of a history of the ACER prepared by Professor W. F. Connell.

This volume contains the papers and reactant statements which were presented at the Invitational Seminar on the Improvement of Measurement in Education and Psychology. A seminar on this topic was considered to be highly appropriate for the anniversary celebrations, as measurement in education and psychology has been one of the main areas of the work of the ACER since its inception.

The seminar was held on 22 and 23 May in the Council Chamber of the University of Melbourne. Sixty-one people attended, including participants from most parts of Australia and from Canada, The People's Republic of China, Finland, Germany (FRG), Great Britain, New Zealand, and the United States. A highlight of the occasion was the presence of Emeritus Professor R. L. Thorndike, of Teachers College, Columbia University, who was especially invited by the ACER to give the opening paper. His visit to Australia was supported by the Australian American Educational Foundation.

The seminar was opened by the President of the ACER, Emeritus Professor P. H. Karmel, whose introductory statement is included in this volume.

It was decided in the planning stages of the seminar that the focus should be on the contribution that latent trait measures can make to education and psychology. In the 1960s and 1970s, psychometricians had devoted much effort to the development of latent trait measurement models. Yet measurement procedures based upon these models had been used for only a short time in the practice of educational and psychological measurement in Australia. Many practitioners in the field of measurement still had little or no knowledge of the features of the various latent trait models. It was thought that the time was opportune to bring some of the theoretical and practical aspects of latent trait measurement procedures to the attention of people in Australia working in or interested in the field of measurement in education and psychology.

Papers on various aspects of latent trait models were sought from appropriate authors in Australia and overseas. These were circulated in advance to participants in the seminar. The seminar itself took the form of a paper presentation, followed by a reactant statement on the paper, followed in turn by general discussion. The edited versions of the papers appear in this volume, together with the statements of reactants. Some of the discussion is caught up in the final paper, which represents the chairman's attempt to summarize the debate emerging from the seminar.

The seminar was undoubtedly successful in raising the level of awareness of many of the participants about theoretical and practical issues in measurement and particularly in latent trait measurement procedures. Since the papers represent original contributions by reputable authors in the field, the ACER believes that they deserve a much wider audience.

July 1981

Donald Spearritt
Vice-President
Australian Council for Educational Research

Acknowledgments

A committee consisting of Dr J. F. Izard, Professors J. A. Keats and B. McGaw, and Mr G. Morgan was set up by the ACER to help with the planning of the seminar. The editor acknowledges their assistance and, in particular, the great assistance of Mr Morgan in the organization of the seminar itself, in preparing appropriate materials, and in the production of this volume. The interest and assistance of Dr J. P. Keeves, the Director of the ACER, has been very much appreciated, as was that of the participants who helped to make the seminar a significant event in the history of the ACER.

Permission to reprint Figures 1 and 2 in the Spada and May paper which were originally published in H. Kluwe and H. S. Spada (Eds), *Developmental models of thinking*, was granted by Academic Press, New York.

Contributing Authors

David Andrich *University of Western Australia, Nedlands, Western Australia, Australia*

Bruce Choppin *National Foundation for Research in England and Wales, The Mere, Upton Park, Slough, Berks., England* (*now* with Centre for the Study of Evaluation, University of California, Los Angeles, United States of America)

Kevin F. Collis *University of Tasmania, Hobart, Tasmania, Australia*

Graham A. Douglas *University of Western Australia, Nedlands, Western Australia, Australia*

John F. Izard *Australian Council for Educational Research, Hawthorn, Victoria, Australia*

Peter Karmel *Chairman, Commonwealth Tertiary Education Commission; President, Australian Council for Educational Research*

John A. Keats *University of Newcastle, Newcastle, New South Wales, Australia*

Roderick P. McDonald *The Ontario Institute for Studies in Education, Toronto, Ontario, Canada* (*now* with Macquarie University, North Ryde, New South Wales, Australia)

Barry McGaw *Murdoch University, Murdoch, Western Australia, Australia*

Regine May *University of Freiburg, Freiburg, Federal Republic of Germany*

George Morgan *Australian Council for Educational Research, Hawthorn, Victoria, Australia*

Charles Poole *University of Melbourne, Parkville, Victoria, Australia*

Glenn Rowley *La Trobe University, Bundoora, Victoria, Australia*

Alan G. Smith *University of Newcastle, Newcastle, New South Wales, Australia*

Glen A. Smith *University of Melbourne, Parkville, Victoria, Australia*

Hans Spada *University of Freiburg, Freiburg, Federal Republic of Germany*

Donald Spearritt *University of Sydney, Sydney, New South Wales, Australia*

Robert L. Thorndike *Teachers College, Columbia University, New York, United States of America*

John D. White *Australian Council for Educational Research, Hawthorn, Victoria, Australia*

Introductory Statement

Peter Karmel

I have been asked, as President of the ACER, to open this seminar. Although I have no technical background in educational measurement, I am a consumer of educational measurements and, with my background as an economist, I have a predilection for measuring things which is similar to yours. I like to quantify attributes and concepts whether they are in economics or education or psychology, and my work in educational policy has usually had a statistical basis. Because of this I am well aware of the conflict one faces when one is trying to measure things. On the one hand there is the need to be precise and to emphasize what is measurable; on the other hand, in respect of policy questions, it is important to maintain a healthy scepticism about what is measurable and about the kinds of inferences that can be made from measurements and mathematical models.

This seminar is the first of three major functions arranged to mark the fiftieth anniversary of the establishment of the ACER. The organization was set up in 1930 through a grant from the Carnegie Corporation of New York. It is therefore most appropriate that the first paper should be given by a distinguished scholar from the United States. While this seminar is of a highly technical kind, being devoted to improving measurement in education and psychology, a second seminar on Societal Change and its Impact on Education is concerned with educational policy in Australia for the remainder of the twentieth century. The two topics provide a nice balance between the interests of the ACER in measurement and its technicalities on the one hand, and in educational policies in a changing social context on the other. The third function, during the Annual Meeting of the Council in October, is the presentation to the Council of a history of the ACER written by W. F. Connell.

It is relevant on occasions like this to issue a warning about divergencies between technical measurement and policy prescriptions derived from measurement. Some of the measures we use are relatively simple, such as the number of students in the last year of high school or the number of staff in institutions of higher education. But many of our measures, and particularly those which will be discussed in this seminar, are statistical constructs which do not precisely portray the concept that they are intended to measure. Since they cannot be used in any simple way, care has to be taken in drawing inferences from them. In economics, for example, we talk about the price level, and in Australia we have the consumer price index (CPI) which governs to a large extent the rate of change of wages in this country. The CPI is a statistical construct which does not measure changes in the price level according to any theoretical definition. Similar difficulties arise with measures of productivity or of the real value of the gross domestic product.

A second matter of concern is that the use of statistical constructs may have unintended results in the practical situations in which they are applied. Measures of educational attainment, for instance, may have consequences of a kind which are not the concern of those who are interested in their technical development. For example, they may lead to the labelling of children with particular kinds of problems, and this in turn can have implications for the way in which different groups in the community are treated, as in some of the Title I programs in the USA.

Thirdly, theoretical and mathematical models may be developed which include concepts of which statistical constructs are measures. Inferences may be drawn from these models and the statistical estimation of their parameters. It is easy to slide into a position where policy prescriptions are being made about a real world which is in fact very different from the theoretical world. In the field of economics, for example, it has become common to measure outcomes against the optimum properties of a world in which there is a free market and free competition. The underlying model, however, rests on a whole series of value assumptions about what is optimum. Further, the optimum properties of that kind of world hold only if the world is made up of a large number of small units. But we know that the world we live in is not made up of a large number of small units; it is made up of quite large aggregations of economic power—large corporations, trade unions, various pressure groups, and so on. To slide from the kinds of policies that one would advocate in a theoretical model to advocating those kinds of policies in the real world is simply not legitimate. My intuition suggests that the difficulties experienced in applying theoretical models in economics to policy questions in the real world are likely to occur also with respect to models relating to educational attainment or psychological testing or measurement.

While it is important to keep these points in mind, it is the purpose of this first seminar to consider some difficult technical problems in the field of measurement in education and psychology. This is a very proper way to begin the celebration of the fiftieth birthday of the ACER, given the role that measurement has played in its activities over the past 50 years.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright - ACER 1982

1
Educational Measurement—Theory and Practice

Robert L. Thorndike

I am honoured to open the program of this seminar celebrating the 50th anniversary of the Australian Council for Educational Research. It represents a major milestone in the history of this distinguished organization.

We could also, perhaps, claim to be celebrating the 75th anniversary of the birth of psychometric theory and practice, for it was just 75 years ago that Binet and Simon published their report of the first workable scale to assess intelligence. It was just 75 years ago, give or take a year, that Charles Spearman published his model of intellectual ability expressed in terms of *g*, a general intellectual factor, together with *s* factors each specific to a single test, and also the model of test score in terms of true score and error that provided the foundation of much of classical test theory. In the same year, in the United States, Edward Thorndike's *Mental and Social Measurements* introduced basic statistical concepts to the educational profession. We are still a relatively young discipline, when compared with the total range of disciplined inquiry, but we have moved far enough along so that we can well afford to pause to consider where we have been, where we are, and where we may be going.

Educational and psychological measurement have, since their inception, involved the parallel streams of practical test development and formulation of theoretical models of test performance. Binet was a pragmatist, assembling in his scale of intelligence a wide variety of tasks that he found useful in the practical task of differentiating those who were making normal school progress from those who were having difficulty in school. His work had little self-conscious theoretical structure. Concurrently Spearman was offering the theory that would provide a rationale for Binet's procedure, in that it postulated a pervasive general factor, *g*, extending through the whole range of cognitive tasks. Throughout the course of the past 75 years these two facets of the enterprise have continued side by side. As the practical test makers have generated test exercises, set scoring procedures, and combined the results into test scores, the test theorists have constructed models to account for how people perform on a test and to explicate what the test scores signify. We are never interested in the bits of behaviour that appear on a test solely for themselves, but always as signifying something more general, more lasting, more fundamental about the individual.

Models have, from the beginning, been developed to account for two distinct, but not unrelated, aspects of test performance. On the one hand, it was observed that two measurements designed to be as nearly as possible measures of the same identical characteristic of a person did not yield identical scores. A theory was required of measurement error. On the other hand, it was observed that measurements of what were designed to be different characteristics of a person were usually not independent, but were in varying degree related. A theory was required to picture the organization of human traits.

The classic theory of measurement error, or test reliability, presented in its essentials by Spearman 75 years ago, viewed a test score as made up of two components, a 'true score' and an error. The true score and error were conceived of as completely independent. The true score was viewed as unchanging from one form of a test to a parallel alternate form and from one occasion to another. The error was considered to be unique to the specific measurement, and to be entirely independent of the error that might equally be expected to appear on another measurement designed to assess that same characteristic of a person. Of course, the true score, as such, could never be directly observed. Its existence and properties could only be inferred from consistency of performance from one test exercise to another or from one test score to another.

This classic model of true score and error dominated the conception of test scores for most of the first 50 years of psychometric theory. It was a productive model in that it led to the formulation of a number of useful relationships. Very early it produced a statement of the relationship between test precision and test length. It permitted the development of estimates of the precision of difference scores and change scores, bringing out their fragile and undependable nature, and at the same time it permitted estimation of the properties of composites of two or more different measures. It led to estimates of the degree to which indices of relationship between measures of different attributes are attenuated by the error of measurement in each.

However, at the same time, this model generated certain problems. Thus, when two forms of a test yielded noticeably different mean scores for a group, which one should be considered to correspond to the true score? Or when a test form yielded higher scores if given second in a sequence of two testings than when it was the first test, which should be considered the true score—when given first or when given second? Again, if different data-gathering procedures— split-test, alternate form, retest – gave different values for the reliability coefficient, which one correctly identified the 'true' proportion of true-score variance?

So a somewhat different model has emerged in the past 25 years. This views any given test performance as being a sample from a universe of behaviour defined in a particular way, and thinks of true score as being the complete universe score that would be approached if the sample could be increased without limit. The difference may seem a somewhat trivial one, but it does make it easier to recognize that we may define different universes to which we may wish to generalize, and that different procedures for collecting test reliability data do imply different universes. As a corollary, it leads us to recognize that 'error' is not a monolithic entity, but involves a number of distinguishable components arising from different sources that can be separately identified and separately analysed. This leads to a components-of-variance model for test scores that encourages us to tease out the several sources of variance, other than between-persons variance, to estimate the magnitude of each, and to plan a data-collecting strategy for future research that will hold these unwanted sources of variance to a minimum. This view of the generalizability of test scores has been most systematically explored by Cronbach and his associates in their 1972 book, *The Dependability of Behavioural Measurements*.

At the same time that a model was being developed for test measurement error, models were also being formulated for what was being measured by true score on tests of different but related functions. The initial formulation was Charles Spearman's, as was the initial formulation of true score and error. It was phrased in the now long-familiar terms of general ability, *g*, pervading a whole set of cognitive measures, and a specific factor, *s*, for each separate measure. This conception of the nature of abilities proved less durable than the true-score-and-error model because the accumulation of empirical data soon showed that it was an over-simplification of the manner in which abilities are organized, and that tests are tied together much more complexly than by a single general factor. Over the past 50 years, many different models have been proposed to account for the observed correlations among test scores. Some have postulated a number of distinct general ability factors. Some have called for the inclusion of group factors, less pervasive than the general ones. Some have introduced second-order factors to make it possible to admit general factors that were not independent but rather were related to each other.

A massive body of statistical theory and computational technique has developed to service the analysis of voluminous sets of test score data that have been gathered to elucidate the nature and relationships of human abilities. Yet the general finding remains that a wide variety of cognitive performances are in fact all related; the view persists that a model must provide some role for a general component of ability; and we can still view tests such as the Binet and its many descendants as devices to estimate an individual's status on one widely relevant cognitive ability. This is not the place to examine in any further detail the structure and substance of different factor analytic models that use the interrelationships of different tests as evidence on which to build a theory of human abilities. Rather, let us turn to the scores on a single test, and to the responses to a single test exercise, and see by what model these may be understood.

The early tests that aspired to measure human abilities were composed largely of exercises that required the examinee to produce or construct the responses. Responses were varied, and each was then scored using some scoring guide that indicated the degree of acceptability of different possible responses. Exercises were developed primarily on the basis of editorial judgment, and little by way of psychometric theory or statistical analysis was applied to the selection of single test exercises or the combination of them into a test.

However, at the time of and shortly after World War I, there emerged in the United States an enthusiasm for selective-response items—true-false or multiple-choice exercises in which the examinee was required to select from among those that were presented to him the correct or best answer. In part because of the pre-established keying of the items, in part because of the susceptibility of such items to ambiguity because of unskilled drafting by the item's author—resulting in unintended interpretation by the examinees of one or more of the response alternatives, preliminary try-out and statistical analysis of the test items became the accepted practice. During the 1920s and 30s a great diversity of statistical indices was proposed to express an item's effectiveness in differentiating the more from the less capable examinees—capable in terms of what the set of items was intended to measure. By 1940, procedures had pretty much stabilized on the correlation, biserial or point-biserial, between item and a total score designed to represent the attribute to be measured by the test. Facility, as represented by percentage of correct responses, and discrimination, as represented by correlation with total score, became standard working criteria for evaluation and selection of test items. Psychometric theory was extended to formalize the relationship between item parameters and test parameters, showing how test parameters can be controlled by a judicious selection of test items. This state of the art prevailed from the time the American Council on Education published in 1951, under E. F. Lindquist's editorship, the first edition of the handbook entitled *Educational Measurement* up untii about the time that I 'rode herd' on the second edition, which came out in 1971.

At the theoretical level, the conception of the nature and function of a test was continuously evolving. As I view the scene, there are at the present time two major competing models to describe what a test score represents, and a number of variations of one of these. The two major models may be designated the 'domain sampling' and the 'latent trait' model. Let us take a look at each of these in turn.

As its label suggests, the domain sampling model starts out with the assumption that there is some definable domain of knowledge, understanding, or skill. In its clearest form, the domain is limited in scope and is precisely defined. Such a domain could be comprised of something like the 100 basic multiplication combinations, or the use of commas in series, or the capitalization of proper names. The domain is viewed primarily as having a certain horizontal extent, with minimal attention to its vertical dimension. The exercises in a test are viewed as being a sample from the defined domain, usually a random sample, though sometimes a stratified sample. The appropriate inference from a test score is considered to relate to the proportion of all the test exercises that might be drawn from that domain on which the individual would be expected to succeed – the individual's completeness of mastery of the domain.

This model has had a good deal of popularity, in the United States at least, over the past 20 years. Tests based on the model are often called 'criterion-referenced tests', to contrast them with traditional norm-referenced tests. When used most appropriately, each of these tests does focus on a highly specific domain, and gives information that is immediately relevant to a specific instructional decision. Has the pupil, or the class, reached a level of proficiency on Skill A that means that further teaching of that skill is not required and that Skill A is available as a foundation for teaching Skill B? In practice, however, the model has been called upon to rationalize much more general assessment and evaluation purposes, such as evaluation of the strengths and weaknesses of a pupil's development or of a school's curriculum. Because of its wide current use in present day edncational measurement, it is important that we examine this model to see what its assumptions are and when and to what extent they are justified.

The basic assumption of any sampling enterprise, whether the universe sampled is one of persons or one of test exercises, is that one can define a universe of objects or events with sufficient completeness and precision so that one can decide in each instance whether a specimen is a member of the defined universe and whether a set of specimens constitutes a representative sample from that universe. The safe procedure for drawing a representative sample is to enumerate all the specimens that comprise the universe and to use some system of randomization to draw a sample from among the enumerated specimens. Clearly the segments of educational achievement, for which such an unambiguous definition, complete enumeration, and random sampling is possible, are relatively few and of relatively limited importance.

Perhaps we can agree that a universe specified as 'Presented with two single-digit numbers in the format 2 × 3 and asked to give the product, gives the correct answer' does consist of 100 enumerable elements, and that some system of random numbers could satisfactorily be applied to draw a random or a stratified sample from this universe. For how many of the significant competencies that education strives to produce is such a procedure possible? We may feel that there does exist such a domain as 'ability to spell' or 'knowledge of biology', but how practical is it to specify the boundaries of such a domain or to enumerate the elements that fall within it? If we cannot specify the boundaries of the domain, how can we be sure that the sample of tasks included in our test covers the full scope of the domain? If we cannot enumerate the elements that fall within the domain, how can we know whether we have sampled randomly from them? And if we are not able to specify the limits of the domain, in what sense is it meaningful to say that a pupil has mastery of it? Within narrow limits it may be reasonable to say that, as of this Friday, Mary shows mastery of the 20 words in this week's spelling lesson, or that, tested at some particular time, John shows mastery of subtraction problems with zero in the minuend. But, for most of the range of significant educational achievements and for any tests that are thought to assess general abilities, the assumptions of a domain sampling model are met at best only roughly and approximately.

We do, of course, prepare course outlines and syllabuses that sketch in broad outline the content and objectives of a program of study. Long before educators began talking about criterion referenced as opposed to norm referenced tests, makers of educational achievement tests used such outlines to guide them in planning the coverage of their instruments. In that sense, there is no conflict between the old-line norm-referenced and new-style criterion-referenced approach. The issue is whether the domain to be tested is sufficiently describable and specifiable for one to be able to assert that it has been sampled in toto in a representative way, and so that some percentage of test items answered correctly can be considered to represent effective mastery of the domain. For much of what we are interested in assessing, this does not seem to be the case.

The alternative to a model based on a domain with lateral extent is one that focuses on a trait dimension, perceived as primarily vertical. In this model, the function of a test is conceived to be to estimate an individual's location on that vertical dimension – not 'How many?' from a domain of tasks, but 'How much?' on a dimension representing a trait. The trait or attribute is a construct rather than an observable. From the psychometric point of view the attribute assessed by a test need not be psychologically simple, though from the psychological viewpoint this would make for clarity. The test tasks may involve quite a complex of functions, but the latent trait model assumes that, to a reasonable approximation, the complex is the same for all the test exercises that make up the test.

The vertical latent trait model seems most obviously appropriate for test tasks that vary widely in difficulty but relatively little in kind – for example, responding to analogies of increasing subtlety, comprehending reading passages of increasing complexity, remembering lists of increasing length. For such attributes one tends not to worry very much about providing precisely defined horizontal boundaries to the trait. Rather the trait is roughly defined by the content and form of the tasks, and perhaps by the factor structure of the task as its relationships to other types of test exercises are studied by factor analytic procedures. Nor are there definable vertical boundaries, because it is usually not possible to specify limits on the facility of the task at its easiest or the difficulty of the task at its most difficult level. A test's function then is to locate each person on a vertical scale of indefinite extent in relation to anchor points provided by tasks scaled for their difficulty or by the performance of his fellows.

When the focus of our concern is educational achievement, the latent trait model is vaguely unsatisfying, because it seems somewhat incongruous to speak of a trait of, for example, competence in history. The unifying attribute seems to belong to the domain of knowledge rather than to the individual examinee. But usually there are no sharp boundaries to the domain: test exercises relating to it *do* differ substantially in difficulty or, at least, in the probability that students will succeed with them; it *is* possible to arrange people on a continuum on the basis of their ability to succeed with tasks drawn from the domain and to arrange the tasks from the domain in a continuum with respect to the likelihood that a person will succeed with that task. So a dimension of 'competence in history' is perhaps one on which different individuals can be placed at different levels. Thus, in broad-gauge measures of achievement, as well as in measures of aptitude (and interest or attitude), it seems plausible to think in terms of a dimension of performance upon which a particular individual may be high or low.

We turn now to alternate models for thinking of a test, together with the items that compose it, as a device for locating individuals on the continuous scale of some latent attribute, where we think of the scale as representing degree or level of that attribute. Our attention must focus first on our model of a test item.

Whereas in the domain sampling model we viewed the test exercise as one element drawn from that domain, and passing the test item as evidence increasing the proportion of the tasks in that domain that the examinee was considered to have mastered, we now view the test exercise as providing a cue as to where on the scale of the latent attribute the individual falls. If he passes the item, the chances are that he falls above the difficulty level that is exemplified by the item. If he fails the item, the chances are that he falls below that level. But it is only a matter of probability, because our model indicates that the likelihood of success on the item increases gradually and continuously as we move up the scale of the latent attribute. Our model specifies a probability function that is a continuous function of the latent attribute—typically the cumulative normal ogive or the logistic, two curves that have nearly identical properties.

Different items may differ in one or all of three parameters that describe these functions. These are, respectively:

1 a parameter that represents the steepness of the function, the *rate* at which the probability of success increases as one goes up the scale of the attribute;

2 a parameter that specifies the location of the function's point of inflexion in relation to the scale of the attribute, representing the difficulty of the item—the level at which just half the examinees are deemed to know the answer;

3 a parameter that represents the lower asymptote for the item—the probability of success for persons at very low levels on the latent attribute.

Current expositions of latent trait theory usually adopt one of two contrasting positions with respect to the role of these three parameters. One school of thought, represented by Rasch and his followers, elects to assume that the steepness parameter may be considered to be uniform over all items and that the lower asymptote may uniformly be considered to be zero. Thus items are deemed to differ only with respect to their difficulty. Then one need only estimate item difficulties, that is, where the inflection point for the item falls on the scale of the latent attribute, to characterize that item fully. The scale values of the items in a set then enable one to estimate scale values for each possible total score based on that set of items. The contrasting school of thought, led by Fred Lord at the Educational Testing Service of Princeton, New Jersey, would contend that one should undertake to estimate all three parameters for each item, and should use the full set of item characteristic curves to estimate the scale values corresponding to different scores, and consequently to the location of different individuals. Let us examine some of the virtues and some of the shortcomings of each of these approaches.

The Rasch one-parameter model certainly has the advantage of simplicity. A first approximation to the necessary scale values can easily be calculated with a hand calculator. With only a single parameter to be estimated for each item, the estimates have some prospect of stability even with pupil samples of modest size, which is to say in the hundreds rather than in the thousands. As Ben Wright has pointed out, in scoring the typical test we *do act* as if all the items had a common steepness parameter – that is, the same correlation with the underlying attribute. We act this way in that we combine items with equal weights and do not try to give greater weight to the more discriminating items – those items with greater item-trait correlations. Furthermore as the process of item selection during test construction has ordinarily weeded out those items with definitely low item-trait correlations, the ones that will show a much flatter item characteristic function, the range of item-trait correlations is reduced a good deal in those items that survive preliminary screening and make it to the final form of the test. Perhaps we do not strain reality too much if we assume that the slopes are all equal.

However, in many instances, the model *is* an oversimplification of reality. We do find that some variation in the steepness parameter does remain in chosen items. For example, when item discrimination indices were compared for two separate groups of pupils (over 2000 in each group) that had been tested with our Cognitive Ability Tests, the correlations from one group to the other of indices within a sub-test ranged from 0.88 to 0.93 with an average value of 0.90. Clearly the items were not all equally saturated with the common attribute that they were measuring. Furthermore, with multiple-choice and especially with true-false items, the assumption of a zero asymptote at the low end of the ability scale is almost certainly incorrect. Though many persons of low ability will omit an item where they do not know the answer, many others (at least in the USA) will guess and, unless the item writer has been more than usually skilful, that guess may be restricted to the two or three more appetizing options. Consequently all examinees are likely to have a considerable probability of hitting the correct answer.

Thus the one-parameter model provides an approximation that may be pretty rough in some circumstances. It seems most defensible in the case of constructed-response items in which the examinee must generate the response, and for tests composed of items that have survived a rigid empirical pre-screening. Especially for these, its computational and conceptual simplicity can make it quite attractive.

The three-parameter model treats as real and significant the differences in steepness and in asymptote of item characteristic functions, as well as differences in difficulty. With the resulting large number of parameters to be estimated for a set of items, really large data samples are called for if the estimates are to show satisfactory stability from one sample to another. Thus use of this model in a practical setting makes sense only for tests in which try-out of the items on large samples – of over 1000, if one accepts Lord's recommendations – is a practical possibility. Alas few of us are so situated. Furthermore estimation calls for high-speed and high-capacity computing facilities. When all of these conditions are met, and especially with the widely used multiple-choice format, it seems likely that the more complex model will provide a better description of each item, even if that information is only partially used in decisions related to or based on the set of items.

One assumption basic to both of these latent trait models is that the parameters of a test item are invariant, depending only upon the properties of the item and not on the group to which it is administered. If an item that is easy, relative to other items, in one group is difficult in another, then any general statement about the difficulty level of the item per se is meaningless. Uniformity seems most likely to prevail for items that depend primarily upon level of maturity and upon broad general background. Problems seem most likely to arise with items that are based upon specific school instruction, especially when that instruction is likely to vary widely from one place to another. Thus difficulty of an item calling for selection of the prime numbers from the set 31, 33, 35, 37, 39 is likely to be *very* much less for a group of 10-year-olds who have just started a unit on prime numbers than for a group who has never received such focused instruction or has received it in the more remote past. On the other hand, difficulty of a matrices or a figure analogies problem, relative to other items of the same type, seems likely to be relatively stable from group to group. We conclude, then, that attempts to express a person's status by the level at which that person falls on a vertical trait dimension is most defensible for ability measures that reflect general growth and the broad range of common experiences.

How will latent trait models influence our practical procedures of test development and our interpretation of test scores? How will we proceed differently from the way that we have in the past when we relied upon item difficulty and item discrimination indices? These questions will be addressed in detail by some of the later speakers at this seminar. I will only hazard a couple of guesses.

For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. The essential feature of a latent trait model is that test score is interpreted as a scale value on the vertical scale of the latent trait, rather than being expressed in normative terms in relation to some reference group of persons. It is more than 50 years now since Edward L. Thorndike and his associates developed the *Intelligence Scale CAVD* as an effort to express level of intellectual performance in an equal-unit vertical scale, but this scale never achieved any great measure of popular acceptance in the testing community. Normative reference rather than absolute scaling has always seemed more meaningful and useful, and I doubt that this will change.

There is, however, one field in which efficient use of item parameters to estimate the individual's location on a trait dimension is likely to prove crucially important. This is in individualized testing, whether by human examiner or by computer. When it becomes important to obtain the greatest possible amount of information about an examinee in a limited amount of testing time, then it is vitally important that each test exercise be one that will yield the maximum amount of information about that examinee. These are the items on which we start out with the greatest amount of uncertainty about whether the examinee will get the item right. They are items for which the item characteristic function is steepest at the ability level where the examinee is believed to lie. Adaptive testing, which progressively refines the estimate of an examinee's status after each item and picks the next test exercise to match the current estimate of that status, is one field in which the item parameters of single items will play a key role.

It will also be true that accumulation of data on item parameters for large pools of items, to the extent that these parameters are stable from group to group and from time to time, will make possible great flexibility in test construction. Such a calibrated item pool will make it quite easy to generate alternate forms of tests, equivalent in difficulty, that can provide estimates of performance expressed on a common scale — estimates that will facilitate studies of growth and change, that will permit testing different candidates with different but equivalent sets of tasks, and that may have a range of other practical uses.

The practice of educational and psychological measurement has evolved gradually over the course of time. Our models gradually become more explicit, better rationalized, and, we hope, more accurate reflections of examinee test-taking behaviour and of the abilities that lie at the back of that behaviour. I have tried to describe to you the present status of certain models. It is highly appropriate at this point in time that we take stock of those models and see what they have to offer for our practice in the years ahead.

REACTANT STATEMENT

Barry McGaw

Professor Thorndike's knowledge of the traditions of educational measurement began, I suppose, at his father's knee. The richness of the subsequent work, and the experience which shaped it, is evident in the perspective which his review has provided for us. The review sets the various theoretical and applied traditions in perspective and leads us neatly to the concluding invitation to take stock of the measurement models now current His paper thus sets the context for our seminar without attempting to pre-empt other speakers with definitive projections of what might prevail and what might disappear from the contemporary body of measurement theory and practice.

I find myself encouraged by the analysis in the paper to commence this discussion with a first attempt at taking stock. The attempt, in classical theory, to provide a theory of measurement error was identified as a significant feature of the early developments. I wonder if that issue might not be helpfully pursued as an important point of comparison of the models.

Since the error variance could not be estimated directly from intra-individual variability on multiple measurements with the same instruments, an indirect means of assessment was required. Classical theory provided the means for using inter-individual variability over two occasions as the basis for estimating error variance. It has probably been unfortunate that the correlation coefficients from which the standard errors of measurements are calculated have been more dominant in the language used to describe the properties of tests and to judge the utility of the measurements they provided.

Speaking generally of a test's reliability can obscure the significance of the assumption that the test has a particular lack of precision which is constant, if not for all applications, then at least for all individuals on a particular application. The possibility of one individual's status not being assessed as precisely as another's is simply not accommodated.

Generalizability theory, as Professor Thorndike's review makes clear, offers a refined conception of error of measurement through more complete specification of the various sources of error. I have found this approach helpful, particularly in the development and use of observation schedules; but it needs to be pointed out that, although this approach allows the estimation of the magnitude of error attributable to different sources, it still presumes that the errors of estimation for all individuals' scores are the same.

I am reminded of a claim I heard recently that the most significant advances in the design of sailing vessels occurred *after* the introduction of steamships. I now wonder whether the refinement of error estimation provided by generalizability theory, within the general framework of classical theory, cannot be seen similarly as the product of zealous attention to detail in a dying cause. I wonder where weekend sailors or designers and owners of 12 metre yachts attempting to win the America's Cup would force me to push that analogy. Sailing boats have not disappeared despite their loss of utility.

Latent trait models provide the means of more precise analysis of error variance. The magnitude of error variance depends on the precision with which the available items can locate each individual's ability. Those whose ability is in a range poorly covered by items are less precisely measured than those for whom there are more items with difficulty levels close to their level of mastery. A simple analogy can be drawn from a measuring stick on which the fine divisions have been erased in some sections.

Given this capacity for more precise specifications of measurement error for each level of measurement, should we not abandon the earlier conceptions, albeit with due acknowledgement of their historic contribution? Professor Thorndike's explication of the historic developments encourages me to press the case for latent trait models to this extent.

How should we then deal with criterion-referenced measurement? Its development was an attempt to free measurement from the circularity of norm-referenced measurement but perhaps an ultimately futile one. Professor Thorndike does speak of both domain sampling and criterion referencing. I would preserve the distinction between them but now refer only to the latter. Items which have been selected from a domain, if they vary in difficulty, can be located on a vertical dimension. The predilection of criterion-referenced measurers for worrying only about whether individuals are above or below some point does not remove the underlying continuum or reduce the value of a fuller view of it. It is a latent trait, and more precise location of individuals on it is surely better than determining only whether they are in one region or another. Can I claim encouragement in Professor Thorndike's analysis for asserting that in criterion-referenced measurement we have another sailing vessel, perhaps even one designed by a Thor Heyerdahl who would eschew not only steam for propulsion but even wood for the hull. I will let those at the ACER who purport to divide the world into literates and illiterates or numerates and innumerates defend themselves.

In characterizing the differences between the main approaches to latent trait measurement theory and practice, Professor Thorndike has set out clearly for us the differences in model complexity and ease of application.

I would like to highlight an important ideological difference which forces me to address not only questions of utility in judging the alternatives.

The three-parameter model is the natural extension of the earlier classical theory. It accommodates the baggage of classical theory—item difficulty and item discrimination and, with the addition of the third parameter, even accommodates the US love of the multiple-choice item.

The proponents of the three-parameter model like to see the one-parameter model as the degenerate version of their own more complex model, perhaps to be used on days when they are feeling simple-minded. That, however, inadequately acknowledges the fundamental view of the developers of the one-parameter model.

They do not include the second-parameter, for item discrimination, because it will accommodate the possibility of individuals of a given ability having a greater probability of being correct on a harder item than an easier one—at a point where the item characteristic curves have crossed. The second-parameter brigade says that items in the world are like that so their characteristics should be represented. The one-parameter brigade says that measurement cannot occur with such instruments so such item behaviour should be identified as failure to fit the model. The battle is thus joined by proponents of the one-parameter model on the ideological ground of beliefs about the nature of measurement.

On the grounds of utility, the battle is joined by supporters of the three-parameter model in terms of questions such as those Professor Thorndike outlined. When, for example, is it worth the effort and the extra data required to extract three parameters?

Proponents of the one-parameter model, however, do not see these as meaningful questions. Once they have stopped asking why one would ever be justified in including more than one parameter, they seek to demonstrate that, with dichotomous responses, there is insufficient information to estimate more than one parameter. For them, it is only by the sleight of hand, which arbitrary constraints in computer programs can conceal, that second and third parameters can be estimated.

Professor Thorndike's overview stimulates me, then, to start the discussion by asserting that the early classical formulations of reliability and their extension to generalizability can be dispensed with, that criterion-referenced measurement as originally introduced led us up a blind alley, and that the measurement theory debate now is lodged in the latent trait domain. The battle is joined by one group claiming that the other's view represents just a special case of its own more complex one. The others, in their turn, claim that the complexity their opponents prefer ought not to be sought but, if sought, cannot be quantified.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

2

Comparing Latent Trait with Classical Measurement Models in the Practice of Educational and Psychological Measurement

John A. Keats

In comparing latent trait theory with true score theory one may apply the criteria of either to the other. This practice will, of course, favour the theory which emphasizes the criteria being used. For example, some have criticized the Rasch model by showing empirically that selecting items in terms of this model will not necessarily produce the most reliable test of a given size from the item pool available. The result is predictable from the models as the true score model and associated methods tend to maximize reliability in some sense whereas the Rasch model requires different characteristics of the items. Alternatively one can argue that, since the relationship between true score and underlying ability is not readily specifiable for most tests, the true score model has defects. Such an argument is based on the criteria which form the basis of latent trait theory.

In the early 1950s, Gulliksen (1950) produced a synthesis of various contributions to true score theory and Lazarsfeld (1950), Lord (1952), and Arbous and Kerrich (1951) produced applications of latent trait theory to attitude scaling, test theory, and accident proneness respectively.

Specifically the model described by Arbous and Kerrich is based on the relation between the proportion, $p(x)$, of subjects having x accidents in a given time and a hypothesized latent attribute, accident proneness (λ) such that:

$$p(x) = \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!} \, dF(\lambda)$$

where $F(\lambda)$ is the cumulative frequency distribution of λ in the population studied. These writers are careful to point out that success in predicting the values of $p(x)$ only makes the accident proneness assumption more plausible but there are other possible interpretations.

In his treatment of the normal ogive latent trait ability model, Lord (1952) assumes that the probability of subjects of ability θ giving the correct response to item g may be written as:

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta-b_g)} \phi(t)dt$$

where a_g and b_g correspond to parameters of item g, $\Phi(t)$ is the standard normal frequency function, and t is a dummy variable which disappears in integrations.

Lazarsfeld (1950) explored the latent linear model which may be written as:

$$p_i = \int_{-\infty}^{\infty} (a_i + b_i x)g(x)dx$$

relating the proportion of subjects giving the correct response to item i, p_i, with the item parameters a_i and b_i, to the underlying variable x.

The true score model has its origins in physical measurement and the study of errors of measurement. It is assumed that a person's score (X_i) on a test may be regarded as consisting of two additive components: the true score (T_i) and the error score (E_i) so that:

$$X_i = T_i + E_i$$

The true score is thought of as a parameter of the person at the time of testing whereas the error score is thought of as due to minor fluctuations caused by irrelevant factors. Gulliksen (1950) produced a systematic account of this model. Ehrenberg (1953) criticized Gulliksen (1950) on the grounds that this account does not make clear what is being estimated and what the basis of estimation is. Rasch (1960) takes up these points and uses them as a justification for his latent trait model.

The questions of definition and estimation of true scores are taken up by Lord and Novick (1968) but their proposed answers were criticized by Lumsden (1978) and more systematically by McDonald (1979).† There appear to be some unresolved questions related to the precise status of the concept of true score even within the context of the model itself. Despite these problems, there is a large and growing literature developing the true score model, particularly addressing the question of the frequency distribution of true scores. Keats and Lord (1962) presented the

† Personal communication, Newcastle, 1979.

beta-binomial model of the frequency distribution of true scores and raw scores. This model has been elaborated by Keats (1964; 1965) and Lord and Novick (1968) and more recently by Huynh (1977) in the context of mastery scores.

In its simplest form, the beta-binomial model assumes that the frequency distribution of number correct raw scores, $g(x)$, can be written in terms of the distribution of true scores, $f(p)$, in the following way:

$$g(x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} f(p)\,dp$$

If the regression of true score on raw score can be represented as a polynomial in x then:

$$g(x) = K \frac{(-n)_x (a_1)_x (a_2)_x \ldots}{x!(b_1)_x (b_2)_x \ldots}$$

where n = the number of items, $a_1, a_2 \ldots; b_1, b_2 \ldots$ are parameters to be estimated and

$$(a_1)_x = \frac{\Gamma(a_1 + x)}{\Gamma(a_1)} = a_1(a_1+1)(a_1+2)\ldots(a_1+x-1) \text{ etc.}$$

and the constant K, $g(0)$, makes the sum of the $g(x)$ values unity. If the regression of true score on raw score is linear, then:

$$g(x) = K \frac{(-n)_x (a_1)_x}{x!\ (b_1)_x}$$

and this form can also be obtained by taking $f(p)$ as a beta distribution, hence the name beta-binomial model. It is worth noting that defining true score in this way implies that it is a latent trait accounting for differences in raw score along the lines of the accident proneness trait defined in the Arbous and Kerrich model. However this is not the usual way of defining true scores. The latent trait ability models of Lord and later of Rasch define ability in terms of performance on an *item* rather than on the total test.

Although there appears to be no general systematic account of the latent trait model comparable to the treatment by Lord and Novick (1968) of the true score model, there are many recent developments of the model in the context of test theory. The main problem tackled is that of estimation of parameters and significance tests for the applicability of the model, e.g. Gustafsson (1979).

Most of the developments in the literature on both models are either statistical in content or deal with problems of application in education and psychology. There has been no attempt to compare the models in the more general context of, for example, cognitive development. Rasch's (1960) claim that his model leads to the estimation of a person *parameter*, for example an ability, can only be justified in the context of *adult* behaviour in which the ability of a person may be thought of as approximately constant over a long period of time. When subjects under, say, 15 years are considered, this 'parameter' as defined and estimated at a particular time will not be the same a year or even six months later. It is important to use the Rasch approach to define the dimension measured by a set of items, but any measure of a person at a given time is only meaningful if it can be related to the value towards which the person is developing. It is this asymptotic value that constitutes a parameter of the person.

A number of writers, Courtis (1933), Bayley (1955), Cattell (1971), and Jensen (1973), have attempted theoretical formulations of cognitive development. Courtis proposed a two-parameter individual model which involved difficult problems of estimation but did not suggest a form for the group curve. Bayley attempted to construct an empirical group cognitive growth curve but encountered difficulties in establishing an appropriate unit of measurement. Anderson (1940), Cattell (1971), and Jensen (1973) separately proposed random accumulation models, with Jensen suggesting a separate consolidation parameter which differed from person to person. These attempts seem to be trying to account for certain known facts of cognitive development separately, namely:

1 Average ability seems to develop at a negatively accelerating rate towards a stable level (Bayley).

2 Individuals develop at different rates towards different stable adult values (Courtis and, in part, Jensen).

3 The older the person becomes, and therefore the closer he or she approaches the adult value, the more stable IQ measures become when taken at yearly intervals. This fact is reflected in the simplex pattern obtained when standardized ability measures at a number of successive age levels are intercorrelated. As pointed out by Anderson (1940), Cattell (1971), and Jensen (1973), such a simplex can be generated by means of random accumulations over time. Jensen's suggestion that people differ in the extent to which they can consolidate these accumulations leads to growing differences between subjects as they become older but these differences become more stable.

These three general findings are jointly consistent with a developmental model with at least two individual differences parameters as well as a measure of ability on a ratio scale and a measure of time since development started. However, for the sake of clarity and system, a developmental model with one individual differences parameter will be examined first and its strengths and deficiencies noted before proceeding to a two-parameter model. A further reason for studying the one-parameter model is that the use of the IQ in one or other of its forms implies a one-parameter model of development. The possibility of more than two parameters will also be explored to the extent of showing how data may be examined to justify more than two parameters. Within the one- and two-parameter models, the applicability or otherwise of a latent trait model or a true score model will be examined. It is important to note that the concern here is to model developmental change, not simply to measure change in the absence of a model. The question of the unreliability of change measures is not relevant to this discussion.

THE PRINCIPLE OF DYNAMIC CONSISTENCY

In discussing key problems associated with formal theory in psychology, Marx (1976) lists the problem of individual versus group functions and notes: 'It is now generally agreed that in terms of mathematical functions representing behaviour processes, data obtained from groups cannot be freely used for individual function' (p. 255). Neither he nor apparently anybody else is sufficiently disturbed about this state of affairs to suggest that it should be a key principle in mathematical models of behaviour that the form of the relationship between behavioural measures, stimulus variables, and individual differences variables should be the same at the group level as it is at the individual level.

If this principle does not apply then the form of relationship obtained with group data may well vary according to the particular sample of individual difference parameters present in the group. Under such circumstances it is hard to see how general laws based on group data can have any general validity at all. Sometimes the principle can be built into the model by specifying how the individual data are to be aggregated into a group function. If this can be done it seems important that it should be done. Consider the basic equation of the Rasch latent ability model (Rasch, 1960) relating the proportion (P_{ij}) of subjects of ability (A_i) giving the correct response to a particular item, to the difficulty (D_j) of the item

$$\text{i.e. } P_{ij} = \frac{A_i}{A_i + D_j}$$

If one considers a test of n items with values of D_j, $j = 1 \ldots n$, not all equal, the true score value corresponding to ability A_i is $nP_{i.}$, where $P_{i.}$ is the mean value of P_{ij} averaged over all n items for subjects of ability A_i. The relationship between $P_{i.}$ and A_i will depend on the distribution of the values of D_j but will *not* in general be of the form:

$$P_{i.} = \frac{A_i}{A_i + D.}$$

where $D.$ is the arithmetic mean of the values of D_j. Thus true score will never be related to underlying ability as defined by this model in the same way as the individual items of different difficulty which generate the score. This conclusion generalizes to all latent ability models using nonlinear item characteristic curves and unequal item difficulties. In the case of the Rasch model it may be noted that, if the harmonic mean of the P_{ij} values, $H(P_{i.})$ were used to define true score,

$$H(P_{i.}) = \frac{A_i}{A_i + D.}$$

which is of the same form as the individual item curves. Of course there would be problems in defining true score in this way.

The above criticism of the true score model can be objected to on the grounds that it arises from the latent trait approach and that one may wish to consider true scores and their frequency distribution without assuming an underlying ability. However application of the principle of dynamic consistency makes it clear that this principle could never be satisfied by true scores for any realistic form of item characteristic curve unless all items have the same characteristic curve. In this extreme case, Birnbaum (1968) has shown that the number correct score is a sufficient statistic for estimating ability irrespective of the form of the item characteristic curve, providing it is monotonic.

In what follows, the principle of dynamic consistency will be regarded as fundamental so that it is possible to discuss the form of individual and group cognitive development curves without obtaining inconsistencies.

THE ONE-PARAMETER COGNITIVE DEVELOPMENT MODEL

The purpose of this model is to relate ability A_{ik}, measured on a ratio scale, at time t_k to time t_k as a variable with one individual differences parameter, c_i and a scaling parameter which will be shown to depend on the units in which ability and time are expressed. For reasons given by Halford and Keats (1978), t_k will be taken from birth. For the sake of simplicity as well as other advantages noted below, the form of the relationship will be assumed to be:

$$A_{ik} = \frac{t_k}{c_i t_k + d}$$

$$\text{or} \qquad 1/A_{ik} = c_i + d/t_k$$

It follows that as t_k increases $1/A_{ik}$ approaches c_i more and more closely. Thus the individual differences parameter, c_i, is a measure of the asymptotic value towards which the ability of subject i approaches as he grows older. It follows that the units u_a in which $1/c_i$ is expressed are the same as those in which A_{ik} is expressed.

It also follows that $H(A_{.k})$ the harmonic mean of A_{ik} over individuals at age t_k, can be expressed as:

$$1/H(A_{.k}) = \bar{c} + d/t_k$$

$$\text{or} \qquad H(A_{.k}) = \frac{t_k}{\bar{c}t_k + d}$$

which is of the same form as the individual curve with parameter $\bar{c}$, the arithmetic mean of the individual differences parameters for the group defined. The group curve, defined in terms of the harmonic mean, satisfies the principle of dynamic consistency.

It further follows that $t_{\frac{1}{2}}$, the age at which the group curve reaches half of its asymptotic value $1/\bar{c}$, can be expressed as:

$$t_{\frac{1}{2}} = d/\bar{c}$$

$$\text{or} \qquad d = t_{\frac{1}{2}}\bar{c}$$

This latter expression indicates that the constant d is expressed in units of $u_t u_a^{-1}$ in dimensional analysis terms. Thus d is a constant which varies according to the units in which time and ability are measured.

The cognitive development model can be related to other measures of ability in common use. For example, the mental age measure used by Binet (1908) can be obtained by noting that:

$$t_k = \frac{dH(A_{.k})}{1 - \bar{c}H(A_{.k})}$$

For a subject with ability A_{ik}, at age t_k, mental age t_m, may be obtained from the formula:

$$t_m = \frac{dA_{ik}}{1 - \bar{c}A_{ik}}$$

provided $A_{ik} < \dfrac{1}{\bar{c}}$

By substituting for the value of A_{ik} one obtains:

$$t_m = \frac{dt_k}{(c_i - \bar{c})t_k + d}$$

Thus if ratio IQ is defined as:

$$\mathrm{IQ}_R = 100 \frac{\text{Mental age}}{\text{Chronological age,}}$$

then:

$$\mathrm{IQ}_R = \frac{100d}{(c_i - \bar{c})t_k + d}$$

IQ_R is a dimensionless constant. For subjects with asymptotic value $1/\bar{c}$, equal to that of the group curve in terms of which mental age is defined, then $\mathrm{IQ}_R = 100$ and is the same at all ages. For subjects whose asymptotes are greater than $1/\bar{c}$, the ratio IQ will increase with age until it is undefined, i.e. $A_{ik} \geq 1/\bar{c}$. Similarly, for subjects whose asymptotes are less than $1/\bar{c}$, the ratio IQ will decrease with age. In general

$$\frac{100}{\mathrm{IQ}_R} = \frac{(c_i - \bar{c})}{d} t_k + 1$$

Two obvious criticisms of the ratio IQ arise from this formulation. The first has been known for many years (see, for example, Thurstone, 1926) and refers to the fact that mental age and therefore ratio IQ are undefined for subjects whose ability has reached or surpassed the asymptotic value of the group curve. In practice this difficulty has been overcome by ad hoc methods (see, for example, Terman and Merrill, 1937) but these simply confirm the breakdown of the model.

The second criticism does not appear to have been noted before. This criticism relates to the significant trend in ratio IQ with age for a subject whose asymptotic value departs from that of the group curve. The question of whether or not such trends occur can be answered in the affirmative from data published by Skodak and Skeels (1949, Table IX, p. 146). These data are for two groups of subjects whose natural mothers' IQs averaged 63 (Group A) and 109 (Group B). It would be expected that individual data would show considerable variability and this is so. However the equation

$$\frac{100}{\mathrm{IQ}_R} = \frac{(c_i - \bar{c})}{d} t_k + 1$$

can be averaged over two subgroups to obtain:

$$\frac{100}{H(\mathrm{IQ}_{RA})} = \frac{(\bar{c}_A - \bar{c})}{d} t_k + 1$$

and

$$\frac{100}{H(\mathrm{IQ}_{RB})} = \frac{(\bar{c}_R - \bar{c})}{d} t_k + 1$$

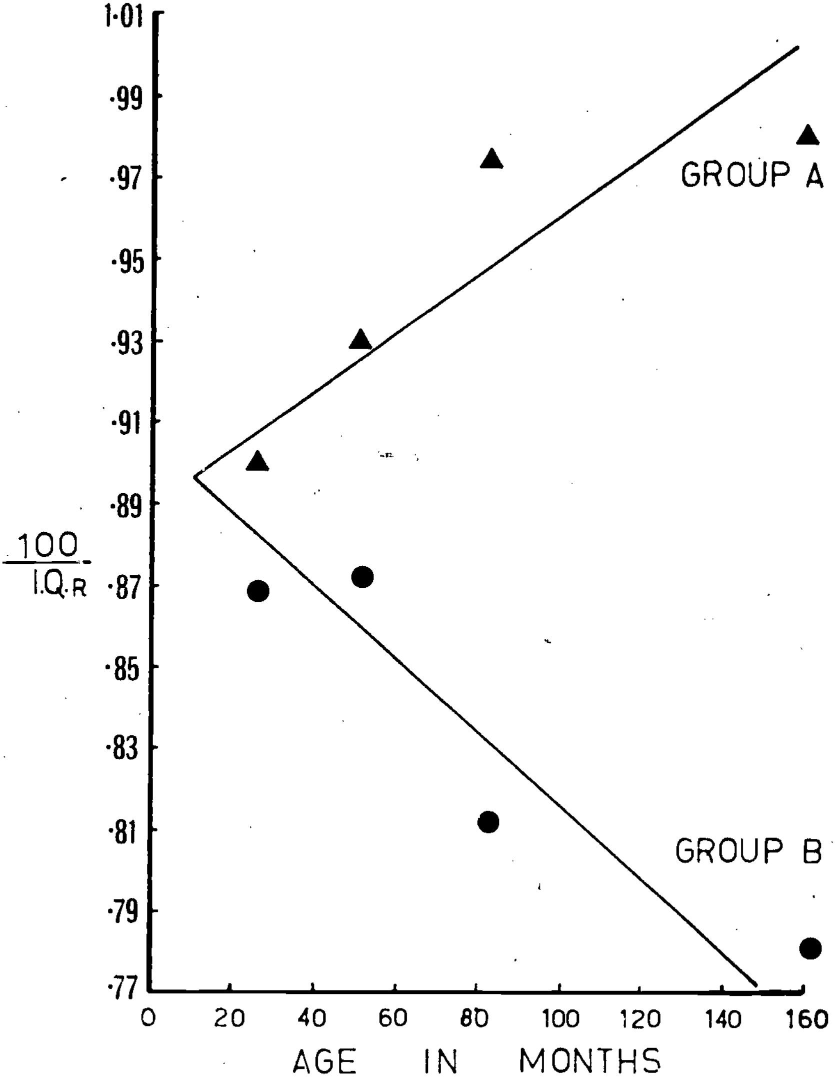


Figure 1 Changes in Ratio IQ with Age in Two Groups of Subjects

The corresponding graphs for group A and group B are given in Figure 1 and show clearly that for group A the slope is positive which implies that $\overline{c}_A > \overline{c}$ and so the asymptotic value for this group is *less* than the average asymptote as would be expected from the average IQ of the natural mothers. On the other hand, for group B the slope is negative which implies that the asymptotic value for group B is greater than average as would also be expected. Thus these data seem to confirm some of the predictions of the one-parameter model and imply that the single parameter, c_i, is at least in part influenced by heredity. On the other hand, the intercept on the y axis should be unity according to the model but this is far from the case. In fact, if individual graphs are drawn and fitted by straight lines in the usual way, almost all of the 19 graphs would have intercepts less than one. Thus the single parameter model is only partially confirmed. This point will be taken up again when the two-parameter model is discussed.

While the adherents to the true score approach were severely critical of the ratio IQ, they were appreciative of the fact that some transformation of number-correct score was necessary in most applications. In many cases the transformation recommended was that of standardizing to a fixed mean and standard deviation. Similar transformations of ability measures have been proposed so that individual values can be related to particular groups.

Because $(A_{ik})^{-1}$ is related to c_i and t_k in a linear fashion, it seems reasonable to define a deviation score or IQ$_D$ in terms of $(A_{ik})^{-1}$. It may be noted that $[H(A_{.k})]^{-1} = E[(A_{.k})^{-1}]$ and so

$$E[(A_{.k})^{-1}] - (A_{ik})^{-1} = [H(A_{.k})]^{-1} - (A_{ik})^{-1}$$

$$= \bar{c} - c_i$$

$$\sigma_{(A_{ik})^{-1}} = \sigma_{c_i}$$

and $$\frac{E[(A_{.k})^{-1}] - (A_{ik})^{-1}}{\sigma_{(A_{ik})^{-1}}}(15) + 100 = \frac{(\bar{c} - c_i)}{\sigma_{c_i}}(15) + 100 = \text{IQ}_D$$

Thus the standard score of the reciprocal of ability at any age level is equal to the asymptotic parameter expressed as a standard score with the same mean and standard deviation.

It follows that IQ$_D$ and IQ$_R$ can be related in terms of the model and the relationship can be simplified by choosing a scale for the ability variable such that $\sigma_{c_i} = 15$. With this convention

$$\text{IQ}_D = \bar{c} - c_i + 100$$

and $$\frac{100}{\text{IQ}_R} = \frac{1}{d}(100 - \text{IQ}_D)t_k + 1$$

While this relationship can be defined in terms of the cognitive development model proposed above and the corresponding definitions of IQ$_R$

and IQ_D, it is of interest to see to what extent this relationship holds for conventional ability tests. It is possible to estimate IQ_D and IQ_R for tests for which norms are available for a number of age levels. For example, ACER Test, Intermediate D, (1947–51) has norms for ages 10–14 years. According to these norms, raw scores corresponding to 100 IQ_D at each of the ages 10.0, 11.0, 12.0, 13.0, and 14.0 are 14, 23, 31, 39, and 46 respectively. Thus 10-year-olds obtaining these scores would have mental ages of 10, 11, 12, 13, and 14 years respectively and ratio IQs of 100, 110, 120, 130, and 140 but their deviation IQs would be 100, 110, 117, 123, and 129. Similar corresponding values could be obtained for 14-year-olds. With these data, $100/\mathrm{IQ}_R$ is plotted against $(\mathrm{IQ}_D - 100)t_k$ and the graph should be linear passing through the point 0,1. Figure 2 displays the graph obtained for the nine points available from these data. It would appear that the simplifications and approximations used in the model have led to a prediction which is not inconsistent with these data.

Although deviation IQs are never defined in this way, the actual values obtained would not differ greatly for the definition above.

If deviation IQ as usually defined were constant, the value obtained would obviously be a good predictor of adult performance. Because of unreliability in measurement, it is not sufficient to show instability in IQ_D to challenge the single parameter approach. Even systematic trend in IQ_D over age in certain subjects would not be sufficient. What needs to be shown is that systematic departures from constancy of deviation IQ are related to other variables.

Two relevant studies on this topic are those of McCall et al. (1973) and Hindley and Owen (1979). The former analysed longitudinal data expressed as deviation IQs for a large group of subjects. Between-subjects differences corrected for mean value were used to define clusters of subjects with similar patterns of change. The largest group (approximately 40 per cent) showed no systematic change over the 10 years studied. A second group revealed an increasing trend in IQ_D up to approximately 10 years of age followed by a decrease to the original value. At all age levels, the average IQ_D values were substantially greater than 100 for this group. Interviews with parents of all subjects were conducted focusing on child raising practices. Parents of this second group were typically 'accelerators', that is they attempted to stimulate the cognitive development of their children beyond or ahead of what was done at school. Other groups of children showed decreasing trends in IQ_D and parental practice in these cases tended to be either repressive or laissez-faire. The essential point is that systematic changes in IQ_D were associated with different kinds of child raising practices.

In the Hindley and Owen study (1979), similar data were analysed individually. For each subject significant departures from constancy, linear

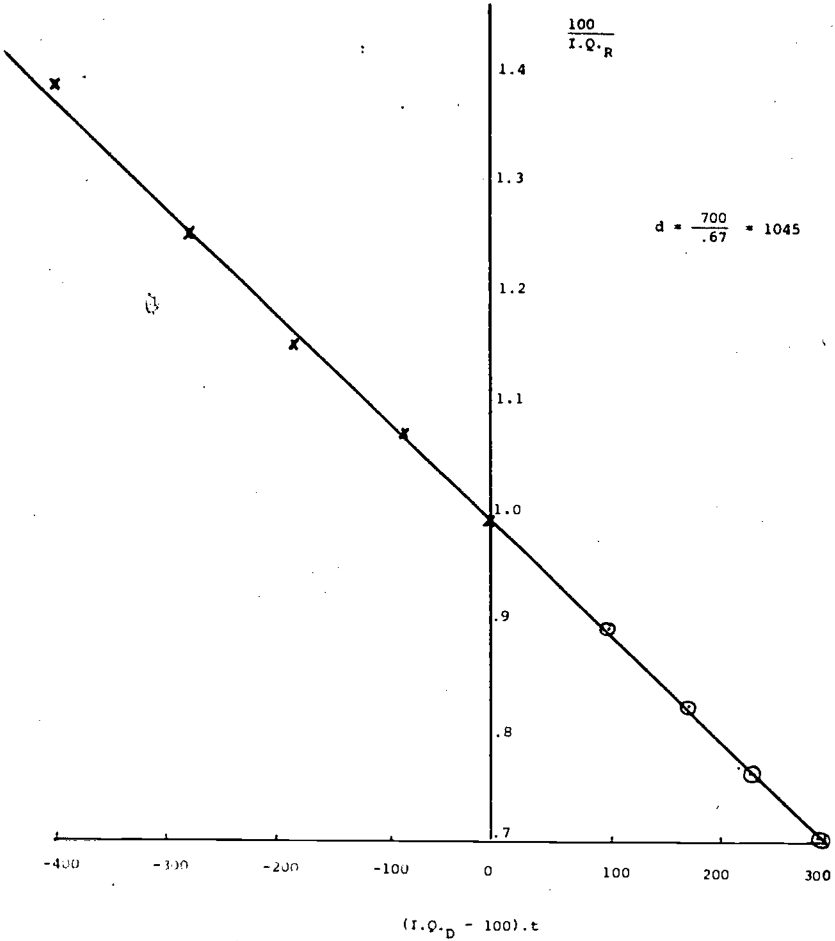


Figure 2 Relationship between Deviation and Ratio IQ on the ACER Intermediate D Test of Intelligence

trend, quadratic trend, etc. were tested for significance sequentially and parameters for linear, quadratic, etc. orthogonal functions estimated where appropriate. Parameters were averaged across subjects for each of three social classes and systematic differences noted. In particular, children with upper-class parents tended to show the same trend in IQ as those in the McCall et al. study with accelerating parents. The Hindley and Owen study is an important confirmation of the McCall study and both show that, for some subjects, IQ_D is variable and the variation is correlated with systematic environmental factors. Thus the one-parameter cognitive growth model which could define a deviation IQ which is constant with age does not account for at least some of the phenomena reliably observed.

The present model may be compared with the consolidation model proposed by Jensen (1973). This model accounts for the simplex structure apparent in correlation tables from repeated measures obtained in longitudinal studies. Jensen's version is explicated in some mathematical detail and is based on the notion of random cumulating increments over time. However the model does not yield an asymptotic growth curve and no way has been suggested for estimating the consolidating factor, F, from test data even though this is intended as the basic parameter determining adult level of performance.

According to Jensen's model, S_n, the performance of a particular subject after n time intervals, is given by:

$$S_n = F_i(G_1 + G_2 \ . \ . \ . \ G_{n-1}) + G_n$$

with G_k a random component from experience in some time interval, F_i the proportion of cumulative experience which is consolidated, and G_n the random component which is the result of current experience, unconsolidated.

If μ_g is the mean of the distribution from which the random components G are drawn, $\overline{G}_i$ the mean of the actual values for a particular individual, i, and t_k is the time measure corresponding to the intervals then

$$S_n = F_i t_k \mu_G + F_i t_k(\overline{G}_i - \mu_G) + \overline{G}_n$$

where $\overline{G}_i$ and μ_G are expressed as experience per unit of time. Thus S_n will increase approximately linearly with time at a rate dependent on F_i, the amount of consolidation.

As t_k increases, the proportion of random variation contributed by the second and third terms of this equation will decrease. Under these circumstances the correlation between S_n and S_{n-1} will increase with n as has been frequently observed in actual data in the form of deviation IQs.

In the case of the one-parameter model, it has been shown that Var $(1/A_{ik}) = \text{Var}(c_i)$ and is independent of age. Thus the correlation between values of $(1/A_{ik})$ at constant time intervals will not vary with age. The single parameter model does not predict the simplex structure.

REVIEW OF PROPERTIES OF THE ONE-PARAMETER GROWTH CURVE MODEL

1 This model does represent a growth curve approaching an asymptote.

2 Individual differences are reflected in the model in differences in the asymptote approached.

3 The model exhibits dynamic consistency when the group curve is defined in terms of the harmonic mean.

4 The model correctly predicts variation in ratio IQ for subjects with above average asymptote as well as for those with below average asymptote.

5 With regard to the above relationships, the model predicts an intercept of unity when $1/\mathrm{IQ}_R$ is plotted against time. The value observed for almost all of the 19 subjects studied was less than unity.

6 The model also predicts that a deviation IQ can be defined which will be constant across age levels and this does not agree with studies that show that environmental factors correlate with systematic variations in IQ_D.

7 The simplex pattern which led Jensen (1973) to propose his consolidation model would not be explained by the one-parameter model.

THE TWO-PARAMETER COGNITIVE DEVELOPMENT MODEL

The need for the introduction of a second individual differences parameter arises from the fact that not only do individuals differ in their ultimate adult level of ability but they also differ in the rate of approaching that level. It has been noted that $d/\bar{c}$ is equal to the age at which the group curve defined in terms of the harmonic mean reaches half of its asymptotic value, $1/\bar{c}$. Thus the parameter d is associated with the rate of development. In the one-parameter model it is implied that the larger the asymptotic value, $1/c_i$, the older the person will be when he or she reaches half of this value since this age equals d/c_i. However, if the parameter d is allowed to vary across individuals, allowance will have been made for the observed fact that individuals do differ in their *rate* of development. Thus the two-parameter cognitive growth model may be written as:

$$A_{ik} = \frac{t_k}{c_i t_k + d_i}$$

$$\text{or} \qquad 1/A_{ik} = c_i + d_i/t_k$$

$$\text{and} \qquad 1/H(A_{.k}) = \bar{c} + \bar{d}/t_k$$

$$\text{or} \qquad H(A_{.k}) = \frac{t_k}{\bar{c}t_k + \bar{d}}$$

and dynamic consistency is still preserved. It may readily be observed that the individual asymptotic value will equal $1/c_i$ and that half of this value will be reached at an age of d_i/c_i. The units of c_i and d_i may be shown to be u_A^{-1} and $u_A^{-1}u_t$ so that the units of d_i/c_i will be u_t as required.

By the same methods as those used in the one-parameter model, it may be shown that

$$\mathrm{IQ}_R = \frac{100\bar{d}}{(c_i - \bar{c})t_k + d_i}$$

$$\text{or} \quad \frac{100}{\mathrm{IQ}_R} = \frac{(c_i - \bar{c})}{\bar{d}}\, t_k + \frac{d_i}{\bar{d}}$$

Thus when $100/\mathrm{IQ}_R$ is plotted against t_k the intercept will no longer be unity unless $d_i = \bar{d}$. In the case of the Skodak and Skeels data, it was observed that the intercept was considerably less than unity for almost all subjects which implies that the rate of development for these children was considerably greater than average since $d_i < \bar{d}$.

The data from this study have been questioned, see for example Munsinger (1975), because of the high ratio IQs reported for most of the children. However, as might be expected on other grounds and is in fact confirmed in the Skodak and Skeels report, the home environments into which these children were adopted were considerably above average in terms of the cognitive stimulation provided for the children at a young age. Such enriched environments lead to a small value for the parameter d_i and a high value of ratio IQ which would tend to overstate the adult value.

When the two-parameter model is written in the form

$$1/A_{ik} = c_i + d_i/t_k$$

it is clear that the influence of the d_i parameter decreases with age. However the d_i parameter is the one associated with rate of growth which can readily be influenced by environmental factors, particularly at young age levels. Thus the correlation between annual ability measures at *young* age levels will be influenced by differences in d_i values as well as c_i values and so will tend to be lower than the corresponding correlation at older age levels which will depend increasingly on c_i.

As noted earlier, such correlations produce the characteristic simplex pattern of correlations observed in many studies and explained by some writers, e.g. Jensen (1973), in terms of random accumulation from experience and consolidation of these experiences. While such an explanation is possible, it is also clear that a process of environmental differences affecting rate of growth towards an asymptote which is influenced by a heredity component can also produce the observed phenomenon. However the current model can also account for the patterns of change in deviation IQ noted by McCall et al. (1973) and Hindley and Owen (1979) and associated with environmental differences.

AN INTEGRATION OF TRAIT THEORY, TRUE SCORE THEORY, AND THE COGNITIVE GROWTH MODEL

From the principle of dynamic consistency, it is clear that true scores can only be explicitly related to ability values when the condition that items have identical item characteristic curves is satisfied. This is an additional restriction to that imposed by the Rasch model and may not always be achievable in practice. However the theoretical and practical advantages of tests with equivalent items when prepared for a number of different age levels are so great that the additional initial effort may be worthwhile at least in some areas. Some of these advantages were noted by Keats (1967).

The first important property of tests with equivalent items is that the number-correct raw score (x) is a sufficient statistic irrespective of the form of the item characteristic curve (Birnbaum, 1968, p. 429). This is obviously an important property which enables one to classify subjects in terms of raw score without particular assumption of the form of the item characteristic curve.

The second advantage of tests with equivalent items is that the conditions for the binomial error model are met. Thus the regression, mean (p/x), of true score on raw score may be written in terms of the frequency distribution of x, $g(x)$ as follows:

$$\frac{g(x+1)}{g(x)} = \frac{(n-x)}{(x+1)} \, \frac{\text{Mean } (p/x)}{1-\text{Mean } (p/x+1)},$$

(Keats and Lord, 1962). Keats (1964) set out procedures based on this formula and the theory of orthogonal polynomials for testing the regression of p on x for significant linear, quadratic, cubic, etc. components. The resulting simple or generalized hypergeometric distribution of x may be written as:

$$g(x) = K \, \frac{(-n)_x (a_1)_x (a_2)_x \ldots}{x! \, (b_1)_x (b_2)_x \ldots}$$

where $$(a_1)_x = \frac{\Gamma(a_1+x)}{\Gamma(a_1)} = a_1(a_1+1) \ldots (a_1+x-1)$$

and $$K = g(0)$$

In the special case in which only the linear component contributes significantly to the regression one may write:

$$\text{mean } \frac{p}{x} = \frac{r}{n} x + \frac{\bar{x}}{n} (1-r)$$

where r is the Kuder-Richardson formula 21. Various bi-variate distributions can also be specified as shown by Keats and Lord (1962) and Keats (1964).

If the regression of true score on raw score is to be constant irrespective of the age sample taken, then

$$\frac{r}{n} \text{ and } \frac{\bar{x}}{n}(1-r)$$

must be the same for each population. This condition will be met if:

$$\text{Var}(x) = \frac{\bar{x}(n-\bar{x})}{(n-r[n-1])}$$

where $\bar{x}$ and Var(x) relate to the same population and r is the constant value taken by KR 21 at each age level. If this condition is met, then cognitive growth could be measured in terms of true score. Failure to meet this condition would imply that the regression of true score on raw score is *not* linear for this particular test and age range.

Another advantage of tests with equivalent items is that, as Birnbaum (1968, p. 458) notes, the maximum likelihood estimate of ability, $\hat{\theta}$, may be written explicitly as:

$$\hat{\theta} = 1 + m \log \frac{x}{(n-x)}$$

if the logistic model is used. By appropriate choice of units and taking $\hat{\theta} = \log A/D$ one has:

$$\hat{A} = \frac{Dx/n}{1 - x/n}$$

$$\frac{x}{n} = \frac{\hat{A}}{\hat{A} + D}$$

where D is the common difficulty parameter of the items. Thus raw score has the same relationship to ability as do the individual items of which it is composed. If in addition true score has a linear relation to raw score, then true score may be related explicitly to ability. In any case the distribution of estimated ability values may be obtained from the distribution of raw scores, $g(x)$, even though the distribution of true scores is unknown. As many recent articles have indicated, there are difficulties in specifying the distribution of true scores if significant departures from the beta function are indicated in the data. One advantage of working with ability values rather than true scores appears clear from this analysis.

According to the two-parameter cognitive growth model, ability A is projective on time (t) with parameters ($1, 0, c_i, d_i$) and as first noted for the case of equivalent items and the logistic assumption, raw score is projective on ability with parameters (n, 0; 1, D). Thus raw score is projective on time with parameters (n, 0; $1 + c_iD$, d_iD); that is,

$$\frac{x}{n} = \frac{t_k}{(1 + c_iD)t_k + d_iD}$$

If two such tests of n items with difficulty parameters D_1 and D_2 are administered to the same subjects at approximately the same time, then the scores on one (x_2) may be chained to the scores on the other (x_1) by the formula:

$$\frac{1}{x_2} = \frac{1}{x_1}\frac{D_2}{D_1} - \frac{(D_2 - D_1)}{nD_1}$$

i.e. reciprocals of raw scores should be linearly equated.

MORE GENERAL COGNITIVE DEVELOPMENT MODELS

The models proposed so far are based on the assumption that $1/A_{ik}$ is linear on $1/t_k$ which leads to the projective relationship of ability on time. For various reasons the observed data may depart significantly from linearity. One obvious departure would occur if environmental factors changed to produce a dramatic change in the value of d_i. If d_i remained stable at the new value, the graph would consist of two or more line segments and could be analysed as such. More gradual and persistent changes in d_i however would produce significant curvature so that:

$$\frac{1}{A_{ik}} = c_i + d_i\left(\frac{1}{t_k}\right) + e_i\left(\frac{1}{t_k^2}\right)$$

might be a possible representation.

Such significant departures from linearity would make the task of predicting and estimating the asymptotic value extremely difficult. To what extent they occur can only be discovered by studies which investigate the need for a three (or more) parameter model of cognitive development. The data for a decision on this question are not available.

CONCLUSION

The present paper has attempted to review the usefulness of latent trait models as opposed to true score models in the more general context of cognitive development. It is clear that, even if only the difficulty of items in a test is allowed to vary, the true score model has difficulty in providing a useful representation of cognitive growth whereas the latent trait model can be readily extended to cover the cognitive growth phenomena. This is true even if cognitive development is much more irregular and idiosyncratic than evidence so far suggests. The evidence available at present is consistent with the notion that at least two parameters are required to represent cognitive development and these parameters could be estimated using two administrations of chained tests of the same ability at widely separated age levels. The most efficient method of estimation has not been explored here.

Although the latent trait approach is clearly superior in this context, it seems unfortunate that some of the advantages of the true score model, for example distribution models, have to be abandoned. Keats (1967) and Birnbaum (1968) noted some of the advantages of equivalent item tests. A further advantage noted here arises from the fact that such tests, and only these, produce a possible reconciliation of the two models.

REFERENCES

Anderson, J. E. The prediction of terminal intelligence from infant and preschool tests. *National Society for the Study of Education Year Book*, 1940, **39**, Part I, 385–403.

Arbous, A. G. and Kerrich, J. E. Accident statistics and the concept of accident-proneness. *Biometrics*, 1951, **7** (4), 340–432.

Australian Council for Educational Research. *ACER Intermediate Test D.* Melbourne: ACER, 1947–51.

Bayley, N. On the growth of intelligence. *American Psychologist*, 1955, **10**, 805–18.

Binet, A. and Simon, T. The development of intelligence in the child. *L'anee psychologique*, 1908, **14**, 1–90.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Cattell, R. B. *Abilities: Their structure, growth and action.* Boston: Houghton-Mifflin, 1971.

Courtis, S. A. The prediction of growth. *Journal of Educational Research*, 1933, **26**, 481–92.

Ehrenberg, A. S. C. On making statistical assumptions. *British Journal of Statistical Psychology*, 1953, **6**, 41–3.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Gustafsson, J. E. *Testing and obtaining fit of data to the Rasch Model.* (Report No. 83). Goteborg: Institute of Education, 1979.

Halford, G. S. and Keats, J. A. An integration. In J. A. Keats, K. F. Collis, and G. S. Halford (Eds), *Cognitive development: Research based on a neo-Piagetian approach.* Chichester: Wiley, 1978.

Hindley, C. B. and Owen, C. F. An analysis of individual patterns of DQ and IQ curves from 6 months to 17 years. *British Journal of Psychology*, 1979, **70** (2), 273–94.

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. *Psychometrika*, 1977, **4**, 601–8.

Jensen, A. R. *Educability and group differences.* London: Methuen, 1973.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. *Psychometrika*, 1964, **29**, 215-31.

Keats, J. A. Survey of test score data with respect to curvilinear relationships. *Psychological Reports*, 1964, **15**, 871-4.

Keats, J. A. An experimental study of cognitive factors. *Australian Journal of Psychology*, 1965, **17**, 52-7.

Keats, J. A. Test theory. *Annual Review of Psychology*, 1967, **18**, 217-38.

Keats, J. A. A proposed form for a developmental function. In J. A. Keats, K. F. Collis, and G. S. Halford (Eds), *Cognitive development: Research based on a neo-Piagetian approach*. Chichester: Wiley, 1978.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika*, 1962, **27**, 59-72.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis: The interpretation and computation of some latent structures. In S. A. Stouffer et al., *Measurement and prediction*. Princeton: Princeton University Press, 1950.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, 1952, **7**.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 1978, **31**, 19-26.

McCall, R. B., Applebaum, M. I. and Hogarty, P. S. Developmental changes in mental performance. *Monographs of the Society for Research in Child Development*, 1973, **38** (3, Serial No. 150).

Marx, M. H. Formal theory. In M. H. Marx and F. E. Goodson (Eds), *Theories in contemporary psychology*. New York: Macmillan, 1976.

Munsinger, H. The adopted child's IQ: A critical review. *Psychological Bulletin*, 1975, **82**, 623-59.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Skodak, M. and Skeels, H. M. A final follow-up study of one hundred adopted children. *The Journal of Genetic Psychology*, 1949, **75**, 85-125.

Terman, L. M. and Merrill, M. A. *Measuring intelligence*. Boston: Houghton-Mifflin, 1937.

Thurstone, L. L. The mental age concept. *Psychological Review*, 1926, **33**, 268-78.

REACTANT STATEMENT

Kevin F. Collis

In speaking to his paper, Professor Keats has really removed any necessity for me to react to the specifics in his paper. He has freed me to speak about the general notions underlying the differences between the two models, true score and latent trait, especially in relation to my own major research interest, cognitive development. It should be noted at the outset that my interest in the models arose from practical problems, initially in the classroom situation, and has never been directed towards the niceties of the mathematics involved.

Like most of us here, I was brought up on the classical true score model, $X_i = T_i + E_i$. However I very quickly began to find it not totally satisfactory for practice in the classroom. This was probably because, as is pointed out in Keats's paper, one was never quite sure of what was being estimated nor of what the basis of estimation was. Quite apart from this, the use of the model led to several undesirable practices, two of which I shall mention briefly.

First, teachers tended to use the various tests based on this model, for example intelligence tests, to 'label' and categorize *individual* children. Once the label had been attached, it was almost impossible for the child to remove it. Most often, and even more sadly, the child and its parents accepted the decision. In classroom practice, once the child had been categorized as of average IQ (say) then many things followed. The individual could be ignored—expectations were determined and a ready excuse was available for any one of a number of quite separate school or home-based problems.

The second undesirable practice which had arisen, partly because of the classical model. was what I like to call the 'carelessness' syndrome. Teachers' classroom tests were, often unconsciously, based on the true score model. They would set a series of items to which the child was to respond, the various correct scores were added, and the total score represented the child's achievement level in the content area concerned. Apart from incorporating many dubious assumptions, this form of assessment focused the teacher's attention on the number correct and on ranking the children in order rather than on why an individual child did not succeed on a particular item. In mathematics teaching, failure to succeed on an item was often put down to carelessness especially if the overall total score was deemed satisfactory. However, in my experience, children were very rarely careless—they seemed, in the main, to take an inordinate amount of trouble to try to follow the model solutions provided by the teacher.

It was largely these undesirable and clearly unfair practices which led me to an interest in what lay beneath the surface. Unwittingly, at first, and then more and more consciously I became involved in what is known in the jargon as latent trait analysis. Let us look at a couple of particular examples from elementary mathematics.

Children can be asked to find 'Δ' in each of the following statements:

(i) $3+4=\Delta+3$
(ii) $7-4=\Delta-7$

Each contains the same number of elements and operations; each uses small numbers. Why then should the first be easily at'ainable by early primary school children and the second not readily achieved until late primary/early secondary school?

The most productive method for finding out seemed to be to talk with the children at various age levels as they attempted to solve the problems. Two strategies for solving the first were in evidence with the younger children—neither of which was a satisfactory strategy for the second. The strategies used were either a low-level pattern seeking—'There's no "4" on that side', or some form of elementary 'counting on', that is, '3 and 4 are 7, so, we need "4" so that 4 and 3 makes 7'. Questioning revealed that the obvious solution of $3+4=7$ followed by $7-3=4$ does not only not occur to these children but will be denied as an appropriate method for solving the problem. Clearly the children needed to be able at least to admit the usefulness of this last strategy if they were to succeed on the second problem. In reality they needed to do more than that. Having obtained '3' from '7 – 4' they needed to have a sufficient overall view of the problem to *add* when subtraction was so fresh in their minds and so strongly suggested in the question.

This increasing ability to solve problems involving more and more complex manipulations of the data could be linked, intuitively at first, to the cognitive growth phenomena which the Piagetians and neo-Piagetians were describing in the literature. Further investigations (Collis, 1975) enabled *logical* links to be made between the data and cognitive development models.

As it turned out when the items which had been devised were analysed using the latent trait model (ACER, 1977) the earlier intuitions were confirmed. Perhaps more significant, in the context of the present paper, psychometricians such as Keats became intrigued by the results coming out of a number of cognitive development studies (Keats, Collis, Halford, 1978) and began to seek a suitable mathematical model to analyse the data in a more objective manner and to reconcile any new model thus devised with the classical model. As Keats's paper shows, a good start has been made in both these areas.

In conclusion, then, my experience suggests that both models have their uses in practice—the latent-trait model set out in Keats's paper is particularly valuable for investigating the developmental concerns currently surfacing in educational and psychological practice. Both models can be abused by their users—especially those who do not fully understand the assumptions underlying the particular model. The classical model has been around a long time and so there is much more evidence available of its abuse. I see this seminar as serving two important purposes; one, developing a basic understanding of the underlying assumptions of the latent-trait model and two, beginning to reconcile two models, one with the other. I believe that Keats's paper has contributed significantly to these purposes and to the theme of this seminar.

REFERENCES

Australian Council for Educational Research. *Mathematics Profile Series: Operations Test.* Hawthorn, Vic.: ACER, 1977.

Collis, K. F. *A study of concrete and formal operations in school mathematics: A Piagetian viewpoint.* Hawthorn, Vic.: ACER, 1975.

Keats, J. A., Collis, K. F. and Halford, G. S. (Eds). *Cognitive development: Research based on a neo-Piagetian approach.* Chichester: Wiley, 1978.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

3

The Use of Latent Trait Models in the Measurement of Cognitive Abilities and Skills

Bruce Choppin

MEASUREMENT SYSTEMS

What motivates most of the work to be described in this paper is a wish to develop a sounder basis for the measurement of educational achievement. I will not dwell much on the short-comings of traditional approaches based on the true score concept except where it is necessary to point out that it does not lead to a system of measurement with the sort of properties that we want.

In the argument, I plan to use the measurement of temperature as an analogy. Temperature is a familiar concept and ideas about some objects being hotter or colder than others must reach back very far in human history. But temperature is an invisible commodity and measurements may be made only indirectly. Turning 'hot' and 'cold' into number values on a scale did not come easily, and even when two temperatures (say the freezing and boiling points of water) are given arbitrary numerical values there is no very obvious procedure for locating intermediate temperatures on the scale (Middleton, 1966). However, the problems of measuring temperature have been largely solved in the last 150 years and the way in which the measurement system developed contains some useful lessons.

It could be argued that human achievement is a very different type of concept. The outcomes of it may be extremely visible, and there would seem to be no need to turn to indirect methods of measurement. For some areas of achievement this is clearly true. If we want to know how fast someone can run, we can time them over a fixed distance with a stopwatch. If we want to know how high they can jump, we set up hurdles and measure them in centimetres or inches. Mental abilities in general, and academic achievement in particular, do not lend themselves to a direct approach. If we want to know how good somebody is at mathematics, we cannot expect a tape measure or stop-watch to give us much help. The best we can do is to take a sample of tasks from the realm of mathematics and, by observing performance on those tasks, infer something about a hypothetical level of performance on a more general ability. It is in this sense that the measurement of academic achievement *has* to be indirect, and the trait itself treated as latent.

What then are the properties we seek in a system of measurement? First, as a matter of convenience, we require that the instruments with which the measurements are to be made shall be usable over a range of values of the variable being measured. A thermometer that works only at one particular temperature has very limited value. An unmarked stick exactly six feet long may be quite useful for dividing people into two groups one of whom all have heights less than six feet and the other all greater, but its value as a measuring instrument will be extremely limited in comparison with a properly calibrated ruler.

Further, one would require that the instrument is not unduly sensitive to factors irrelevant to what is being measured. Neither thermometers nor rulers should react noticeably to changes in humidity or barometric pressure.

More fundamental perhaps is the requirement that instruments should be to some extent interchangeable. It should not matter which of several available thermometers is used to measure the temperature of a room or the temperature of a cup of coffee. The results obtained should not depend upon which thermometer is chosen, and this has implications for the calibrations employed. In itself calibration is not a difficult task. Any set of marks on a ruler can be treated as calibrations of length and any set of marks on a thermometer as calibrations of temperature. The raw score achieved on a test can reasonably be regarded as a calibration of performance. The problem arises when consistency among the calibrations is required so that instruments themselves may be used interchangeably.

Cross-calibration procedures have something in common. To calibrate two thermometers one against the other, you might use both to measure the temperature in several situations (say freezing and boiling water and a number of points in between) and observe carefully the readings on each thermometer. To cross-calibrate two tests, a straightforward procedure would be to give two tests to a number of people and to observe the raw score of each person on each test. In this way a table could be developed to show how the raw score on one test was related to the raw score on the other. This, in a limited sense, is a basis for interchangeability since if a person's score on one of the tests is known it would always be possible to predict his score on the other.

Unfortunately this does not work too well in practice. Firstly, the errors of measurement usually present with the test scores are of such a magnitude that in a real-life situation a single raw score on test A is likely to correspond to a whole range of raw scores on test B. (The same sort of thing happens with thermometers but the measurement error there is so small that it is usually ignored.) Secondly, all pairs of instruments have to be brought together for the cross-calibration, and this rapidly becomes impracticable as the number of instruments for measuring a particular variable is increased. If I develop a new test of arithmetic in a world where 100 other tests of this topic already exist, then in theory I need to carry out 100 cross-calibration experiments in order to make my new test fully a part of the measurement system. To solve this problem for temperature, constructors of thermometers make use of the apparently regular, though different, expansion properties of solids, liquids, and gases, as temperature is increased. Most thermometers make use of these expansions so that an indirect measure of temperature is obtained by making a direct measurement of length. Equal changes in length are said to correspond to equal changes in temperature, and this makes the construction of a variety of types of thermometer relatively straightforward. Calibration is carried out against a standard thermometer at only two points on the scale, the rest of which is marked off in equal intervals of length. With this system, one can use a number of different thermometers with confidence that a reading of 47 degrees on one of them means more or less the same thing as a reading of 47 degrees on any of the others. The consistency achieved by real-life thermometers is frequently exaggerated. Mercury-in-glass and platinum-resistance thermometers which agree at 100°C will differ by about 9°C at 300°C (Nelkon and Parker, 1968). Neither is necessarily true or false; they represent two facets of an inherently inconsistent system. Though less dramatic, similar inconsistencies occur among liquid-in-glass thermometers. Many different liquids all with different properties were tried during the first hundred years of thermometry and the general agreement in the 18th century to standardize on mercury as the liquid seems to have been arbitrary (Eysenck, 1980).

With mental tests, a major effort to get around the calibration problem has come to be known as norm referencing. Here a hypothetical scale of performance is defined by the distribution of ability within a particular population. If a particular test is administered to the whole population (or a representative sub-sample of it), then it is possible to define a transformation of raw test scores into, for example, percentiles. Thus a score of 30 on test A may be held to be equivalent to a score of 36 on test B, if both translate to the same percentile value for the same population. As past experience has shown that a normal distribution of ability is a reasonable hypothesis to hold for most populations, the procedure to establish percentile norms for a new test need not be arduous. (Age norms take these procedures one step further by using the average performance of each of a whole set of different sub-populations).

It is instructive to consider norm referencing for temperature. Suppose thermometers were all calibrated in terms of the percentage of days that were cooler than a particular temperature. They would have their uses, but clearly also their limitations. They would not be much use for measuring the temperature of cups of coffee or human body temperatures. In the context of weather, the calibrations would be meaningful only in the restricted context of the climate where the calibration was carried out, and even then they would not be much use for night-time temperatures. Further one might well say, 'Today is a cool day considering the whole year, but is it cool for the end of May?'. Normed calibrations would not contain much information about that.

These limitations are just those that restrict the usefulness of the norm-referenced standardized test. The calibrations are only strictly relevant in the context of the reference population and, in real-life situations, it is almost always true that the population of interest will *not* be the one on which the calibration was carried out. Human populations are not particularly easy to define and, in particular, the characteristics of student populations usually change quite rapidly, so that a standardization carried out one year may already be noticeably inaccurate twelve months later. Further the idea of a single population is often unhelpful. Individual children need to be considered in terms of their own characteristics, and not merely as representatives of a national population. Sex, ethnic origin, and educational background may all be crucial to the interpretation of a particular test performance.

If norm referencing is not the answer, then perhaps we should seek some theoretical basis for test interpretation analogous to that which turned the problem of the measurement of temperature into essentially a problem of the measurement of length. Just as the relationship of the expansion of materials with temperature can be expressed (approximately) in mathematical form as an equation so we seek a formal mathematical representation of the way that performance level on an achievement trait translates itself into observable performance on a mental test. It should be clear by now that the model implied by adding up the number of correct responses and using the resulting score as a measure of achievement will not do. In general the same person would get different scores on different tests even though his achievement level remains constant, and so we must somehow build information about the content of a particular test into our model.

A single test item is analogous to the six-foot long uncalibrated measuring rod we mentioned earlier. It can divide people into two groups: those who can answer it correctly and those who cannot. A test is like a bundle of rods of different lengths. A height-measuring system based on them must take account of which rods are in the bundle.

In the last 20 years or so, a number of alternative models of test behaviour have been advanced. In this paper I shall concentrate on one of them which seems to have by far the widest range of application. This is the Rasch model which relates the probability of a person responding correctly to a particular item to a function of just two parameters: the ability of the individual and the difficulty of the item (Rasch, 1960).

It is usually written:

$$\text{Probability}\left\{a_{\nu i} = 1\right\} = \frac{W^{(\alpha_\nu - \delta_i)}}{1 + W^{(\alpha_\nu - \delta_i)}} \qquad (1)$$

where $a_{\nu i} = 1$ represents the event that person ν responds correctly to item i; α_ν is a parameter measuring the ability of person ν and δ_i is a parameter measuring the difficulty of item i. W is a constant which determines the size of the units of measurement. For measures in logits, W is set at e. For measures in brytes or wits, W is set at $3^{0.2}$ which is about 1.2457.

Certain assumptions are built into this model. The first is that probability of a particular response being correct does not depend on which other individuals are attempting the same item, or on the pattern of responses these individuals might give. More importantly perhaps, it assumes that the probability of a correct response to a particular item does not depend on which other items make up the test, in which order they appear, or what responses were given to the items that preceded the one under examination. This assumption is known as 'local independence'. Secondly the model assumes that the individual's response is conditioned by his ability to answer questions in this area, but not by his motivation, his tendency to guess, his degree of hunger, or indeed *any* other personal attribute. Thirdly, the model assumes that one and only one item parameter (difficulty) affects the outcome and that other item characteristics (such as reliability or discrimination) are not relevant.

In both Britain and the United States, the educational literature still occasionally produces an outraged statement by a respected traditionalist that he has looked at what underlies all the fuss over the Rasch model, and he has found that the model simply is not true. If this can be taken to mean that the Rasch model does not exactly represent the behaviour of real people in actual testing situations, then let me hasten to agree. The Rasch model is a gross simplification, deliberately designed to provide an approximate representation of reality, not an exact one. That indeed is the virtue of scientific models. Is Charles' law about the expansion of gases true? No, of course not. Neither is Van de Waal's Equation of State a true account of the behaviour of gases under changes in temperature. It is more accurate than Charles' law, but still an approximation. Newton's laws of motion are themselves no more than an approximation. They work very well most of the time, but are woefully inadequate in some circumstances.

More telling is the criticism raised repeatedly by Goldstein and others in the United Kingdom (e.g. Goldstein 1979) that the Rasch model is unsound because its basic assumptions are untenable. That they are untenable is not really open to dispute, but I would argue that this again is not a sufficient reason for dismissing the model. After all, the use of flat maps to represent portions of the earth's surface involves assumptions about the preservation of relative areas and distances that we know to be untrue. Exactly what does a scale of 1:1 000 000 or 10 miles to the inch imply? Yet two-dimensional maps are almost everywhere regarded as being useful aids to personal navigation. Our experience to date with the Rasch model suggests that it is quite robust with regard to violation of its assumptions. Even when items have a built-in dependence upon another and when parameters such as discrimination vary widely, the results obtained when the Rasch model is used to measure people show at most only very minor inconsistencies.

There are, I would submit, three separate reasons for adopting the Rasch model as the basic scaling technique for measures of achievement. They are:

1 It is mathematically simple and convenient to use. Methods of estimating the parameters are relatively straightforward and they do not require vast amounts of data.

2 The model is in fact a direct extension of current testing practice which adds up the number of correct responses and uses this as a measure. In fact a number of authors have shown that under normal circumstances the raw score on a test is a sufficient statistic for the ability of the person achieving it. That is, all the information about the person's ability contained in the set of responses he gives is concentrated in the raw score. Similarly all the information about the relative difficulty of items is contained in the set of facility indices (i.e. the proportion of correct responses item by item). Thus, if we have a complete data matrix resulting from each of a particular group of people attempting all the items in a particular test, where one is reported in the matrix for a correct response and zero for an incorrect response, then the marginal sums of this matrix contain all the information necessary for calibrating these items and measuring the people. The Rasch model provides support for the use of raw scores for a variety of measurement purposes such as the ranking of students. There is a one-to-one monotonic relationship between the raw score and the underlying latent trait scale.

3 The Rasch model does appear to predict the behaviour of real test

items and real people with considerable accuracy (given the enormity of the simplifying assumptions).

A straightforward illustration of this is given below. When one item is more difficult than another, this manifests itself by a tendency for people who succeed on it also to succeed on the other item. Information about the relative difficulty of the two items is contained in the respective success rates, for a given group of people.

The Rasch model leads to the somewhat surprising (but easy to remember) result that, for any two items (*i*, *j*) measuring the same ability, the ratio of the number of people who respond correctly to *i* and incorrectly to *j* (say b_{ij}), to the number who do the opposite (say b_{ji}) should be constant – a measure of the relative difficulty of items *i* and *j* – no matter what the ability level or distribution of the people (Choppin, 1978).

In the notation of equation (1)

$$W^{(\delta_j - \delta_i)} = \frac{\text{Probability}\,\{a_{\nu i} = 1,\ a_{\nu j} = 0\}}{\text{Probability}\,\{a_{\nu i} = 0,\ a_{\nu j} = 1\}} \qquad (2)$$

$$\text{and } (\delta_j - \delta_i) \text{ is estimated by } \frac{\log\ b_{ij} - \log\ b_{ji}}{\log W}$$

To illustrate this, I have looked at four items used in the 1971 IEA Science survey (Comber and Keeves, 1973). The data I have are from six separate samples of about 1000 pupils ranging from one of eighth grade pupils in non-academic streams to one of twelfth grade academic pupils. Traditional facility values for a single item vary widely from one sample to another reflecting the varying abilities of the pupils. Two of the four items are in chemistry and two in biology. For each pair, the values of b_{ij} and b_{ji} were counted in each sample. The results are plotted in Figure 1. The relative difficulty of the two items, according to the Rasch model, is given by the slope of the line joining the origin to the sample point. You will note the consistency from one sample to another despite the extreme variation in ability.

These data are not 'cooked'. The items were drawn at random from the set that were administered across the wide age group. Discrimination indices range from 0.14 to 0.36. Both biology items were said to be measuring 'understanding' but, of the chemistry pair, one was classified as 'knowledge of facts' and the other as 'higher mental processes'.

So, as well as illustrating the sample-free aspect of Rasch relative difficulty, these data give some insight into the robustness of the model. The structure holds up well even when the departures from the underlying assumption of homogeneity and uniform discrimination are quite substantial.

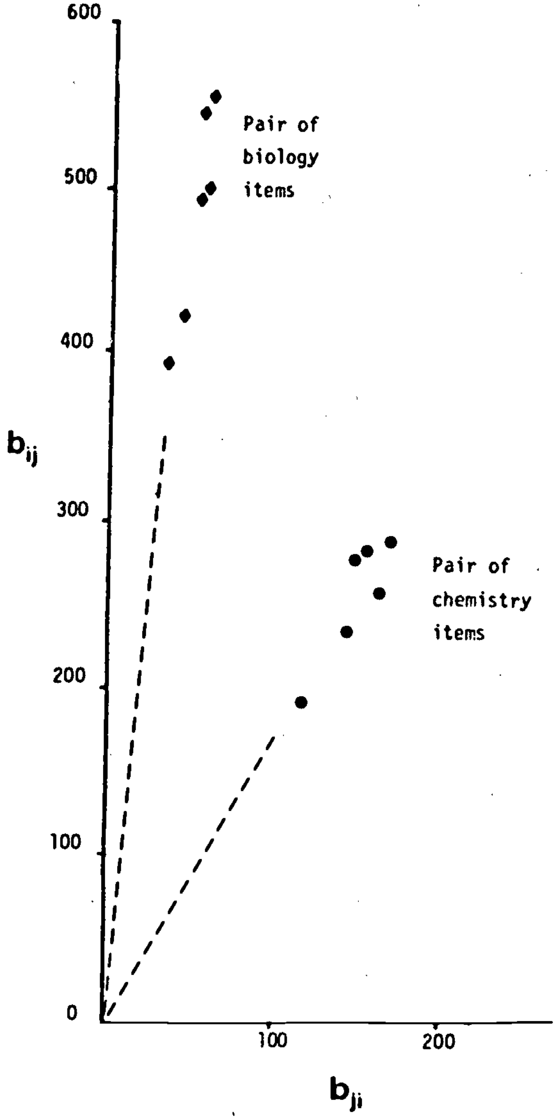


Figure 1 Results for Four Science Items on Six Widely Differing Samples Each of about 1000 Students (For a pair of items within one sample, b_{ij} is defined as the number of individuals who responded correctly to item *i* and incorrectly to item *j*.)

MORE COMPLEX MODELS?

In the United States the Rasch model has often been referred to as the 'one-parameter' model because it uses only one parameter to describe a test item (the difficulty parameter). A 'two-parameter' model has been proposed and investigated from a mainly theoretical point of view. Like the Rasch model it relies on a single parameter to describe the person taking the test (ability) but, in addition to difficulty, it introduces a parameter to represent the item's power of discrimination. Largely in response to theoretical objections resulting from the use of the multiple-choice format, a 'three-parameter' model has been suggested. This still retains a single parameter for the person involved, but adds another item parameter representing the 'guessability' of the item (i.e. the limiting probability that a person with absolutely no relevant ability would still respond correctly to the item). Though there is a substantial amount of published discussion of this model, it has been little used in practice.

In passing it should be noted that there has been discussion about the advantages of using a probability function based on the normal curve (proposed by Lord, 1952) rather than the logistic (exponential) functions adopted by Rasch and Birnbaum. In quantitative terms it appears that this would make very little difference to the results obtained and the logistic form of model is now generally preferred because of its relative mathematical simplicity.

The reason for interest in these more complex models is clear. The simplifications made by the Rasch model are rather extreme, and a more complex model could be expected to provide a better fit to real data.

Against this one must set the disadvantages of losing the simple one-to-one relationship of latent trait measure with raw score. The more complex models require very lengthy computation in order to score the test. Even to rank candidates on a very short test, these models will almost inevitably require the use of a computer.

Secondly, while quite usable estimates of the Rasch model parameters can be obtained from as few as 30 candidates attempting a test, the more complex models seem to require samples running into the thousands in order to obtain a similar degree of reliability in parameter estimation. Observations come as single bits of evidence, and it seems to be difficult to squeeze more than one item parameter out of a single bit. It is for this reason, I suspect, that Wright found that, in practice, the more complex models fitted real data less well than did the Rasch model.

The orthodox view among Rasch scalers is that it is better to avoid the problems which prompt the introduction of extra item parameters. A good test from the Rasch point of view (and hence also from the point of view of those who use traditional test statistics) is one which avoids substantial amounts of guessing and items whose discrimination parameters vary widely. Where the Rasch model is used as an aid to test construction, it is usually employed in this mode.

PARAMETER ESTIMATION

It is now nearly 20 years since Wright and I began to play around with different computational algorithms for solving the Rasch model. A good many other people have come in on the act since then, and it would be fair to say that by now this particular topic is fairly thoroughly explored and documented.

That there is a problem requiring a solution at all results from the stochastic or probabilistic form of the Rasch model itself. It gives only the probability of a correct response to an item whereas the actual observation is of success or failure. There is unfortunately no way to observe probabilities directly.†

In general, however, if we have *N* people responding to the *k* items in a test we have a total of *Nk* bits of information to estimate the *N*+*k* parameters (*N* abilities and *k* difficulties). The problem is to find values of the parameters that best explain the set of observations, and then to check that this explanation is good enough to justify the use of the Rasch model in these particular circumstances. Many ways of fixing the values of the parameters have been suggested; some precise but computationally rather long-winded (Andersen, 1973); some rough and ready but easy to calculate (Wright and Stone, 1979). In general they can be grouped into two separate categories: 'least squares' and 'maximum likelihood'.

The least squares approach is based on the idea of minimizing the discrepancies and deviations from the model. It is the approach suggested by Rasch in his book, and was the first method to be investigated in any detail.

The maximum likelihood approach begins by specifying the probability of a particular set of observations, given particular values for the parameters. The procedure then calculates the values of the parameters that make this probability function a maximum. In practice the log likelihood function is maximized as this is computationally rather easier, and leads directly to estimates for the standard errors of the parameters.

Some but by no means all maximum likelihood methods produce a systematic bias in the parameter estimates obtained (an approximate correction factor has been proposed to take care of this). Some but by no means all maximum likelihood methods are computationally lengthy and hence rather expensive. My own experience suggests that, although the quickest maximum likelihood method (producing unbiased results) can-

† In this case (as in many others) it would be quite impracticable to face an individual with the same test item on a large number of occasions in order to estimate the probability by way of the relative frequency of success.

not compete for speed with the shortest of the least squares approaches, it is somewhat more stable with badly conditioned data. Where both methods are applicable to the same set of data, it is comforting that the results usually agree to within a tenth of a standard error, and I have never seen a case where real (as opposed to artificial) data gave rise to least squares and maximum likelihood estimates that differed by more than half a standard error.

EXTENSIONS TO THE BASIC RASCH MODEL

Most of the published accounts of the use of the Rasch model refer to the standard situation where each of a group of people attempts all the items in a test, and where each item response is scored either right or wrong. The last few years, however, have seen a great deal of work in developing extensions to the basic model to cope with more complex testing situations. I will consider three such extensions.

Incomplete Observation Matrices

If every person in a group attempts every item in a test, the data can be arranged as an N by k rectangular matrix of ones and zeros, where one represents a correct response and zero an incorrect response, as in Figure 2.

In this case the $(N + k)$ marginal values (i.e. the row and column sums)

		Items						
		1	2	3	4	5	6	
Persons	A	1	1	0	1	0	0	3
	B	1	0	1	1	1	0	4
	C	1	0	1	0	0	0	2
	D	1	1	0	1	0	0	3
	E	0	1	0	0	0	1	2
	F	1	1	1	0	0	0	3
	G	1	0	1	1	1	0	4
	H	1	1	0	1	1	0	4
	I	1	1	0	0	0	0	2
	J	1	1	1	1	0	0	4
	K	0	0	1	0	0	0	1
	L	1	1	0	0	0	0	2
		10	8	6	6	3	1	Margins

Figure 2 Hypothetical Data Matrix for Results of 12 Persons on a Six Item Test ($N = 12$, $k = 6$)

contain enough information to estimate all the parameters. However, many real-life situations occur in which this matrix is incomplete. Although it will then have less than Nk bits of information, it will still usually have more than enough to develop estimates for the $N+k$ parameters. With missing data, the parameters cannot be estimated from the margins.

There are several different situations in which omissions not resulting from a candidate's *inability* to answer a question can occur. One is through some irregularity in the test administration. Some candidates may be given test booklets containing printing errors; one may be taken ill half-way through an examination.

Another such situation is a test or examination in which the candidate is allowed a choice of questions. This procedure is fairly standard in British public examinations where a typical rubric might run 'Answer questions 1, 2, and any five from numbers 3–15'. It is widely believed that such a procedure is fair to candidates who may (through no fault of their own) have missed being taught some parts of the curriculum. Be that as it may, it can easily result in a situation where, although an examination consists of N items, no student vector contains more than n responses – and of course different students choose different questions.

A third situation in which the data matrix is not complete occurs when different tests are quite deliberately given to different students. Sometimes this method is adopted to prevent students from copying the answers of their neighbours, and hence to increase the overall security of the examination, but it is also a quite legitimate way of increasing the range of items on which achievement data are gathered without unduly lengthening the time devoted to testing. One example of the latter kind occurred in the IEA 1971 Survey of Science Achievement which I mentioned earlier. For students in the pre-university year six different forms of the test were prepared. Each consisted of a basic sub-test of 60 items which appeared in all forms, and six more advanced items drawn from an additional pool of 36. This ensured that, while no student was asked to respond to more than 66 items, achievement data on a total of 96 items were obtained. Another pertinent example occurs in the work the NFER is doing as part of a national monitoring of standards for the government's Assessment of Performance Unit (APU). For the assessment of primary school mathematics, 26 different forms of test were used, each one containing only one-thirteenth of the total pool of items. (The test was so arranged that each item appeared in two separate forms.) When observations from these types of testing are arranged in a *persons-by-items* matrix (see Figure 2), it is clear that large parts of the matrix will be empty. Yet it would still be desirable to be able to calibrate *all* the items one against the other, and to measure the achievement of *all* the people.

Another, and rather more novel, type of 'missing data' occurs as a result of editing an observation matrix, often as a result of a Rasch scaling analysis. For example, after estimating item difficulties and the abilities of candidates, it is possible to use the model to examine the probability or improbability of each separate item response. This method has been used to identify lucky guesses (a student correctly answers a question that appears from the rest of the data to be far too difficult for him). Further, one may discover that a group of testees respond in apparently eccentric fashion to perhaps a sub-group of items, suggesting that perhaps those particular items are not appropriate for them. It is possible, in this way, to identify instances where the item response appears not to be a good indication of the candidate's overall level of achievement and these instances can be edited out before a second analysis of the persons-by-items matrix. This matrix will now have some holes.

How is the incomplete matrix analysed? The answer comes from the result mentioned earlier for pairs of items. The ratio of the probability of getting item *i* correct and *j* incorrect to the probability of getting *i* incorrect and *j* correct is a simple function of the relative difficulties of items *i* and *j*. When faced with an incomplete observation matrix, we decompose the test into all possible sub-tests of length two. Any individual who responds to more than one of the original *k* items contributes to the estimation of the relative difficulties of at least some of the possible item pairs. Once all the items have been calibrated, it is relatively straightforward to look at the set of responses a particular person gave, and derive from them an estimate of that individual's ability (Choppin, 1978). Both maximum likelihood and least squares methods work in this analysis of incomplete observation matrices and, although over the years my preferred method has usually been that of maximum likelihood, I am now coming to the conclusion that in routine testing situations the non-iterative least squares algorithm is going to prove the more reliable.

Partial Credit

Suppose that instead of being scored zero/one, a set of item responses has each been scored on a scale from zero to five depending on the degree of 'correctness'. Tests of this sort are not unknown in Great Britain, though I do not know whether you have to deal with them in Australia.

Two different methods have been developed for analysing such data with a Rasch model. The first, stemming from the work of Wright and his colleagues in Chicago, is to treat each item in the example given above as replaceable by five dummy items each of identical difficulty and scored zero/one. The score actually achieved on one of the original items is thus taken to be the raw score obtained by summing the scores on the sub-set of dichotomous dummy items. Since the dummy items associated with one real item are assumed all to have the same difficulty value, the margins of the observation matrix may be collapsed somewhat before the conventional parameter estimation method is applied.

An alternative approach, and the one that I have been developing, is to convert the actual score awarded (on the 0–5 scale) into a fractional score on a continuum from zero to one. Thus in the example quoted above, a score of 4 would be converted to one of 0.8, indicating that the candidate has achieved four-fifths mastery of that particular question (and one-fifth lack of mastery). On a second item for which the score is three, I deduce a degree of mastery 0.6 and lack of mastery 0.4.

A simple extension of the Rasch model to accommodate partial scores between zero and one, in lieu of dichotomous scores, replaces the probability function by an expected value. From this we can use the scores achieved on two items by the same individual to give an estimate of the relative difficulty of the items.

$$\text{Expected value}\quad E\left[\frac{(a_{\nu i})}{5}\right] = \frac{W^{(\alpha_\nu - \delta_i)}}{1 + W^{(\alpha_\nu - \delta_i)}} \tag{3}$$

This equation is analogous to equation (1) above. If we call the two items in our example *i* and *j*, then the relative difficulty of the items is estimated by

$$\frac{1}{\log W}\;\frac{\sum_\nu a_{\nu j}(5 - a_{\nu i})}{\sum_\nu a_{\nu i}(5 - a_{\nu j})} \tag{4}$$

The items are then calibrated by looking at all possible item pairs as in the missing data method outlined above.

I have not much more to report about the use of non-dichotomous scoring systems at the moment. Both methods of analysis are in use. Sometimes they produce virtually identical results (e.g. where all the items in a test are scored on the same scale). On other occasions the results may be somewhat different, and it is up to the analyst to decide which of the approaches is the more sensible in that particular case. The first method I described weights questions according to the maximum number of marks awarded for them. The second method weights all the questions equally.

Early results that I have had suggest that non-dichotomous scoring can give much more information about a candidate from a limited number of item responses. It can thus substantially reduce the standard error of measurement. On the other hand it is in general rather harder to meet the Rasch model requirement of homogeneous discrimination levels when non-dichotomous scoring is used.

Markers or Judges

This extension to the basic model is intended to cope with the situation where several different judges provide ratings of the quality of some performance. I think that the best way to present this will be to describe in some detail an actual problem, and the progress made so far towards its solution. It arose in connection with background work carried out for the Assessment of Performance Unit that I mentioned earlier. In our national monitoring of writing skills it was deemed appropriate to have each writing sample graded by expert judges. Since it would clearly be impracticable to have any one judge consider all the writing samples obtained in a large national survey, we decided to explore the feasibility of recruiting a pool of expert graders on whose judgment we could rely, and whose variation on a severity-leniency scale could be minimized. To accomplish this the following experiment was conducted.

Seven hundred and fifty students provided writing samples for analysis. Eleven separate writing tasks had been defined and each student was asked to respond to two of them, one from the set (task 1–task 10) and also task 11. In the experiment four markers were used, although only two marked each candidate's papers. Marker one was never paired with marker three, and marker two never with marker four. Apart from this all combinations of markers appeared with approximately equal frequency.

Each marker was required to grade each task on four separate criteria. For the sake of brevity these criteria will be referred to as content, grammar, style, and orthography. All grades were made on a 1 – 5 scale with 5 being the best work.

Thus for each of the 750 candidates, we had 16 scores (two tasks × two markers × four criteria). A Rasch scaling was carried out in order to estimate the marker parameters and the task/criteria parameters. These parameters were estimated to provide the best possible fit of the data to the model. First,

$$E\left\{\frac{(X-1)}{4}\right\}=\frac{e^{(\alpha-\delta-m)}}{1+e^{(\alpha-\delta-m)}} \tag{5}$$

where α is the writing ability parameter for a student, δ is the difficulty level parameter for a criterion on a particular task, m is an adjustment for marker severity and X is the grade awarded. The expression $(X-1)/4$ converts the grade to a fractional score on the zero/one interval. This is analogous to the procedure I described in the preceding section on partial credit. For the present experiment it should be noted that the estimation has no solution unless at least some interactions of task/criterion are graded by more than one marker, and some individual pupils are graded by the same marker on more than one task.

If these conditions are met

$$W^{(\delta_j-\delta_i)} \text{ is estimated by } \frac{\Sigma(X_i-1)(5-X_j)}{\Sigma(5-X_i)(X_j-1)} \tag{6}$$

where *i*, *j* are indices indicat ng particular task/criterion combinations (henceforth called items) and the summations are taken over all pairs of pupil and marker for which grades for both items *i* and *j* exist.

Similarly,

$$W^{(m_g-m_f)} \text{ is estimated by } \frac{\Sigma(X_f-1)(5-X_g)}{\Sigma(5-X_f)(X_g-1)} \tag{7}$$

where X_f and X_g are the grades awarded by marker *f* and marker *g*, and the summation is taken over all pupil responses to individual items for which both marker *f* and marker *g* provided grades. These two sets of equations were analysed using both least squares and maximum likelihood methods. In no cases were the resulting calibrations different from each other by more than 0.1 wits. All the results reported below are drawn from the maximum likelihood analysis.

Averaging the results over criteria and tasks gave the overall difficulty levels shown below:

(i) *Tasks*

Tasks	Mean difficulty
task 1	50.8
task 2	49.6
task 3	50.5
task 4	49.4
task 5	49.7
task 6	50.8
task 7	49.7
task 8	49.4
task 9	49.5
task 10	50.3

(ii) *Criteria*

Tasks 1 10	criterion (a)	content	– adjustment	0.7 wits
	criterion (b)	grammar	– adjustment	zero
	criterion (c)	style	– adjustment	+ 0.3 wits
	criterion (d)	orthography	– adjustment	+ 0.4 wits
Task 11	criterion (a)	content	– 51.5 wits	
	criterion (b)	grammar	– 51.4 wits	
	criterion (c)	style	– 48.7 wits	
	criterion (d)	orthography	– 49.5 wits	

The interpretation of these results was that, for example, task 3 was on average one wit harder to score well on than task 9 (50.5–49.5) and that, for tasks 1–10, the *content* grading was about one wit more lenient than the *style* grading (−0.7)−(+0.3).

Task 11 was reported separately because it was common to all students, and the pattern of marking was substantially different. Whereas 'content' was the easiest criterion on which to score for tasks 1–10, for task 11 it was the most difficult. Both content and grammar were marked more severely for task 11 than for other tasks or other criteria. Was this perhaps the result of overexposure of markers to responses on this task? In any case, since everyone took task 11, variations on it from the other tasks did not introduce any bias into the estimation of student attainment.

The variations present in tasks 1–10 *did* have implications for measures of student attainment. If the outcome was taken to be a score derived from adding the eight separate grades provided by each of the two markers, then we can see that half of these grades depend to some extent on the choice of task. For a student of approximately average ability, one point on each grading scale is equal to about 4 wits. Hence the discrepancy of 1 wit between tasks 3 and 9 suggests a difference of about a quarter of a score point for each grade awarded. The sum of this would produce a difference of about 2 points in the total score; e.g. a student who took tasks 3 and 11 and got a total of 47 would be expected to score 49 on tasks 9 and 11. Although most discrepancies are smaller than this, some are larger, suggesting that (Rasch) scaling of the raw results was highly desirable.

The Results – Markers

Adjustments for variations of severity of markers are shown below:

	Criterion				
Marker	Content	Grammar	Style	Orthography	All criteria
1	0.4	0.7	0.9	0.2	0.2
2	0	0.6	1.8	0.3	0.5
3	0	0.3	0.4	0.2	0.1
4	0.4	0.2	2.3	0.3	0.6
			(SE = 0.5)		(SE = 0.3)

These results showed that there were no significant marker effects except on the criterion, style.

On style the discrepancies were substantial:

On average, marker 1 was 1 wit too severe,
marker 2 was 2 wits too severe,

marker 3 was about right,
marker 4 was 2 wits too lenient.

The structure of the data meant that it was not possible to test the consistency of these results over each task separately, but it was possible to compare grading standards on task 11 with the remainder.

	All criteria		Style	
Marker	Tasks 1-10	Task 11	Tasks 1-10	Task 11
1	−0.3	0.6	−0.3	2.2
2	0.7	0.4	1.6	2.1
3	−0.2	0	0.2	−1.0
4	−0.1	−1.0	−1.4	−3.3
	(SE = 0.4)		(SE = 0.7)	

From this it appeared that the main discrepancies between markers occurred on task 11. The difference between markers 1 and 4 on task 11 'style' was 5.5 wits whereas on the other tasks it averaged only 1.1 wits.

Since all students responded to task 11 the effect of differential marker severity on total score was considerable. The effect of having markers 1 and 2 rather than 3 and 4 was about 2 score points on task 11 and 1 score point on the other task, or about 3 points in total. Overall the results appear to confirm that the criterion 'style' was not being used in an acceptable fashion by the markers. The rest of the calibrations appear satisfactory. Marker calibration has been achieved, and on the basis of this set of data it seemed reasonable to conclude that there were no systematic differences of severity between the standards adopted by the four markers in this experiment.

MONITORING OVER TIME

My colleagues at the NFER are now heavily engaged in various aspects of Rasch scaling and the chief reason is the British version of national assessment that I mentioned earlier, the APU. It is appropriate therefore to spell out the precise nature of the problems raised by our work for the APU, how we are proposing to solve them, and the sorts of results we hope to produce.

The APU program includes the monitoring of achievement standards within our school system through the administration of tests to random samples of children at two or three different age levels. At the moment the cycle in each subject is an annual one (e.g. in mathematics, we administer tests to a sample of 10-year-olds each May, and to a sample of 15-year-olds each October) but it is possible that the testing frequency will be reduced in the future. The aim of the program is to provide a detailed description of the attainment of children in our schools on a wide variety of test material, to try to identify external factors associated with low patterns of performance over time. The aims are very similar to those which motivated the National Assessment for Educational Progress (NAEP) in the United States, and in Britain too we have had a debate over the merits of a purely descriptive approach in contrast with attempts to identify causal links between academic performance and background variables. However, for the moment, I shall concentrate on the issue of monitoring standards over a period of years, an issue that requires solutions to some interesting technical problems.

The first problem appears on the surface to be straightforward; we cannot use the same items in the tests year after year. One reason for this is that we are required to report results on an annual basis, and it is regarded as essential that we publish at least some of the test items in each of these reports so that the general public will understand and be able to interpret the performance statistics that we give. But once an item has been published in a report it becomes accessible to teachers who wish their pupils to perform well. As a result it may be expected that it will receive special treatment in a substantial proportion of classrooms and thus will no longer function as a good indicator of achievement across the curriculum. A more subtle difficulty inherent in the continued use of the same test items is that changes in curriculum do occur, albeit fairly slowly. Test items, that seemed entirely appropriate when the first mathematics tests were put together in 1978, may well seem much less appropriate for use in 1988. Our commitment is to test each year what is currently being taught in the schools. Our tests are supposed to remain up to date and the deliberate and repetitive re-use of the same items, although it would facilitate the comparison of test scores between years, would also guarantee that the tests steadily lost validity.

At one stage our then Secretary of State for Education was reported to be extremely sceptical about the existence of any satisfactory alternative. She felt that comparisons of performance from year to year would lack credibility with the general public unless they were based on the same test items. But, of course, there are ways to handle this problem. One fairly neat procedure is that adopted for the Scholastic Aptitute Test in the United States, wherein each test carries a small section which does not contribute to the reported score, but which contains items that appeared in the preceding year's tests. This provides a basis for equating standards from one year to the next using regression techniques. Thus a reported verbal aptitude score of 500 in 1980 should represent the same standard of performance as a score of 500 on the 1979 test, and this score of 500 was itself equated to a score of 500 in 1978. The year-to-year linking, always on the basis of current test material, is probably sound over a small number of years during which the population of test candidates does not change too dramatically. One suspects though that quite substantial errors may well accumulate over a period of 10 to 20 years, which was the sort of time scale at which the APU was aiming.

We preferred to approach the problem by using a bank of items scaled along the appropriate latent trait. To achieve a wide coverage of the curriculum we used some 600 items in the first survey of mathematics at age ten although no individual child was asked to attempt as many as 50. From the 600, we were able to discard a proportion the next year because they had been used as examples in our reporting, because they had proved not to have very good measurement characteristics or because they were judged no longer very relevant to what was being taught. (On the second cycle, this last category was largely non-existent, but we expect the number of items involved to grow as the years go by.) The discarded items are replaced with new items developed by a panel of practising teachers and curriculum experts, so that the next year's testing has a substantial overlap with the preceding year, but it has been changed sufficiently to bring it up to date. The new items, together with the survivors from the previous year, are scaled back on to the original latent trait scale, so that we feel we can make valid comparisons of standards from one year to the next. Further, since we can keep a careful watch on the performance of individual items over a number of years in order to ensure that they are performing in a consistent way, we feel more confident of our ability to make comparisons over a ten-year-period, or even longer if necessary.

This brings us to the second problem I would like to consider, the question of what to report. A simple statement to the effect that standards have gone up or down by so many points since the year before is unlikely to be of any help to anybody (except perhaps certain politicians). The important findings are tied up in the ways teachers and children are reacting to a changing curriculum, to the changing emphases being placed on topics within that curriculum, and in the resulting changed pattern of performance. Our aim in future APU work must be to try to quantify the changes in the pattern of performance so that they can be related to changes in curricular emphasis and, hopefully, to changes in the country's perceived educational needs.

If we are to attempt this, then we are forced to confront the reality that attainment in what we think of as a single school subject (be it mathematics, science, or English language) is in reality a multi-dimensional group of separate attainments. By this I do not mean to imply that, for example, attainment in geometry is uncorrelated with computational accuracy, understanding of algebra, or skill at using a slide rule. There is ample evidence from past research that these traits are quite highly correlated.

In an *n*-dimensional space there are two ways of looking at this situation. The first is the bundle of traits all pointing in roughly the same direction but each being assessed separately by its own set of items. The second is to take this *n*-dimensional cigar and decompose it into orthogonal axes. Here the major axis is a sort of conglomerate 'general mathematics performance' and the minor axes represent not 'geometry' but rather 'the way in which geometry is different from general mathematics'. Of the two approaches to analysis and reporting, I prefer the second for the following reasons.

If we choose to work with discrete sub-tests in geometry, computational skills, etc. each with its own latent trait scaling, we can link performance from year to year within a sub-trait without problem. We know however that, in time, particular traits will be emphasized more in the way in which school mathematics is taught, and others will be emphasized less. This may be reflected in increases in performance on certain traits and reductions on the others, but the link between the two will be hard to establish. Further it would be difficult to incorporate changes in the assumed structure of school mathematics; for instance, we may be required to combine two sub-traits into one, or break one up to make two new ones. With the second approach to analysis, the unit is the individual item within the item bank. Each item will have a loading (i.e. a difficulty level) on the general mathematics scale, but also an indication of the extent to which it measures one or more sub-traits such as geometry. In this case, the entire collection of 'live' items at any point in time defines the effective mathematics curriculum that is being assessed. Uneven patterns of performance that result from the multi-dimensional nature of mathematics achievement show up in departures from the Rasch model. These departures (or residuals once the model estimates have been subtracted from the data) can be analysed to provide the details of the *n*-dimensional pattern of performance. Mead (1976), in the United States, has developed some useful pointers here. The results of the testing then, whether for an individual pupil or a large group of pupils, can be expressed in terms of an overall level of performance in mathematics together with a profile showing relative areas of strength and weakness. If it is decided to redefine the trait structure either by combining existing traits or by splitting to develop new ones, then this is readily accomplished merely by re-classifying the items. Past data sets can then be reanalysed in terms of the new structure in order to look for evidence of change.

In case all this seems rather abstruse, let me try and summarize here. Over the years I expect mathematics performance to change. Not only will its overall level move up or down when measured against the current requirement being placed on the school system, but also the very structure of the curriculum, and the pattern of emphases placed on various component parts, will be continuously changing. Our *n*-dimensional cigar will be slowly changing its shape in ways in which it would be virtually impossible to predict. We want not only to be able to identify the existence of these changes, and their general direction, but also to quantify them in real performance terms. We want to say not only whether performance in school mathematics is going up or down, but how the definition of school mathematics as evidenced by the pattern of performance of pupils is responding to the changing needs of society.

It is too early to say if we shall be able to achieve this. We are still in the process of establishing base lines from the initial surveys, but if you should feel inclined to invite me back in ten years time I may have something more definite to report.

ITEM BANKING

That I leave this topic to the end does not imply any lack of importance. It is my belief that the future of educational testing lies largely with item banks. They clearly will have a much wider application than just in national monitoring programs, although such programs are currently providing the substantial resources necessary for item bank development. I chose not, in this paper, to concentrate on the general issues surrounding item bank use, preferring to consider in some detail alternative measurement models, but most of the points I have discussed will have direct relevance to the item bank user.

The great advantage of an item bank lies in its flexibility of operation. The simple and rapid construction of custom-tailored parallel tests to order—long or short, hard or easy, wide or narrow, all with known psychometric properties and with calibration tables generated automatically—can improve the quality and impact of educational testing everywhere.

As some of you know, I am involved in trying to set up an international network of item banking centres which will exchange technical know-how and actual test materials between different nations. Why are we trying to do this? Because to be really effective item banks need to be large, and this means they will be expensive to create from the beginning. Sharing materials can save money. Further, the existence of internationally agreed criteria and conventions for classifying items, for reporting psychometric parameters and so on will greatly assist cross-national comparisons, evaluation, and accreditation.

On the smaller scale, my particular hope is that item banking can be developed to provide the classroom teacher with really good diagnostic instruments to clarify the learning difficulties of his or her pupils. In the past this has tended to be a neglected area because of its inherent difficulties, but latent trait scaling coupled with item banking may provide the answer.

REFERENCES

Andersen, E. B. *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forskningsinstitut, 1973.

Choppin, B. H. *Item banking and the monitoring of achievement*. (Research in Progress Series No. 1). Slough: NFER, 1978.

Comber, L. C. and Keeves, J. P. *Science education in nineteen countries*. New York: Wiley, 1973.

Eysenck, H. J. *The structure and measurement of intelligence*. Berlin: Springer-Verlag, 1980.

Goldstein, H. Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 1979, 5, 211-20.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, 1952, 7.

Mead, R. The assessment of fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago, 1976.

Middleton, W. E. *A history of the thermometer*. Baltimore: Johns Hopkins Press, 1966.

Nelkon, M. and Parker, P. *Advanced level physics*. London: Heinemann, 1968.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement*. Chicago: MESA Press, 1979.

REACTANT STATEMENT

Glen A. Smith

I welcome this opportunity to contribute to this invitational seminar, by commenting on Dr Bruce Choppin's interesting and practical paper. While this is a conference on measurement, it is directed to a particular field, education and psychology, and as such it should touch on practical aspects and examples. Dr Choppin's paper does this, while still being soundly based on theory. We have been shown glimpses of latent trait modelling in the field (or rather, classroom) and it is here I feel that the model should face its closest checks. Its utility must go beyond keeping the mathematically inclined among us happy and debating.

Dr Choppin raised the analogy of temperature as with accepted measurement properties, but let me take the analogy a little further. As practical measurers, we are more concerned with fuzzier traits, like 'comfort'. Temperature is certainly one dimension of comfort—all people are uncomfortable at very high and very low temperatures. If we were to use a thermometer, with its good metrical properties, to measure comfort, we would miss our mark, while getting data that fitted the Rasch model. Some people are comfortable at 20°C, others at 26°C (it probably would not be 'culture fair', in any case). We still need to be critically aware of the name/trait distinction, as I am sure Dr Choppin is, but I see the point overlooked so often that it is worth reiterating.

I am interested to hear that NFER is using a latent trait model to measure achievement—an area where I think it may be less applicable than others, for learning does presuppose exposure to the material tested, and this can differ between groups, e.g. schools, and give data not fitting the Rasch model while still truly measuring achievement. This question can be treated empirically, and perhaps the data being collected by NFER will show the assumptions made to be warranted. It should give a test of the robustness of the model; if it uncritically fits any data, with deletions of the occasional item, we need to think deeply about what the model is giving us. This is possibly more relevant to think about for diagnostic testing—an area that I think does not need Rasch modelling, and by its nature—identifying low achievement areas for individuals—possibly does not hold the essential assumptions.

Dr Choppin gave details of several other interesting applications and extensions of the model which are in the exploratory stage, and I will be interested to follow their developments, especially the analysis of incomplete data matrices with its exciting application to multiple raters, a common problem in assessment. I am also interested in hearing more about the index, b_{ii} b_{ni}, especially its sensitivity to deviations from the model.

I would finally like to reinforce Dr Choppin's comments on item banking with its close ties to computerization of testing, something of high interest to me. I agree that the future of educational testing will be tied closely to item banking, and I hope that it can work worldwide, drawing internationally on the expertise of NFER and like bodies.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

4

The Linear Logistic Test Model and its Application in Educational Research

Hans Spada and Regine May

INTRODUCTION

One of the central questions of educational research with regard to test data is the assessment of learning effects. Psychometric analyses based on the Rasch model (Rasch, 1966) avoid some pitfalls of applying classical test theory (cf. Fischer, 1974; Rost and Spada, 1978). But this approach, which results in the measurement of the variable 'student ability' and its change over time and also the variable 'item difficulty', is still deficient in several ways.

First, this approach gives no answer to the question of how differing problem difficulties can be explained from a cognitive psychological viewpoint. Second, it gives no analysis of how changes in individual ability are to be understood. In other words, a theory and method for analysing item difficulty and ability change in a psychological and educational context is lacking. This paper demonstrates a way of overcoming these deficiencies. At the same time it should be emphasized that the main problem of applying the Rasch model and models based upon it is that they are at variance with the assumptions of several well-known psychological theories of learning and development. This will be demonstrated later in the paper. As yet, insufficient attention has been directed to these questions in English-speaking countries.

In Austria and Germany in the early seventies, Rasch's ideas were extended by developing logistic models to go beyond the quantification of item difficulties and student abilities (Fischer, 1973; Scheiblechner, 1972; Spada and Scheiblechner, 1973). The intention was to represent explicitly the effects of thinking and learning processes by means of these models.

In the following pages, one of these probabilistic test models will be described, the so-called Linear Logistic Test Model (LLTM). The LLTM was first discussed by Scheiblechner (1972). Cox (1970) proposed a similar model but without the explicit notion of interindividual differences. Fischer (1973; 1974; 1976) studied the statistical properties of the LLTM and derived the estimation equations on which the computerized algorithm (cf. Fischer, 1974) is based; this was also used in our studies. Fischer (1978) gives a thorough review of the work done in this area.

THE LINEAR LOGISTIC TEST MODEL (LLTM)

The basic idea which led to this model is the following:

1 The difficulty of the items of a test, e.g. an achievement test, is traced back to ('explained' by) the number and difficulty of the cognitive operations needed and used for their solution; and/or

2 the effect of the conditions of teaching and learning (e.g. instructional measures) with which the subjects were faced before taking the test (and possibly also of those conditions which were of relevance while the test was taken) are quantitatively assessed.

The LLTM makes it possible to realize this idea by decomposing the item parameters into linear combinations of more elementary parameters corresponding to the difficulty of cognitive operations or to the effect of instructional measures, etc. The estimation of the elementary parameters is based on the same principles as the estimation of the item parameters in the Rasch model, and the validity of the decomposition can be tested statistically by methods similar to those used for testing the fit of the Rasch model (cf. Andersen, 1973).

The LLTM is a Rasch model with an additional marginal condition. Therefore the model is characterized by the following equation:

$$p(+ \mid v, i) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)} \qquad (1)$$

$$\text{with } \sigma_i = \sum_{j=1}^{m} f_{ij}\,\eta_j + c$$

The probability that student v solves item i correctly is represented, according to the Rasch model, as a logistic function of two parameters, namely ξ_v, characterizing the ability of the student to solve problems of this kind, and σ_i, characterizing the difficulty of the item. But what is denoted by f_{ij} and η_j? In the context of analysing the problem-solving process, the item parameter σ_i is seen as a linear function of the number and difficulty of the cognitive operations leading to a correct solution. Therefore, in this case, f_{ij} denotes the hypothetical frequency with which operation j is needed. The parameter η_j characterizes the difficulty of operation j and c is a normalizing constant. In a later part of this paper we shall see that in studies in educational evaluation the parameters η_j are often introduced to quantify the effect of instructional measures on the item difficulties.

Since the LLTM is a Rasch model with a linear marginal condition, it shares many of its characteristic features with this model. The number of correct responses is a sufficient statistic for the ability parameter, just as in the Rasch model. The structural parameters η_l are estimated by a set of conditional maximum likelihood equations, which do not include the ability parameters. The estimates of the parameters η_l are therefore 'sample free'. in the same sense as the item parameter estimates in the Rasch model. A precondition for an application of the LLTM is that the matrix $F = (f_{ij})$, which is a k (number of items) times m (number of elementary parameters) matrix, is (a) of rank m and (b) specified before the estimation procedure.

To provide detailed information about the advantages and the problems of applying the LLTM, three different empirical studies which were carried out at the Institute of Science Education at Kiel (West Germany) are summarized. In the first study the LLTM was used as a model of thinking and intellectual development in the area of balance scale tasks (Spada, 1976; Spada and Kluwe, 1980). In the second study the effect of different instructional measures was estimated in connection with an instructional unit on nuclear power plants (Spada, Hoffmann, Lucht-Wraage, 1977). In the third study the LLTM was used to develop an instructional unit on problems of 'recognizing functional relationships' *and* to assess its effects (Häussler, 1978).

The discussion of these investigations will show that there are problems in the assumptions of the Rasch model itself.

THE LLTM AS A MODEL OF THINKING AND INTELLECTUAL DEVELOPMENT

Structural Assumptions

In the first study to be reported here, the development of the concept of proportion was investigated by means of the LLTM and a deterministic model of qualitative change (Spada, 1976; Spada and Kluwe, 1980). Only the results of balance scale problems analysed by means of the LLTM will be discussed. These problems represent one form of proportional tasks. They have been frequently used by developmental psychologists since Inhelder and Piaget (1958).

By studying the relevant Piagetian literature and by observing children who solved balance scale tasks, hypotheses were deduced about the cognitive operations applied by children in reaching the correct solution, and a sample of balance scale tasks with specified task structures was constructed. The term 'psychological structure' of a task denotes in this

Table 1 Four (of Eight) Cognitive Operations Assumed to be Relevant for the Solution of the Balance Scale Tasks Used in the Investigation

	Operation
1	Attention to and deduction from different amounts of weights
2	Attention to and deduction from different lengths of the lever arms
3	Compensation of a change of the amount of one weight or the length of one lever arm in the same modality on the other side of the bar
4	Compensation of a change in the other modality on the other side of the bar

context the type and number of cognitive operations which enable a person from a certain population to solve a task.

Altogether a set of eight cognitive operations for solving balance scale items was defined. Table 1 shows four of these operations.

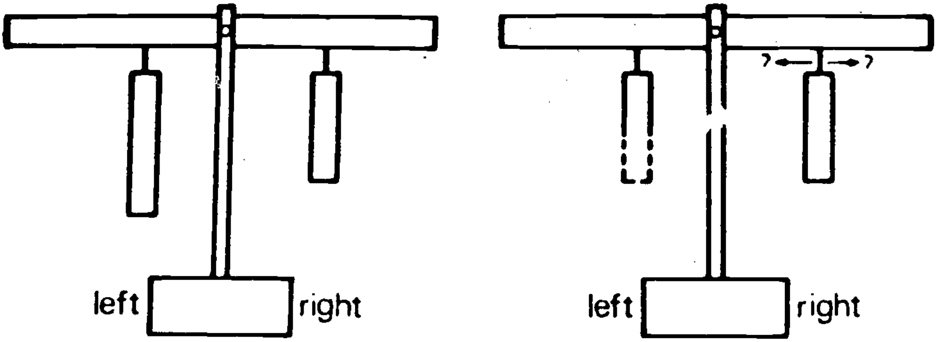
It was attempted to present tasks whose solutions involved certain subsets of the eight postulated cognitive operations. Twenty-four tasks corresponding to different combinations of these operations were constructed. Figure 1 shows one of these tasks. It is supposed that operations 1 and 4 are relevant in the solution of this task. It was hypothesized that the student was thinking in the following way: because the weight on the left side is reduced, the bar will be unbalanced. To compensate for this change on the left side, the weight on the right side has to be moved inwards. Analogously, other tasks were related to other combinations of operations.

For every item i, a vector of the task structure f_i was defined, consisting of ones and zeros, where a one at the j-th position denotes the presence of operation j in item i and a zero denotes its absence. The structure of the sample item referred to in Figure 1 is (1, 0, 0, 1, 0, 0, 0, 0). The hypothetical structure of all items under study can be summarized in a task structure matrix $F = \|f_{ij}\|$.

Quantitative Developmental Assumptions

In terms of the LLTM, the development of the ability to solve proportional tasks can be considered to take the following form.

Developmental change is reflected in the LLTM as a purely quantitative change in the student parameters ξ_v. Development then is comparable to global learning. This type of learning leads to higher solution probabilities for all tasks of one homogeneous class of problems. For all children and adolescents under study (with regard to the investigation on balance scale tasks, the age range between 11 and 16 years) it is supposed, therefore, that the task structure remains constant and that the operation parameters are invariant. That is, correct solutions are assumed to be



The drawing shows a balance scale with weights. The weights are hung in such a way that the scale is in equilibrium.

Now the weight on the left-hand side of the bar is decreased. In order to keep the bar in equilibrium, the weight on the right-hand side of the bar must:

stay in the same position. □
be hung further inward. □
be hung further outward. □
I do not know. □

Figure 1 A Balance Scale Problem Used in the First Investigation
It is hypothesized that Operation 1 and Operation 4 are applied to obtain a correct answer (Adapted from Spada and Kluwe, 1980, Figure 1.1.)

based on the same cognitive operations irrespective of age; the operations are assumed to have a constant rank order with regard to difficulty.

Figure 2 presents the functional relationship that is expected on the basis of the model equation between task solution probabilities and children's task solution ability, the latter corresponding to the developmental level. Given medium person ability, the probability of a correct solution of the structurally most complex of the three tasks is very small. With higher person abilities, the solution probabilities of the three tasks approach each other and become approximately one for very high values. The functional relationship can be understood to represent a more precise version of a model proposed by Flavell (1971) to describe the developmental change of intellectual abilities.

Empirical Findings

In the investigation reported here, a pencil and paper test was used (Spada, 1976). Twenty-four balance scale tasks were given. The sample included 949 male and female students from ages 11 to 16 attending Ger-

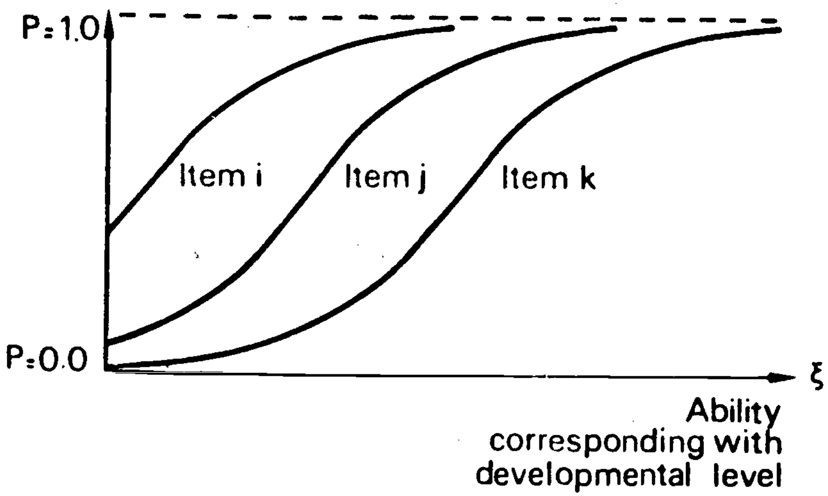


Figure 2 The Functional Relationship Postulated in the LLTM between Task Solution Probability and Person Ability (or 'Developmental Level')

The item characteristic curves of three tasks with the following task structure vectors f_i = (01000000), f_j = (01100000) and f_k = (01100100) are shown. The abscissa distances between the three item characteristic curves—which are parallel to one another—were computed from the data and reflect the differing task difficulties. (Reprinted from Spada and Kluwe, 1980, Figure 1.4)

man secondary schools. The test was administered in classrooms. (In another investigation real balance scales were used in individual sessions (Spada and Kluwe, 1980)).

Conditional likelihood ratio tests were used to test the assumption of sample free parameter estimates and thus the validity of the Rasch part of the LLTM. Some of the tests showed significant divergences between the estimates of the item parameters computed from the data derived from different groups of students. It must be emphasized, however, that the parameter estimates did not differ widely and that the statistical tests indicated significant divergences essentially because of the large sample of subjects (significant results would not have been obtained with less than approximately 670 subjects). There is the question, nevertheless, of what shortcomings might be responsible for these results of testing the Rasch part of the LLTM.

In the next step, the operation difficulty parameters η_i were estimated. Based on these parameter estimates and the task structure hypotheses f_{ji}, estimates of the item parameters σ_i were computed and compared with those item parameter estimates resulting from an application of the parameter estimation algorithm of the Rasch model itself. A graphical comparison indicated a good correspondence between the two different sets of parameter estimates. However, a conditional likelihood ratio test of the linear marginal condition of the LLTM showed that the differences were significant. This meant that at least some of the task structure hypotheses were not valid and/or that the formalization of the hypotheses by means of the linear logistic model was not without problems.

Quite another approach to testing the validity of LLTM results on operation difficulties was taken by Nährer (1977, cf. also Fischer, 1978). He constructed new items for some of the tests which had been analysed by means of the LLTM in various research studies. Nährer predicted the difficulty of the new items by using the published results on the estimates of the operation difficulties and the task structure hypotheses. The new items were then given to a new sample of subjects and the item difficulty parameters were estimated from these data by means of the Rasch model. Nährer reports a good correspondence between both groups of parameter estimates in most of the cases, especially for tasks from the field of mechanics, e.g. rotation mechanism problems (Spada, 1977). It was possible to predict the difficulty of the newly constructed items quite well, although in this case also the fit of the LLTM was far from perfect. Unfortunately no reanalysis was carried through for the balance scale problems discussed in this paper.

Nährer's study indicates another interesting field for applying the LLTM, namely the controlled construction of items with predictable difficulty, e.g. for item banks and individualized testing (cf. Fischer and Pendl, 1980). The LLTM based on valid task structure hypotheses allows us to define in a precise way what might be understood by the notion of a 'domain of tasks'. It is the homogeneous class of items which can be constructed on the basis of the set of analysed operations.

Shortcomings of the LLTM *and* the Rasch Model

There are, of course, aspects of the task structure hypotheses which are open to question. They do not encompass all features of the problem solving process. Nothing is stated about sequential or temporal characteristics. Encoding, decoding, and memory features are virtually neglected in their present state of development. But the different tests of fit of the LLTM have shown that it is necessary to discuss some of the assumptions of the structural part (the marginal condition) *and* the Rasch part of the LLTM in more detail.

There is a serious drawback to the LLTM when applied in this context. The task solution probabilities cannot be understood as the products of the corresponding operation probabilities, because in general

$$\frac{\exp(\xi_v - \sum_{j=1}^{m} f_{ij}\eta_j + c)}{1 + \exp(\xi_v - \sum_{j=1}^{m} f_{ij}\eta_j + c)} \neq \prod_{j=1}^{m} \left\{ \frac{\exp(\xi_v - \eta_j)}{1 + \exp(\xi_v - \eta_j)} \right\}^{f_{ij}} \tag{2}$$

This contradicts the 'product rule', which is a familiar assumption in probabilistic automata theory, as proposed by Suppes (1969) and applied, for example, in the analysis of arithmetic teaching in elementary school, by Suppes and Morningstar (1972). The product rule states that the probability that a student solves an item correctly is the product of the probabilities that he carries out correctly all operations necessary for the solution (for a thorough discussion, see Spada, 1977).

Equation (2) makes it clear that this problem is not specific to the LLTM but also occurs in the Rasch model itself if that model is applied—as is usual—at the level of item data and not at the level of operation data. It may be that some of the negative results in applying the Rasch model in this study and in other investigations are due to the fact that the task solution probabilities cannot be understood as the products of the probabilities of a correct application of the corresponding operations. Furthermore the assumption of a *linear* combination of the operation parameters in the logistic function is rather arbitrary in the sense that it does not reflect psychological hypotheses about underlying cognitive processes. Statistical reasons were decisive in the choice of this type of function, since only the logistic function in the framework of the Rasch model involves the important advantage of 'sample-free' measurement.

Another criticism of the LLTM can be made from the viewpoint of developmental psychology. The assumptions that the task structure is the same for all children in the sample and that developmental change can be represented as a quantitative change of the person parameters are at variance with numerous developmental findings. Again this criticism applies equally to the LLTM and the Rasch model, if the models are used to analyse data from children of differing age levels.

The balance scale tasks of this investigation are in some respects similar to those analysed in the experiments of Piaget and his co-workers (Inhelder and Piaget, 1958). In addition, their results inspired the formulation of the task hypotheses which were tested in this study. While

Piaget's theory assumes that the solution algorithms vary and become more and more complex in the course of development, applications of the Rasch model and usually also of the LLTM are based on the assumption that correct solutions result from the same cognitive operations for all individuals tested.

In Piaget's theory, it is emphasized that qualitative structural changes take place in intellectual development. In applications of the LLTM, inter- and intra-individual differences are often understood as differences in the degree of mastery of the *same* solution algorithm, in contrast to developmental theories, where these differences are explained in terms of *different* solution algorithms.

In the field of information processing theories, the work of Siegler (1976) is relevant to the present discussion. In his rule assessment approach, cognitive development is also characterized as the acquisition of increasingly powerful rules for solving problems. In his extensive study of problem solving with balance scale tasks, Siegler (1976) postulated four different algorithms, each algorithm corresponding to one developmental stage. His theory predicts for every stage a certain pattern of correct and false answers. These answer patterns are used to assess the developmental level (with regard to this class of tasks) of each child. One of the more interesting findings which substantiates Siegler's assumptions is that children moving from stage II to stage III show a striking decrease in the number of correct answers with regard to one class of items, the so-called conflict-weight-items. It is assumed that at this developmental level the children have learned to pay attention at the same time to both dimensions of balance scale tasks, namely weight and distance, but do not yet know exactly how these variables are related. The consideration of both dimensions without exactly knowing how to combine them leads to an increase in the number of incorrect answers, because the sub-class of conflict-weight-items was constructed in such a way that the items can be answered correctly (without full insight into the problem) by reference to weight alone.

The Rasch model does not fit data of this type, nor does the LLTM, if only one matrix of task structure hypotheses is assumed for all subjects of the sample studied (cf. May, 1979). In principle, it would be possible to represent such structural changes (e.g. the acquisition of new sequences of operations) in the LLTM. This could be done by specifying different task structure hypotheses for different children and for the same child at different stages of development or learning. A prerequisite would be to have well-founded hypotheses about such structural changes with regard to each subject.

Empirical falsification of the Rasch model and the LLTM, which would be inevitable with this type of data, could also be avoided by excluding all tasks from the test sample whose difficulty does not decrease in a monotonic manner with age. This approach seems defensible from a diagnostic viewpoint. It is more problematic if the main interest is in a cognitive analysis of the developmental process.

In summary, it can be said that structural changes of the problem solving process caused by development or learning contradict the homogeneity assumption of the Rasch model and the LLTM (with only one task structure matrix). If inter- or intra-individual differences result from such structural changes in a sample of subjects, deviations will be detected in graphical and statistical tests of the model.

We refer finally to two groups of psychological models of human knowledge, and of its acquisition, storage, and use. These relate to models based on semantic networks (Dorner, 1976; Norman and Rumelhart, 1975) and on production systems (Newell and Simon, 1972) (cf. also Anderson, 1976; Greeno, 1978). In these models cognitive processes are represented in such a way that the assessment of structural changes is also of special importance in the measurement of change.

As a consequence we have to face the fact that the great majority of psychological developmental and learning theories postulate cognitive changes which would make the emergence of homogeneous item samples an exceptional, surprising result. In reality, however, the multiplicity and complexity of factors influencing test behaviour often lead to a falsification of these theories and to a reasonable fit of the probabilistic Rasch model under appropriate item construction conditions.

EDUCATIONAL EVALUATION: EXPERIMENTAL DESIGNS WITH BINARY DATA

In educational evaluation some parts of the learning history of each individual are usually known. The central aim of such investigations is often the assessment of the effects of different teaching strategies on the learning outcome. The most relevant type of learning effects, which can be assessed by means of the LLTM and traced back to instructional factors, are global learning effects. Global learning leading to higher solution probabilities for all tasks of one homogeneous class can be represented in the model either as an increase in the value of the person-parameters (i.e. individual abilities to solve tasks of this type correctly) or as a general decrease in the values of the item-parameters (i.e. item difficulties). For technical reasons we shall use the second form of representation of global learning effects.

Let us consider the following very simple experimental educational design. An instructional unit is applied in two different variants in two samples of students. A third sample (control sample) does not receive this type of instruction. Each sample comprises about four classes. Of interest is the effect of each of the instructional methods on the improvement of some ability or skill of the students referred to in one of the learning objectives. This ability or skill is assessed by means of samples of items of one homogeneous item class given before instruction (pre-test) and after instruction (post-test). (For a detailed discussion of evaluation problems of this type, see Rost and Spada, 1978.)

Equation (3) shows—for the general case of t tests (test 1 = pre-test)—how the LLTM might be applied in problems of this type. The linear marginal condition is introduced to quantify the effects of the instructional methods under study. The difficulty of the items of test t (e.g. the post-test) is traced back to the difficulty of the items of the pre-test (before instruction), to the effects of the instructional methods and to a trend parameter, characterizing non-instructional general effects between the pre-test and test t on the ability under study. Table 2 illustrates for our simple example how the corresponding matrix F of the LLTM is set up.

$$p(+ \mid v, i, t) = \frac{\exp(\xi_v - \sigma_{it_{(v)}})}{1 + \exp(\xi_v - \sigma_{it_{(v)}})} \tag{3}$$

$$\text{with } \sigma_{it_{(v)}} = \sigma_i - \delta_{t_{(v)}}$$

$$\text{and } \delta_{t_{(v)}} = \sum_{a=1}^{s} f_{t_{(v)}a}\eta_a + \tau_t$$

$p(+ \mid v, i, t)$ is the probability that student v solves task i correctly at test (time) t.

ξ_v is the ability parameter of student v.

$\sigma_{it_{(v)}}$ is the difficulty of item i at time t (test t) after that type and amount of teaching (i.e. instructional method in our example) that took place in the class with student v between test 1 and test t.

σ_i is the (hypothetical) difficulty of item i before instruction (test 1).

$\delta_{t_{(v)}}$ characterizes the effect of that type and amount of teaching (i.e. instructional method in our example) that took place in the class with student v between test 1 and test t (with $\delta_1 = 0$, $t = 1$ . . . pre-test).

η_a characterizes the effect of instructional variant $a (a = 1, s)$ on the ability.

$f_{t_{(v)}a}$ denotes the amount of teaching method a between test 1 and test t (in our example, 1 denotes that the method was used with student v, 0 that it was not used).

τ_t characterizes non-instructional general effects between test 1 and test t on the ability, and is a trend parameter.

In the example,

$\delta^{*}_{i(t)} = \eta_1 + \tau_t$ in the case of instructional variant 1, and
$= \eta_2 + \tau_t$ in the case of instructional variant 2, and
$= \tau_t$ with no instruction.

Based on this approach, it is possible to compute 'sample free' estimates of the effect parameters of the instructional methods and of the trend parameter, even in the absence of random sampling with approximately equal ability distributions in the different sub-samples of students. This property is of great importance in our example and in general in educational evaluation, because instructional methods are usually tested with classroom groups, that is, the analysis is based on cluster sampling. Cluster sampling leads to an underestimation of error variance in analyses of variance and thus to an overestimation of the statistical significance of the instructional effects. Using the LLTM, the significance of the effects and of differences between them can be tested statistically by means of conditional likelihood ratio tests. These tests do not rely upon the variance of the ability parameters; they are conditional tests, in which the ability parameters do not even enter.

The use of the LLTM in this context has the additional advantage (common to most of the Rasch model applications) that it is not necessary to give the same test at the different points in time. Provided that some items are given repeatedly, the other test items can be selected in such a way that their difficulty is adapted to the achievement level of the students at the time the test is given. If one is not interested in the general trend effect, but only in the question of the differences of the effects of the instructional methods, it is even possible to present different item samples in the different tests, as long as all items measure the same ability, thus meeting the assumption of homogeneity.

This approach was also used by Spada, Hoffmann, and Lucht-Wraage (1977) to evaluate the effects of an instructional unit and of four additional instructional measures. The instructional unit was entitled 'Nuclear Power Plants—Dream or Nightmare?'. A Rasch-scaled situation test was developed (cf. Spada and Lucht-Wraage, 1980), which made it possible to assess several variables simultaneously, which corresponded with the objectives of the instructional unit. The four groups of instructional measures were: introduction of a model person, activation of alarm in subjects, the stabilization of attitudes, and group instruction on the basis of interaction structure analyses.

The usual procedure in empirical curriculum development is to lump all good ideas together and to measure their joint effect by evaluating the resulting instructional unit. In this particular classroom experiment, a different approach was chosen. With the traditional ideal of a factorial design in mind, the four measures were combined. Each of the 16 combinations (– , – , – , – ; + , – , – , – ; . . . ; + , + , + , +) was then applied in at least one class.

The LLTM was used to analyse the data for all students in the 22 classes involved in the experiment and to estimate the effectiveness of each instructional measure, the general effects of the instructional unit, and the item difficulties and student abilities for each of the variables of the situation test. The results of the study cannot be presented here because of space limitations, but are available elsewhere (Spada, Hoffmann, and Lucht-Wraage, 1977).

Mention must be made of two serious drawbacks in applying the LLTM in this way in the context of educational evaluation.

1 In contrast with the application of multivariate analysis of variance, no simultaneous analysis of all dependent variables is possible.

2 It has to be assumed that the learning effects are not person-specific but depend only on type and amount of teaching. In other words, the model is only valid if all inter-individual differences in learning outcome which are not attributable to ability prior to instruction can be traced back to global effects of the individual instruction or learning histories of the students in the course of the educational experiment. This assumption of global learning, which refers to the level of the parameters employed and not directly to the reactions of students, is quite restrictive and should be tested when the LLTM is applied in this manner.

Fischer (1977; 1978) referred to another problem in connection with the use of the LLTM and the Rasch model, namely, the restrictive assumption of item homogeneity, and developed similar logistic models, the Linear Logistic Models with Relaxed Assumptions (LLRA), which are not based on this assumption.

INSTRUCTION AND EVALUATION: THE STUDY OF HÄUSSLER (1978)

A study undertaken by Häussler will serve as a final example of an application of the LLTM in an educational context.

Haussler (1977; 1978; 1981) developed and evaluated two different teaching programs to improve the ability of adolescents to solve tasks of the type 'recognizing functional relationships'. He made use of both the structural aspect of basing teaching on task structure hypotheses and the assessment aspect. Some of the different ways of applying the LLTM, as discussed in the preceding pages, were therefore combined in his investigation.

Häussler's investigation was based on 356 students in a pilot study and 1037 students in the main study, aged 12 to 16. He used the LLTM firstly to describe by means of task structure hypotheses the constituents of the solution algorithms used by the students to solve the tasks, and secondly to measure the effects of the teaching programs.

The task structure hypotheses were deduced by observing and interviewing students solving such problems and by considering some of the conceptions of Scandura (1973). The hypotheses were then tested by means of the LLTM. In line with the preceding discussion of shortcomings of the LLTM as a model of thinking, it is not surprising that the fit of the LLTM (and of the Rasch model) again proved to be not satisfactory in some cases. On the other hand the contribution of the task structure analyses in producing a basis for psychologically well-founded teaching methods was substantial.

Those problem solving operations which were used by students to solve individual problems correctly, before any training in this special field was provided, were denoted by Häussler as 'spontaneous' algorithms. The first teaching program was based on these spontaneous algorithms. All of the algorithms identified by the task structure analyses have one procedure in common: they involve manipulation of the data in such a way that an invariant quantity is produced. This common procedure was used to synthesize algorithms which include many of the spontaneous algorithms as special cases (cf. Häussler, 1978). As part of the second teaching program these 'synthetic' algorithms were taught; they are more comprehensive, theoretically superior, higher-order algorithms.

Haussler (1978, 1981) ascertained that both programs yielded statistically significant, substantial and relatively long-lasting positive effects. Figure 3 summarizes the results of the estimation of the effects of the two teaching programs. The LLTM was used to estimate 2 (teaching programs) × 4 (points in time of testing) × 3 (subsamples of tasks) = 24 instruction effect parameters δ (cf. Equation 3).

The students were tested prior to instruction (T_1), immediately after instruction (T_2), and six weeks after instruction (T_3) in either teaching program A (spontaneous algorithms) or teaching program B (synthetic algorithms). Some students were given a short refresher program after the six-week period; these are designated (T_4) in place of (T_3). Three groups of tasks were given in the four testing phases. Group 1 tasks were used during instruction to practise the different algorithms. Group 2 tasks could be solved by an algorithm similar to the one learned during instruction. Group 3 tasks could be solved only by inventing a new algorithm. A most interesting result was that the teaching of the spontaneous algorithms turned out to be more effective with Group 3 tasks. Presumably, this strategy had forced the students at the outset to consider the possibility of being confronted with new problems and to

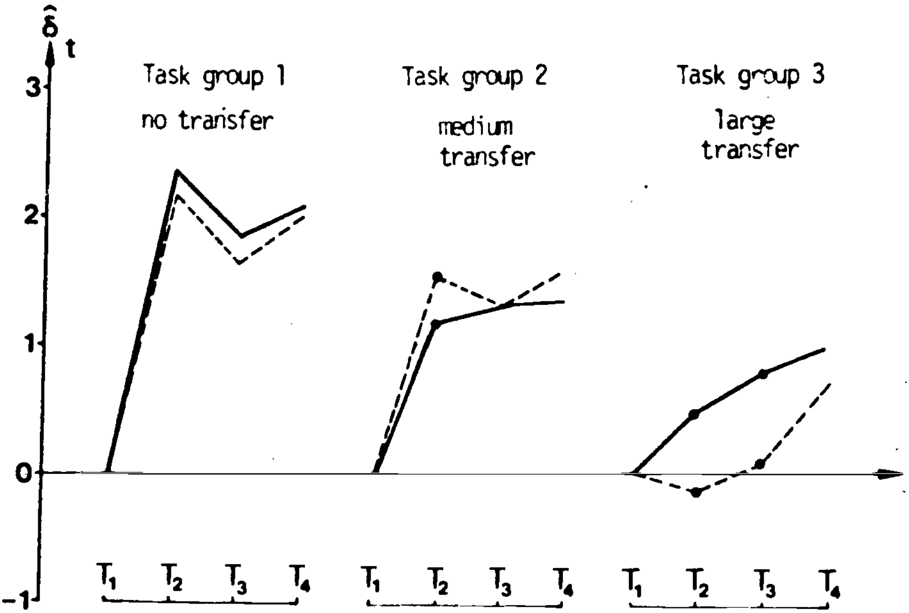


Figure 3 Graphical Representation of Instruction Effects δ_t
The solid line corresponds to teaching program A (spontaneous algorithms), the broken line to teaching program B (synthetic algorithms). Significant differences between A and B are marked with a dot. (Reprinted from Häussler, 1978, Figure 6.)

develop solution algorithms on their own—along the lines of the learned spontaneous algorithms—in order to cope with these problems. The presence of interactions between teaching effects and certain subsamples of tasks draws attention to the restrictiveness of the usual assumption in applying the LLTM and the Rasch model, namely that learning effects are postulated to be constant for all items.

SOME CONCLUDING REMARKS

After twenty years of development and discussion the logistic models originating in the basic ideas of Rasch are largely accepted as valuable tools in educational measurement. In recent years, however, doubts have been expressed by several authors about the validity of these models as psychological models of cognitive processes in learning and development. In this paper we have tried to consider questions of educational measurement and of psychological theorizing simultaneously. We have demonstrated that the LLTM, as one of these logistic models, and the Rasch model itself cannot provide a completely acceptable basis for educational measurement if the various critical psychological arguments are taken seriously.

Nevertheless the LLTM seems to be an interesting tool in cognitive research and educational evaluation, because it makes it possible both to measure inter- and intra-individual differences and at the same time to analyse general regularities which are often hidden behind these differences. In practice, this statement holds only if the restrictive assumptions of the model are not falsified by the data under study. As a consequence, the LLTM and the Rasch model should be applied only in those cases in which the validity of their assumptions is plausible and is tested sufficiently. It is our hope that this paper has added some insight to a more restricted and better controlled use of the LLTM and the Rasch model.

REFERENCES

Andersen, E. B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, **38**, 123–40.

Anderson, J. R. *Language, memory and thought*. Hillsdale, NJ: Erlbaum, 1976.

Cox, D. R. *The analysis of binary data*. London: Methuen, 1970.

Dorner, D. *Problemlösen als Informationsverarbeitung*. Stuttgart: Kohlhammer, 1976.

Fischer, G. H. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 1973, **37**, 359–74.

Fischer, G. H. *Einfuhrung in die Theorie psychologischer Tests*. Bern: Huber, 1974.

Fischer, G. H. Some probabilistic models for measuring change. In D. de Gruijter and L. van der Kamp (Eds), *Advances in psychological and educational measurement*. New York: Wiley, 1976, 97–110.

Fischer, G. H. Some probabilistic models for the description of attitudinal and behavioural changes under the influence of mass communication. In W. H. Kempf and B. Repp (Eds), *Mathematical models for social psychology*. Bern: Huber and New York: Wiley, 1977, 102–51.

Fischer, G. H. Probabilistic test models and their applications. *German Journal of Psychology*, 1978, **3**, 298–319.

Fischer, G. H. and Pendl, P. Individualized testing on the basis of the Rasch model. In L. van der Kamp, W. F. Langerak, and D. N. M. de Gruijter (Eds), *Psychometrics for educational debates*. New York: Wiley, 1980, 171–88.

Flavell, J. Stage-related properties of cognitive development. *Cognitive Psychology*, 1971, **2**, 421–53.

Greeno, J. A study of problem-solving. In R. Glaser (Ed.), *Advances in instructional psychology*, (Vol. 1). Hillsdale, NJ: Erlbaum, 1978, 13–75.

Haussler, P. Investigation of mathematical reasoning in science problems. In H. Spada and W. F. Kempf (Eds), *Structural models of thinking and learning*. Bern: Huber, 1977, 263–79.

Haussler, P. Evaluation of two teaching programs based on structural learning principles. *Studies in Educational Evaluation*, 1978, **4**(3), 145-61.

Häussler, P. *Denken und Lernen Jugendlicher beim Erkennen funktioneller Beziehungen*. Bern: Huber, 1981.

Inhelder, B. and Piaget, J. *The growth of logical thinking from childhood to adolescence*. New York: Basic Books, 1958.

Kluwe, R. and Spada, H. (Eds). *Developmental models of thinking*. New York: Academic Press, 1980.

May, R. *Wie entwickelt sich das Verständnis von Proportionalitat?* Diplomarbeit (unpublished), Universität Konstanz, 1979.

Nährer, W. *Modellkontrollen bei Anwendung des linearen logistischen Modells in der Psychologie*. Philosophische Dissertation (unpublished), Universität Wien, 1977.

Newell, A. and Simon, H. *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Norman, D. and Rumelhart, D. (Eds). *Explorations in cognition*. Reading, Berks: Freeman, 1975.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, **19**, 49–57.

Rost, J. and Spada, H. Probabilistische Testtheorie. In K. J. Klauer (Ed.), *Handbuch der Padagogischen Diagnostik*. (Bd. 1). Düsseldorf: Schwann, 1978, 59-97.

Scandura, J. M. *Structural learning I: Theory and research*. New York: Gordon & Breach, 1973.

Scheiblechner, H. Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift fur experimentelle und angewandte Psychologie*, 1972, **19**, 476–505.

Siegler, R. S. Three aspects of cognitive development. *Cognitive Psychology*, 1976, **8**, 481-520.

Spada, H. *Modelle des Denkens und Lernens*. Bern: Huber, 1976.

Spada, H. Logistic models of learning and thought. In H. Spada and W. F. Kempf (Eds), *Structural models of thinking and learning*. Bern: Huber, 1977, 227-62.

Spada, H., Hoffmann, L. and Lucht-Wraage, H. Student attitudes towards nuclear power plants: A classroom experiment in the field of environmental psychology. *Studies in Educational Evaluation*, 1977, **3**, 109–28.

Spada, H. and Kluwe, R. Two models of intellectual development and their reference to the theory of Piaget. In R. Kluwe and H. Spada (Eds), *Developmental models of thinking*. New York: Academic Press, 1980, 1-31.

Spada, H. and Lucht-Wraage, H. A paper and pencil situation test to assess attitudes: An analysis of reactions to open-end items based on the model of Rasch. In L. van der Kamp, W. F. Langerak, and D. N. M. de Gruijter (Eds), *Psychometrics for Educational Debates*. New York: Wiley, 1980, 277-89.

Spada, H. and Scheiblechner, H. Stichprobenunabhängige Denkmodelle beim syllogistischen Schlussfolgern. In G. Reinert (Ed.), *Bericht 27. Kongress der Deutschen Gesellschaft für Psychologie Kiel 1970*. Göttingen: Hogrefe, 1973, 868–76.

Suppes, P. Stimulus response theory of finite automata. *Journal of Mathematical Psychology*, 1969, **6**, 327-55.

Suppes, P. and Morningstar, M. *Computer-assisted instruction at Stanford, 1966 1968: Data, models and evaluation of the arithmetic programs*. New York: Academic Press, 1972.

REACTANT STATEMENT

Glenn Rowley

There are two things on which I want to comment. One is the title of the paper, and the other is the paper itself. The title I have had for some time and the paper for very little, so I may give more considered comments on the title of the paper than on its content. Another point I would like to note is that there has been an unannounced change in the title. The program leads us to expect a paper on 'The Linear Logistic Test Model and its Application to Educational Evaluation'. The paper we have before us has 'evaluation' replaced in its title by 'research'. I am not sure whether the alteration was inadvertent or resulted from a change of plans, but it does change one's expectations quite dramatically. While it seems to me that the use of latent trait models in evaluation is an area that, if it proves to be successful, is going to be very important for evaluators, it is also an area in which there are many problems. So I want to begin by making a couple of points about evaluation, and about how it differs from research.

Firstly, as distinct from research, evaluation is an activity which is, or I would argue should be, conducted by or in conjunction with teachers, for the benefit of teachers, and ultimately for pupils. I am not sure that that is always the case with research, and I am not even sure that it should be. I am quite sure that evaluation is not always conducted in that way, and perhaps this has some implication for the kinds of measurement that we can make use of in evaluation. Barry McGaw introduced the analogy of the sailing ship, and I think it can be taken much further. It does seem to me that we have to live with the fact that what we are doing in faculties and schools of education throughout this country is sending our teachers out, not in battleships or even in sailing ships, but in row boats without oars. We rarely provide our teachers with enough training to cope adequately with the demands made on them by traditional norm-referenced measurement procedures. Our teachers usually know little about criterion-referenced measurement (except, frequently, that they are in favour of it), and I seriously wonder whether measurement based on latent trait models can yield results which are meaningful to the practitioner. Yet, if measurement is to have an impact on practice, it must yield results which are meaningful to practitioners, for they are the people for whom we need to provide assistance and with whom we need to communicate.

At the same time, Barry McGaw, I thought, dismissed criterion-referenced measurement rather too easily. I do not think that we can dismiss criterion-referenced measurement at all. It may disappear in the sense that measurement specialists lose interest in it and turn their attention in other directions; but it will not disappear in the sense that teachers are going to keep right on using it. Perhaps they may not call it criterion-referenced measurement, but if, as a teacher, I have taught a given area of content or towards a given set of objectives, the first thing I will want to find out is whether my students can do certain things. No matter whether the measurement experts are there to help me or not, that is what I will be trying to do as a teacher, and that is what my testing will be directed towards. So criterion-referenced measurement will not disappear, even though we may not help the teachers as much as we should.

Secondly, there is an assumption involved in all latent trait models which is fine for measurement and for psychology, but which gets us into trouble when we try to use the models for evaluation. This is the assumption that Professor Thorndike spelt out, to the effect that we are never interested in the behaviours themselves, but only in the underlying trait which those behaviours represent. That very same point of view is expressed at more length in a piece from a test manual published by Educational Testing Service. This is not latent trait modelling; this is traditional classical measurement, as practised in the bastion of norm-referenced measurement, circa 1969:

> When we use a test we are measuring indirectly by taking a series of 'readings' (one for each test question), *not* of the characteristic that we are trying to measure, but rather various indicators of that characteristic. Then we must try to *infer* something about the characteristic itself from the indicators we have collected. In a way a test is like radar, where observations of a series of 'blips' on a screen are used to infer various characteristics of some unseen object. (*SCAT-STEP Series II Teachers Handbook*, 1969)

I want to say that when I am evaluating programs, or when I am evaluating my own teaching, I am very interested in those blips. I want to know whether my students can do these particular tasks and those particular tasks, and if they cannot, what can be done about it. So, from my point of view, the blips on the screen are very important. If the blips on the screen enable me to develop scores which measure an underlying trait that I have managed to increase, or at least stop from decreasing, then that is an added bonus. But I remain very interested in the blips on the screen, and I am not at all ready to dismiss them. It seems to me that, in evaluation, very often teachers want to know in what areas they have succeeded and in what areas they have failed, and the notion of a single latent trait may perhaps not have so much appeal. Certainly I would say that the measurement of a single latent trait has much more appeal to researchers than to evaluators.

A third feature of evaluation is that very often the evaluator is interested in the performance of a group rather than in isolating the performance of a single individual. That, I think, makes evaluation quite often a distinct activity from research which is more inclined to focus on individuals and how an individual tackles a problem. This is the focus of Professor Spada's paper, and this is why I think it appropriate that the title refers to 'research' rather than 'evaluation'. In considering his paper, therefore, I am viewing it as a contribution to research (and I think it is an important one) rather than to evaluation.

Firstly, it does seem to me that the notion (and this is, perhaps, an over-simplification of what has been done) of analysing item difficulties as a way of understanding the processes involved in solving problems and in developing cognitive skills is a very important way of tackling those kinds of problems. What we have seen is a demonstration that item difficulties measured in the metric that the Rasch model provides can be analysed successfully in this way, and that this can lead to useful information and even understanding. What I would like people to think about exploring is whether item difficulties measured in traditional metric, or in other metrics that may be devised, can be treated in similar ways. It seems to me that we have a situation where we are interested in knowing what factors might make an item more difficult or less difficult, and the metric in which we measure item difficulty is something that people can legitimately differ about. I do not know which is the best metric in which to measure item difficulty for these purposes, although some of the properties of the Rasch model may make it particularly advantageous.

In opening the paper, Professor Spada made a comment about the measurement of change. He said psychometric analyses of such data based on the Rasch model avoid the many pitfalls of applying classical test theory. I hope they do, but I am not yet convinced. Every time I listen to its advocates talking about the Rasch model, I have to keep reminding myself that one of the nice properties of the model is that the total score is a sufficient statistic for estimating ability. Another way of putting that is that the ability estimates that you finish up with are really a transformation of the total score—of the number of items answered correctly—and therefore the error of measurement associated with those will be carried along intact through the transformation. Errors of measurement do not go away when you use a latent trait model. Some of the major problems of measurement of change come about partly because errors get confounded with one another, and partly because the errors loom very large in comparison with the amount of change which has taken place. Applying transformations of one kind or another does not remove that problem—it is still going to be there, and I do not know any way of overcoming it. There are also problems of metric when we measure change. One thing that we cannot often do is equate a change of so many points at one part of a scale to a change of so many points at another part of the scale. Given that the use of the Rasch model corresponds to a transformation of metric from one to another, it may well be that those problems are at least reduced, if not eliminated, but I have never seen it argued that this is so.

I suppose the other question that I want to raise is whether the sort of research that has been described could be tackled in other ways, and whether other ways are better – whether, for instance, the same questions could be addressed via variance component analysis on item difficulty indices, be they Rasch model or classical or whatever. Are item difficulties affected by this or that instructional treatment, by this or that characteristic of the item? There are other ways of asking those questions which may lead to better or worse answers. By what criteria do we judge them to be better or worse answers? What I have tried to do here is to raise a number of questions which may be taken up, if of interest, or followed up in quite different directions if that seems more appropriate.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

5

Using Latent Trait Measurement Models to Analyse Attitudinal Data: A Synthesis of Viewpoints

David Andrich

INTRODUCTION

I have chosen to demonstrate how a Rasch latent trait model synthesizes two common approaches to attitude measurement. There are two reasons for this choice. Firstly, because the two approaches to be considered appeared in the literature around 1930, the time the Australian Council for Educational Research was founded, a presentation with some historical flavour seemed appropriate. Secondly, Dr Keeves's statement on objectives for this conference commenced as follows:

> During the past two decades there has been much effort expended by psychometricians in the development and perfection of latent trait measurement models. Yet it is only within the last five years or so that measurement procedures based upon these models have begun to make an appreciable impact on the practice of educational and psychological measurement in Australia. A few practitioners in these disciplines have become acquainted with these procedures, but most still remain unacquainted with the features of the various latent trait models. Consequently, for the most part, traditional measurement procedures, developed during the first half of this century, are still being used.

Therefore it seemed that explicit connections to familiar traditions would make the material less esoteric. The price for this apparent advantage is that sometimes characteristics of the more familiar approaches have to be rearranged and viewed from a somewhat different perspective.

After a brief review of the two traditional approaches, the main model of the paper, the Rasch model for ordered response categories which is called *the rating response model*, is presented. Because it has been presented elsewhere in more detail, this exposition is relatively brief. The next section shows that the main features of the two traditional approaches, both theoretical and practical, are also covered by the Rasch rating model.

The development of this model, particularly with its emphasis on the explicit elimination of parameters in what is called the Rasch tradition of model construction, is traced. It is argued here that, without this perspective of parameter elimination, the model is most unlikely to have been constructed. This section also attempts to show that the rating model can retain characteristics which usually are seen as mutually exclusive to the two approaches because it is set in a framework apart from the other two. To help make this point, some further less obvious but no less important connections with the established approaches are described. While it has not been developed explicitly in that way, readers may see illustrative glimpses of a paradigm shift, in the sense of Kuhn (1970), in such a presentation of the Rascn tradition. This is not coincidental. I did have an eye to Kuhn's thesis when structuring this paper. A brief summary is then provided.

THE THURSTONE AND LIKERT TRADITIONS FOR STUDYING ATTITUDES

The following discussion on the relatively well-established frameworks for studying attitude is circumscribed in two ways. Firstly, it deals only with the two most common traditions, one associated with Thurstone which appeared formally in the late 1920s, and the second associated with Likert which appeared in the early 1930s. Other approaches to data collection and its modelling, together with definitions of the concept of attitude, are covered in books such as Dawes (1972) and others. Secondly, only certain key characteristics of these traditions, which are well known but which set the relationships among traditions in context, are highlighted. These restrictions in scope are directed by the concern with basic principles in both traditions rather than with their detailed elaborations which can be found, for either or both, in text books such as Edwards (1957), Torgerson (1958), Oppenheim (1966), and Bock and Jones (1968).

The work of Guttman (1954) is not considered here, partly because of space and partly because it is not used as commonly as other traditions. However, it is conjectured that, with further developments, the main features of Guttman's formulation could be covered by the rating model.

The Thurstone Tradition

Rational Scales

By analogy with studies in psychophysics, but with no reference to a physical continuum, Thurstone (1927a, 1927b) defined the concept of a discriminal process when a person reacts to a statement and formulated the law of comparative judgment. The application of this law associates with each statement a real number, called the *affective value*, which indicates the relative degree of a particular affect the statement arouses. Although no physical continuum was available, the notion of a proper linear scale with well-defined intervals was not abandoned; indeed it was stressed (Thurstone, 1928). Consequently emphasis was placed on both the evidence that the statements could be placed on a single continuum and on estimating the relative affective values of the statements. A collection of statements conforming to a linear continuum was taken to define a 'rational scale'. This term will be used throughout for a linear scale with interval or additive properties.

The Pair Comparison Design

The law of comparative judgment is based on the design of 'pair comparisons', in which persons compare statements with respect to their intensity relative to some particular attitude variable. (Thurstone developed applications of this design primarily in terms of social values and predictions of choice, but it can equally well be applied to attitude statements. Thurstone (1928) describes an alternate method for scaling statements for purposes of attitude measurement of individuals. This is discussed later in this section.) This law is generally formulated as follows (Thurstone, 1927a; Bock and Jones, 1968):

1 On encountering statement i, a randomly selected person from a population perceives it to have a real value d_i on the affective scale which, over a population of persons, may be defined by

$$d_i = \delta_i + \epsilon_i \qquad (1)$$

where δ_i is the hypothesized scale value of the statement and ϵ_i is an error component associated with the person. In the population of persons d_i is a continuous random variable which is normally distributed with expected value δ_i and variance σ_i^2.

2 In comparing statement i with statement j, the person reports statement i to have the greater value if $d_i - d_j > 0$. In the population, this difference

$$d_{ij} = d_i - d_j = (\delta_i - \delta_j) + (\epsilon_i - \epsilon_j) \qquad (2)$$

is a continuous random variable, normally distributed, with expected value

$$E[d_{ij}] = \delta_i - \delta_j \qquad (3)$$

and variance

$$V[d_{ij}] = \sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 - 2\varrho_{ij}\sigma_i\sigma_j. \qquad (4)$$

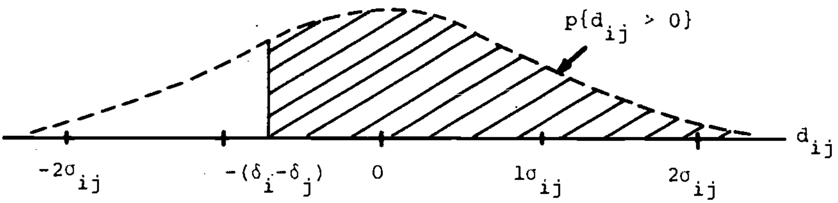


Figure 1a Probability that $d_{ij}>0$ for Fixed $(\delta_i-\delta_j)$ in a Pair Comparison Design

This difference process for a fixed $\delta_i-\delta_j$ is graphed in Figure 1a in which the shaded region represents the probability that $d_{ij}>0$. In data, the proportion of persons who judge that statement *i* has a greater effect than statement *j* is an estimate of this probability and the estimate of $\delta_i-\delta_j$ is the corresponding normal deviate. The probability that $d_{ij}>0$, as a function of $(\delta_i-\delta_j)$, may be expressed as

$$p\{d_{ij}>0\,|\,\delta_i, \delta_j, \sigma_{ij}\}=\phi\{\alpha_{ij}(\delta_i-\delta_j)\} \qquad (5)$$

where ϕ is the cumulative normal distribution with mean zero and variance unity, and where $\alpha_{ij}=1/\sigma_{ij}$ is the discrimination. This probability is graphed in Figure 1b.

The consequence of the assumption that a person is randomly selected from a population and that the inter-individual differences form part of the error is that, when estimated from a body of data, the relative scale values of items actually describe the *population* and indicate nothing specific about any individual person beyond what can be inferred from the person's membership of that population. For example, Thurstone (1927b) scaled social values with respect to criminal offences for a particular population of persons. If the scale values in another population had been different, it would have been inferred that the populations were different with respect to their opinions regarding the offences. This issue is amplified later.

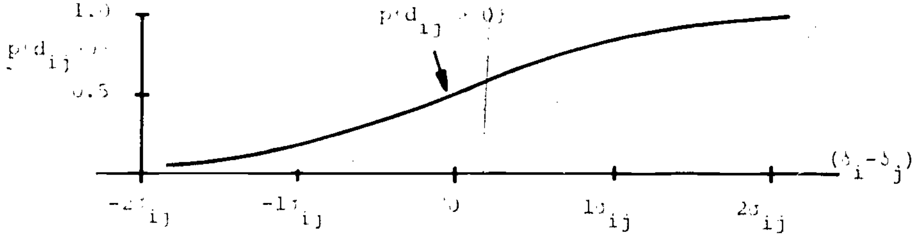


Figure 1b Probability that $d_{ij}>0$ as a Function of $(\delta_i-\delta_j)$

The pair comparison design, which has received a great deal of attention in the literature both at a technical level (David, 1963; Bock and Jones, 1968; Davidson and Farquhar†, 1976) and at a more theoretical and philosophical level (Bradley, 1976) has the drawback that it is extremely time-consuming. As a result, models for incomplete designs (Bock and Jones, 1968, Ch. 7) have been described. Adaptations and applications of the law of comparative judgment to the much simpler design of rank ordering from which dependent pair comparison can be inferred were also developed by Thurstone (1931).

The Equal Appearing Interval Design

With a scale having proper interval level properties, it was a small step to realize that not only could populations be compared with respect to the nature of the scale they generated but that, if they generated the same scale, then the populations could also be compared for location and dispersion. In this case, the scale takes on an additional characteristic, that of a *measuring instrument*.

For the explicit purpose of constructing an attitude measuring instrument in which many statements had to be scaled, the somewhat different design of 'equal appearing intervals', which has particular significance for this paper, was also developed by Thurstone (1928). In this design, people order a collection of statements, of which some 20 are finally required, into a number of groups which they consider appear equally spaced on an affective continuum.

The model for the classification, displayed in Figure 2, is a straightforward adaptation of the one shown in Figure 1b and again involves assuming that a continuous random variable d_i is induced when a person encounters a statement. Then if $\tau_1, \tau_2, \ldots \tau_k, \ldots, \tau_m$ designate the m boundaries or thresholds separating the $m+1$ ordered categories or intervals, the response corresponds to the interval in which the value of the random variable falls. If δ_i is again defined to be the affective value of statement i, then the generalization of (5) is given by

$$p\{d_i > \delta_i - \tau_k | \delta_i, \tau_k, \alpha_i\} = \phi\{\alpha_i(\delta_i - \tau_k)\}, \tag{6}$$

which is the probability that a randomly selected person will place the statement *in or below* a particular category. The estimate of each of these probabilities is the proportion of persons who classify the statement in or below a given category and, by transforming these estimates to normal deviates, the scale values of both the statements and the category boundaries can be estimated (Edwards and Thurstone, 1952).

† This paper is a bibliography of some 350 papers on the topic of pair comparisons.

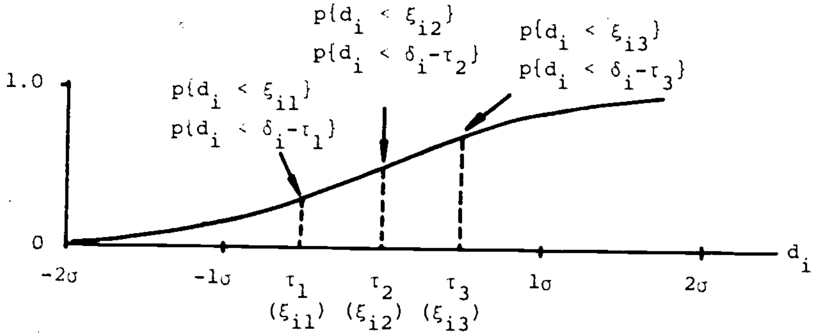


Figure 2 Probability that α_i is Less than $\delta_i - \tau_k$ for each τ_k in the Equal Appearing Interval Design or Probability that d_i is Less than ξ_{ik} in Likert-style Statement

It becomes evident from (6) that, in the equal appearing interval design, the affective value of each statement is compared with the thresholds or category boundaries. This contrasts with the pair comparison design where the affective values of two statements are compared with each other. Notice that no random variation is associated with the thresholds, but only with the statement.

With a further rearrangement and redefinition so that a person v has his ability parameter β_v compared with the item difficulty δ_i (Lumsden, 1977), equation (6) becomes

$$p\{d_{iv} > \beta_v - \delta_i | \beta_v, \delta_i, \alpha_i\} = \phi\{(\alpha_i(\beta_v - \delta_i)\} \tag{7}$$

This model has been extensively studied for dichotomous responses to achievement test items (Lord, 1952; Kolakowski and Bock, 1970).

There are three aspects of this extended Thurstone framework that are especially relevant here. Firstly, Thurstone stressed the importance of the invariance of the scale with respect to people to be measured and made a clear distinction between the construction of a scale and its use for measurement as follows:

> It will be noticed that the *construction* and the *application* of a scale for measuring attitude are two different tasks. If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. (Thurstone†, 1928; 1959, p. 228)

† Many of Thurstone's papers are reproduced in Thurstone (1959).

Thurstone described how these assumptions can be tested empirically by taking persons of known different attitude and comparing the relative scale values of statements obtained from the two groups. Secondly, he also recognized the complementary requirement that a person's measure should be independent of specific statements in the set. With respect to an achievement testing situation, he made this point as follows:

> It should be possible to omit several test questions at different levels of the scale without affecting the individual score. (Thurstone, 1926, p. 446)

Thirdly, while recognizing the importance of person-attitudes, both as a possible source of contamination in scale construction, and for the comparison of two groups with respect to location and dispersion, Thurstone never explicitly formalized a person-effect. Consequently, the procedure for 'measuring' persons with a scale, though practical and sensible, was essentially ad hoc. The procedure is to ask the person either to agree or disagree with each statement in the scale and then the measurement is taken to be either the mean or the median of the scale values of statements the person endorses. To establish invariance over statements, the statements must be equally spaced on the continuum.

The Likert Tradition

Data Collection and Scoring

Perhaps the most popular design for studying attitude is that from Likert (1932) in which persons respond directly to statements by indicating the degree of intensity with which they approve or disapprove of them. It is the same design as that of Thurstone for attitude measurement but, instead of simple endorsement or rejection, the responses have *degrees* of endorsement or rejection. To distinguish it from the pair-comparison and other data collection designs, this will be called the 'direct-response' design.

The basic response set of Likert is {Strongly Approve, Approve, Undecided, Disapprove, Strongly Disapprove}, though variations, most of which also have five categories, are readily devised; for example, another set is {Always, Often, Sometimes, Seldom, Never}. The statements are similar to those constructed by Thurstone, but they are not first scaled. Instead the ordered categories are simply scored with successive integers and a person's attitude value is taken as the sum of the scores of all statements.

This approach is popular because it is simple, it focuses directly on the attitude of persons, empirical researchers find it satisfactory, and it is theoretically undemanding. In explaining his own design, Likert wrote the following with respect to those of Thurstone:

> A number of statistical assumptions are made in the application of his attitude scales, e.g. that the scale values of statements are independent of the attitude distribution of the readers who sort the statements, assumptions which, as Thurstone points out, have not been verified. The method is, moreover, exceedingly laborious. It seems legitimate to inquire whether it actually does its work better than simpler scales which may be employed, and in the same breath to ask also whether it is not possible to construct equally reliable scales without making unnecessary statistical assumptions. (Likert, 1932, p. 7)

Thus, while arriving directly at person-attitude as if obtained by a measuring instrument, Likert rejected the necessity, spelt out by Thurstone, that the instrument's operating properties be invariant across different groups which are to be measured. He rejected, also, recourse to any formal statistical modelling of a response process.

Likert did originally investigate derivation of empirical weights for the categories, not statements, and in doing so, assumed that the distribution of responses across categories was normal. As scale values for the categories, he also used the normal deviates corresponding to the cumulative distribution. It is interesting to note that Likert observed that the distribution of responses across categories was often skewed. He took this to be, primarily, a function of the attitude distribution of the group involved. For example, with respect to a statement which had the distribution {1, 1, 3, 8, 87} in one group, he noted that it had the less skewed distribution {4, 3, 17, 18, 58} in a group with attitudes known to be different from those in the first group In this context he wrote:

> On the basis of this experimental evidence and upon the results of others, . . . it seems justifiable for experimental purposes to assume that attitudes are distributed fairly normally and to use this assumption as the basis for combining the different statements. The possible dangers inherent in this assumption are fully realised. This assumption is made simply as part of an experimental approach to attitude measurement. It is a step which it is hoped subsequent work in this field will either make unnecessary or prove justifiable. Perhaps this assumption is not correct; its correctness can best be determined by further experiment. (Likert, 1932, p. 22)

There is no attempt here to define the population in which the distribution is normally distributed but it is interesting to note firstly, that distributional assumptions among people were mentioned, and secondly, that Likert hoped these would prove unnecessary. It is also interesting that no mention was made of the effect of statement scale values on the distribution of responses among categories.

Following such scaling of categories, Likert investigated the weighting of categories by successive integers. A comparison of scores of persons obtained by a sum across items of the empirically determined weights, and those obtained by a simple sum of the integral weights, provided almost perfect correlations. On such evidence, Likert concluded that it was adequate to use the simpler weights.

The procedures for checking the consistency of statements in evoking unidimensional responses is also less formal than in the Thurstone tradition and parallels traditional test theory (Gulliksen, 1950; Lord and Novick, 1968) to such issues. Thus correlations between person-scores on odd and even statements of a questionnaire and correlations between person-scores on a statement and on the total set of statements are employed. In addition a discrimination-type index, obtained by comparing the scores of a statement on two extreme groups defined by their total scores, is often calculated. These indices are all formally much less rigorous than those for the pair-comparison and equal-appearing-interval designs which are exact probability statements regarding the quality of fit.

Although the Likert format has proved extremely satisfactory, both in terms of easy application and traditional reliability criteria, two related issues continue to be questioned: firstly, the adequacy of integer scoring, and secondly, the correctness in considering the Undecided middle category to represent an attitude between Approve and Disapprove. The first pertains to the belief that scoring by successive integers depends upon equal distances between successive categories while the second pertains more to the question of unidimensionality. Thus it is considered that a person may respond to the Undecided category for reasons such as failure to understand the statement, indifference, or ignorance, as well as some kind of neutrality (Dubois and Burns, 1975). Therefore, because an expression of attitude is considered more informative, there is an intuitive appeal in constructing statements which do not attract responses in the middle category. However, that is the very reason for concern regarding the weighting of the middle category.

Statistical Models

In advances with latent trait theory, the threshold concept with respect to ordered categories within statements, as investigated by Likert, has been formalized following the cumulative normal ogive procedure of Thurstone outlined above. Samejima (1969) developed the mathematical machinery for the case of statements and persons while Kolakowski and Bock (1972) have written a computer program to execute data analyses. However, these authors and others seemed to have concentrated on achievement items with more than one category rather than attitude items. Two further points in relation to this development need to be made.

Firstly, in achievement testing the threshold parameter is taken to be different on each item so that instead of the additive structure $\delta_i - \tau_k$ between item and threshold parameters, there is a different threshold parameter ξ_{ik} for each item. In a Likert-style attitude questionnaire, this would correspond to Likert's original approach to scaling categories rather than statements. Secondly, these authors and others shift from the cumulative normal to the logistic distribution. This is done because in maximum likelihood parameter estimation, which tends to be used, the explicit logistic is far more tractable than the implicit normal. For the pair comparison design, the logistic analogue to (5) takes the form

$$p\{d_{ij} > 0 \mid \delta_i - \delta_j, \alpha_{ij}^*\} = \frac{\exp\{\alpha_{ij}^*(\delta_i - \delta_j)\}}{1 + \exp\{\alpha_{ij}^*(\delta_i - \delta_j)\}}, \qquad (8)$$

while, for the ordered category situation, the analogue to (6) is

$$p\{d_i > \delta_i - \tau_k \mid \delta_i, \tau_k, \alpha_i^*\} = \frac{\exp\{\alpha_i^*(\delta_i - \tau_k)\}}{1 + \exp\{\alpha_i^*(\delta_i - \tau_k)\}} \qquad (9)$$

For a separate scaling of categories for each statement, $\delta_i - \tau_k$ in (9) is simply replaced by say ξ_{ik} so that different thresholds ξ_{ik} always pertain to different statements.

It is well known that, with the constant factor adjustment of $\alpha^* = 1.7\alpha$, the numerical values of the cumulative logistic and normal differ by less than 0.01 over the entire domain of the variable (Johnson and Kotz, 1972) and this numerical equivalence to the normal has given further justification for use of the logistic. Thus use of the logistic is made firstly for algebraic convenience and secondly because any differences in statistical results, apart from the unit which is often automatically adjusted, are negligible. Bradley and Terry (1952) and Luce (1959) considered the logistic model for pair comparison data in which α_i^* effectively is unity, while Birnbaum (1968) and Jensema (1974) have considered it for achievement items by analogy to Lord's (1952) consideration of the normal ogive model.

THE RASCH RATING MODEL FOR ORDERED RESPONSE CATEGORIES

The Rasch models are formalized immediately for the *direct response* design in which a response is assumed governed by *both* the affective value of a statement and the attitude value of the person. Rasch (1961) immediately specifies the condition, required by Thurstone, that the relative scale values of statements be independent of the scale values of persons. (Rasch reached the significance of this requirement, and the possibility of its realization, quite independently of Thurstone's writings. The main steps by which he did arrive at these issues are documented in Rasch (1977). The connections to Thurstone are made here for purposes of exposition of the relationships among approaches.) Rasch also makes explicit the symmetrical condition that the relative scale values of persons should be independent of scale values of the statements. Complementary conditions, which are also made explicit, are that the relative scale values of a pair of statements should be independent of the scale values of any other statements, and that the relative scale values of any pair of persons should be independent of the scale values of any other persons. Such requirements may be satisfied within some explicit *frame of reference* which includes defining the *class or set of persons*, the *class or set of statements*, and any other relevant conditions.

Specifically Objective Comparisons

In both the Thurstone and Rasch specifications, the stress is on relative rather than absolute scale values. In the Rasch system, this characteristic is often enunciated in terms of comparisons.

Rasch defines comparisons between the scale values of any two statements or any two persons, which depend only on the values of the two statements or the two persons being compared, to be specifically objective. The 'objective' term arises from the feature of independence of all other values in the system except the two being compared, while the 'specific' term is used to indicate that this objectivity is relative to some specified frame of reference (Rasch, 1977).

These specifications immediately give the scale values of statements and persons an explicit generality. That is, one can say, for example, that with respect to a class of statements and persons, and without further qualification, statement A has a greater affective value than statement B. Analogously, one can say that within the same frame of reference, and without further qualification, person C has a greater attitude than person D. Evidence for the difference in affective values between two statements is the same irrespective of person attitude; therefore the evidence from every person must point to the same difference. Again analogously, evidence for the difference in attitudes between two persons is the same irrespective of statement affective value; therefore the evidence from every statement must point to the same difference.

The requirement of objective comparisons can be used explicitly to check which statements and persons are so related and then to define classes of statements and persons which may be termed 'mutually conformable'. Further implications of this specification are considered in the next section.

The Model for Dichotomous Responses

In the context of statement and person scaling, the ordering of statements and persons on a linear continuum, as articulated by Thurstone, is immediately assumed. Accordingly, let β_v and δ_i be real numbers which characterize the scale values of person v and statement i respectively. Then the response of person v to statement i is governed by some function $\theta_{vi} = \theta[\beta_v, \delta_i]$. To begin with, consider only a dichotomous response of endorsement or rejection rather than the Likert response which has degrees of endorsement or rejection. The response, while governed by θ_{vi}, is not completely determined by it; therefore a random variable Y_{vi} which takes on the value $y_{vi} = 0$ for one response (rejection say) and $y_{vi} = 1$ for the other (endorsement) may be defined. Associated with the respective values are the probabilities

$$p\{y_{vi} = 0 \mid \beta_v, \delta_i\} = f_0(\theta_{vi})$$

and

$$p\{y_{vi} = 1 \mid \beta_v, \delta_i\} = f_1(\theta_{vi}) = 1 - f_0(\theta_{vi}). \qquad (10)$$

Both the function θ, determining the structural relationship of β_v and δ_i, and the function f giving the probability distribution, must be specified. The necessary and sufficient condition required to satisfy specific objectivity (Rasch, 1968) is that $\theta_{vi} = \exp(\beta_v - \delta_i)$, or its equivalent, and that $f_1(\theta_{vi}) = \theta_{vi}/(1 + \theta_{vi})$. Entering these functions into (10), gives

$$p\{y_{vi} = 0 \mid \beta_v, \delta_i\} = 1/\psi_{vi}$$

and

$$p\{y_{vi} = 1 \mid \beta_v, \delta_i\} = \{\exp(\beta_v - \delta_i)\}/\psi_{vi} \qquad (11)$$

where

$$\psi_{vi} = 1 + \exp(\beta_v - \delta_i)$$

in which the logistic form of the model becomes evident. This is known as Rasch's simple logistic model (SLM) for dichotomous responses, and the consequences and illustrations of why this model permits the elimination of the person parameters while estimating the contrasts among statement parameters is well documented (e.g. Rasch, 1960, 1961, 1968; Andersen, 1973a, 1973b; Wright, 1968, 1977; Fischer, 1973, 1976). In his paper, Douglas discusses these statistical issues in some detail. The key feature of the model is that the estimation of statement parameters involves firstly identifying a set of *sufficient statistics* for the person parameters and secondly conditioning on these statistics so that the person parameters are eliminated from the resulting probability expressions.

For completeness and clarity, a small example may be useful. If two statements $i = 1, 2$ whose parameters δ_1 and δ_2 are to be compared are responded to by a person v with parameter β_v, the response set is the ordered pair (y_{v1}, y_{v2}) with the sample space $\Omega = \{(0,0), (1,0), (0,1), (1,1)\}$.

With each element of the sample space the total score

$$r_v = \sum_i y_{vi}$$

is an observable statistic. The pattern of responses and this statistic are:

(y_{v1}, y_{v2})	r_v
(0, 0)	0
(1, 0)	1
(0, 1)	1
(1, 1)	2

The statistic r_v can be seen to *partition* the sample space into the sub-spaces $\Theta_1 = \{(0,0)\}$, $\Theta_2 = \{(0,1), (1,0)\}$ and $\Theta_3 = \{(1,1)\}$. Now if within a sub-space so defined—that is, conditional on the sub-space or equivalently conditional on that value of the statistic—the distribution of elements is independent of a particular parameter of the model, then that statistic is said to be *sufficient* for that parameter.† It is sufficient and no further information regarding the *value* of the parameter can be obtained by taking account of the pattern of responses. In the case of the simple logistic model for two statements, it can be shown that

$$p\{(y_{v1}, y_{v2}) \mid r_v = 1, \delta_1, \delta_2\} = \frac{\exp(-\delta_1 y_{v1} - \delta_2 y_{v2})}{\exp(-\delta_1) + \exp(-\delta_2)} \tag{12}$$

which is independent of the person parameter β_v. Therefore irrespective of the attitude values of the persons, which can be expected to be different, each response within the sub-space is a *replicate* of each other with respect to the same parameters, δ_1 and δ_2, which are to be compared. When sufficient statistics for parameters can be obtained so that one set is eliminated while another set is estimated, the parameters are often said to be *separable*.

The Model for Ordered Polychotomous Responses

The generalization of this model for more than two ordered categories, which is of interest in this paper, has been described in Andrich (1978a, 1979), Wright and Masters (1980) and Masters (1980), where the latter

† The principle of sufficiency was observed by R. A. Fisher in 1922 (Rasch, 1960). Rasch studied with Fisher in 1935 while these ideas and the associated theory of maximum likelihood were being developed. In his work, Rasch shifts the emphasis of the sufficient statistic from estimation to the elimination of parameters. Andersen (1973b) developed the theory of conditional inference.

authors give a slightly different rationale from the one presented here. Therefore the exposition below will be relatively brief.

First, suppose that every statement again has an affective value δ_i and that the categories qualify this affective value. Specifically, let $\tau_1, \tau_2, \ldots \tau_m$ be real values designating thresholds or boundary points between categories where these threshold values are on the same scale as the statement affective values, and $\tau_k > \tau_{k-1}$ for $k = 2, \ldots, m$. The thresholds are taken to *qualify* the statement and therefore effectively increase or decrease its affective value. Accordingly, suppose the thresholds and statements values are related additively and enter the model in the form $\delta_i + \tau_k$.

Second, suppose a response process of the form (11) at each threshold. Then with the addition of the threshold parameter, the response at threshold k is modelled by

$$p\{y_k = 0 \mid \beta_v, \delta_i, \tau_k\} = 1/\psi_{vik}$$

$$p\{y_k = 1 \mid \beta_v, \delta_i, \tau_k\} = [\exp\{\beta_v - (\delta_i + \tau_k)\}]/\psi_{vik} \qquad (13)$$

where

$$\psi_{vik} = 1 + \exp\{\beta_v - (\delta_i + \tau_k)\}.$$

Third, consider for the moment, and for simplicity, the case of only two thresholds and three categories. If the responses at each threshold are assumed instantaneously to be statistically and experimentally independent, the set of possible outcomes or the sample space Ω for responses to the two thresholds is the set of ordered pairs $\Omega = \{(0,0), (1,0), (1,1), (0,1)\}$, where the first member of each ordered pair indicates the response at the first threshold. The probabilities of these outcomes are given by†

$$p\{(0,0) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} = 1/\psi_1\psi_2$$

$$p\{(1,0) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} = [\exp\{\beta_v - (\delta_i + \tau_1)\}]/\psi_1\psi_2$$

$$p\{(1,1) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} = [\exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}]/\psi_1\psi_2 \qquad (14)$$

and

$$p\{(0,1) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} = [\exp\{\beta_v - (\delta_i + \tau_2)\}]/\psi_1\psi_2.$$

After considering each threshold separately, the person must bring the two processes together and, in doing so, recognize the ordering of the thresholds and the categories. Consequently the person must recognize that the pair (0,1) reflects a response above the second threshold and below the first, that is, an incompatible pair of responses. Therefore suppose the response (0,1) is not recorded if it occurs instantaneously, and that it is reconsidered and eventually distributed in one of the compatible pairs of responses. (Note that if these responses at each threshold were spaced in time so that memory, say, played no part, such an outcome

† A vector variable is set in bold type face, e.g. $\boldsymbol{\tau}$ for $(\tau_1, \tau_2, \ldots, \tau_m)$.

could occur. For example, in grading or rating a paper as excellent or not on one occasion, and fail or not fail on another with respect to a three level category set of {fail, pass, excellent}, it would be possible for the paper to be rated fail on one occasion and excellent on another.)

Fourth, suppose that each response in the sub-set $\Omega' = \{(0,0), (1,0), (1,1)\}$ of compatible responses retains the same relative probability as in the full space Ω. The appropriate probabilities are obtained simply by normalizing the probabilities with respect to Ω'. After some algebraic rearrangements, these are given by

$$\begin{aligned} p\{(0,0) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} &= 1/\gamma_{vi} \\ p\{(1,0) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} &= [\exp\{\beta_v - (\delta_i + \tau_1)\}]/\gamma_{vi} \\ p\{(1,1) \mid \beta_v, \delta_i, \boldsymbol{\tau}\} &= [\exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}]/\gamma_{vi} \end{aligned} \qquad (15)$$

where $\gamma_{vi} = \gamma(\beta_v, \delta_i, \boldsymbol{\tau}) = 1 + \exp\{\beta_v - (\delta_i + \tau_1)\} + \exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}$.

To simplify (15), first define the random variable X_{vi} to take the integral values x corresponding to the vector responses in Ω' according to

$$\begin{aligned} x &= 0 \text{ for } (0,0), \\ x &= 1 \text{ for } (1,0), \\ \text{and} \quad x &= 2 \text{ for } (1,1). \end{aligned}$$

Clearly, the value x_{vi} indicates the number of thresholds exceeded, where $x_{vi} = 0$ indicates that none has been. Second, observe that

$$\beta_v - (\delta_i + \tau_1) = -\tau_1 + (\beta_v - \delta_i)$$

and that

$$\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2) = -\tau_1 - \tau_2 + 2(\beta_v - \delta_i).$$

Third, define the τ combinations according to $\kappa_1 = -\tau_1$, $\kappa_2 = -\tau_1 - \tau_2$. Then (15) may be simplified to

$$\begin{aligned} p\{X_{vi} = 0 \mid \beta_v, \delta_i, \boldsymbol{\tau}\} &= 1/\gamma_{vi}, \\ p\{X_{vi} = x \mid \beta_v, \delta_i, \boldsymbol{\tau}\} &= \exp\{\kappa_x + x(\beta_v - \delta_i)\}/\gamma_{vi}. \end{aligned} \qquad (16)$$

Finally, generalizing to m thresholds and $m+1$ categories, and defining $\kappa_x = 0$ for $x = 0$, (16) becomes

$$p\{X_{vi} = x \mid \beta_v, \delta_i, \boldsymbol{\tau}\} = \exp\{\kappa_x + x(\beta_v - \delta_i)\}/\gamma_{vi} \qquad (17)$$

where $$\gamma_{vi} = \sum_{k=0}^{m} \exp\{\kappa_k + k(\beta_v - \delta_i)\}.$$

Equation (17) is the rating response model to which the rest of the paper is primarily devoted.

The most interesting and important feature of (17) is that the non-negative integral values of the random variable appear conveniently in the probability distribution and make it a member of the exponential family. Special cases of this distribution are the well-known binomial and Poisson distributions in which $(\exp \kappa_x) = \binom{m}{x}$ in the former and $(\exp \kappa_x) = 1/x!$ in the latter. The model of equation (12) for dichotomous responses is also clearly a special case. Masters (1980) gives a further account of the models of this family in relation to Rasch models, as does Douglas at this conference.

Next, and as a consequence, the sufficient statistic for the person parameter β_v becomes simply $r_v = \sum_i x_{vi}$, which is identical to the result in the dichotomous case. Analogously, the sufficient statistic for the statement parameter is $s_i = \sum_v x_{vi}$ and the sufficient statistic for the category coefficient parameter is $T_x = \sum_v \sum_i I_{vi}^{(x)}$ where I is an indicator variable which takes the value 1 if a response is in category x, and 0 otherwise. That is, T_x is the total number of responses, over all persons and all items, in category x. The various arguments which actually eliminate the parameters through a conditioning on these statistics will not be pursued here. (A distinction between a set of jointly sufficient statistics for a set of parameters and individually sufficient is not made here though, in developing statistical machinery, the distinction can be important (Barnard, 1974).)

In addition, the category characteristic curves have intuitively defensible and appealing properties. These can best be observed from the example shown in Figure 3 in which curves are drawn for the case of four thresholds. Firstly, the thresholds are equally spaced about an origin of zero. While this is not necessary, the thresholds can always be centred about zero (Andrich, 1978a) without any loss of generality. This makes it convenient for the interpretation of thresholds as qualifying affective values of statements, some decreasing and some increasing their affective values. Secondly, as β_v gets larger than δ_i, the probability of a high score increases. Thirdly, the response with the highest probability corresponds to the interval in which $\beta_v - \delta_i$ falls. Finally, although the assumption of independence of decisions at thresholds as a first step in deriving the model might seem 'counter-intuitive' initially (McCullagh, 1980), it is clear from equation (17) that in the final response process, the response in any category is dependent on all thresholds.

An application of this model, for the case $m > 1$, which is identical in principle to the now relatively well-known application for $m = 1$, is provided in Andrich (1978b) and therefore formal aspects of the analysis of

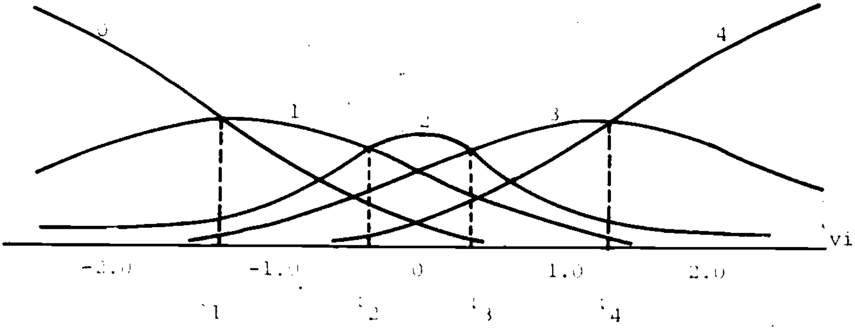


Figure 3 Theoretical Category Characteristic Curves for Four Thresholds

data will not be presented here.†

Further implications of the model are discussed in the next section in which explicit connections to the Thurstone and Likert approaches to attitude measurement are made. Before proceeding to those connections, it is significant to note that this model goes beyond the Likert-style questionnaire situation. The mode. may be entertained for any rating situation, which is very common in the social and biological sciences. These considerations have been explained for a contingency table context in which the dependent variable is a rated variable (Andrich, 1979).

THE RATING MODEL AND THE THURSTONE AND LIKERT TRADITIONS

To demonstrate the full level of unification that the rating model approach brings to the Thurstone and Likert perspectives and practice, a brief comparison of their characteristics is summarized first.

Comparisons and Contrasts between the Thurstone and Likert Traditions

A comparison of the Thurstone and Likert approaches reveals an interesting contrast, despite their virtually simultaneous development. The Thurstone approach is characterized by (i) providing statement scale values, (ii) being time consuming, requiring a judgment group to obtain

† A Fortran IV computer program which analyses data according to this model is available, at a nominal cost, from the Measurement and Statistics Laboratory, Department of Education, The University of Western Australia, Nedlands, Western Australia, 6009. An ERDC grant to the author and G. A. Douglas (principal investigators) supported the development of the program and is gratefully acknowledged.

the scale values and requiring that scale values be independent of the attitudes of the judgment group, (iii) being statistically rigorous in establishing a continuum, and (iv) somewhat ad hoc in obtaining person scale values which, asymmetrically, do depend on the distribution of statement scale values. The Likert approach is characterized by (i) not providing statement scale values, (ii) being relatively simple to apply, not requiring a judgment group and making no mention of any independence of attitude values of groups, (iii) not having a statistical model and therefore not having statistical rigour in establishing a continuum, and (iv) being direct in obtaining person values.

Spanning these Traditions with the Rating Model

Now consider the Rasch rating model in relation to these issues. First, for the purposes of person measurement, statements may be responded to in either the Thurstone tradition of rejection or endorsement, or the Likert tradition of degrees of rejection or endorsement, where the latter is seen simply as an extension of the former.

Secondly, and as a consequence of the model and not for reasons of either conceptual or numerical approximation as in the Likert tradition, the successive categories are scored with successive integers where the first category is scored by zero. Thus in the Thurstone case of a response of rejection or endorsement, the scores are 0 and 1 respectively, while in the Likert case these are extended to 0, 1, 2, 3, 4 in correspondence to the extended responses. Furthermore, although the score must be transformed to place a person's attitude on a rational scale, because the total score of a person across statements is a sufficient statistic for the attitude, the first stage of summarizing the responses is the same as in the Likert approach.

Thirdly, statement scale values which are independent of the attitudes of the person as required in the Thurstone tradition, are estimated. The advantage of scale values for statements even for Likert style formats is that they help define the continuum in a more tangible way. How this feature is exploited is shown in the next section.

Fourthly, because the data collection design is of the Likert style, it is simple and does not require the time-consuming involvement of a judgment-group.

Fifthly, the statement scale values do not have to be equally spaced on the continuum because the attitude estimate is independent of the scale values of the statements. A more or less judicious choice of statements with respect to their values can be made with a view to minimizing the error of measurement, just as in choosing items of appropriate difficulty in achievement testing (Wright and Douglas, 1977).

Sixthly, and as in the original Likert approaches with respect to the weighting of categories, thresholds between categories are estimated. However, the threshold values do not need to be equally spaced to justify integer scoring.

Finally, while employing the simple Likert-style data collection design, exact probability statements as with the Thurstone models can be made regarding fit of data to the model.

From the above list, it should be transparent that not only does the Rasch rating model synthesize and account for the apparent differences between the two traditional approaches to attitude measurement, but that in doing so, it retains the best theoretical and practical features of both. Most importantly, it does this simply and elegantly.

THE RASCH TRADITION

The above list of points deals with the most obvious relationships among the Thurstone, Likert, and Rasch approaches. An interesting issue to consider is that the rating model, which has so many characteristics consistent with theory and practice of the two traditional approaches, even where these appear in conflict (Ferguson, 1941), was not motivated with any explicit intention to reconcile the two approaches. None of the three papers which are most directly concerned with the evolution of the model (Rasch, 1968; Andersen, 1977; Andrich, 1978a) deals with real or simulated data or make reference to Thurstone or Likert. This section traces the development of the rating model through these papers, stressing the importance of the emphasis on sufficient statistics, and shows that the independent development of the model facilitated its having properties of the other approaches.

Sufficiency

Rasch's paper generalizes his SLM for dichotomous responses to the case of polychotomous responses and in the first instance, the model is of m dimensions for person and statement parameters. Briefly, if persons and statements are characterized in the first instance by vectors $\boldsymbol{\beta}_v = (\beta_{v1}, \beta_{v2}, \ldots, \beta_{vm})$ and $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{im})$ respectively, in which the maximum number of independent vectors is one less than the number of categories as in the dichotomous case†, and if the set of values of the discrete random variable X is extended from $\{0,1\}$ to $\{0,1, \ldots x, \ldots m\}$, where no meaning other than a naming of the categories for identification is

† The vector parameters are denoted by bold type face as in '$\boldsymbol{\beta}_v$' and '$\boldsymbol{\delta}_i$'. When a specific element of a vector is considered, the bold type face is replaced by an extra sub-script denoting the specific element, as in β_{vx} and δ_{ix}. A unidimensional or scalar parameter is recognized by having neither bold type face nor an element subscript as in β_v and δ_i. The term 'dimension' is not used strictly in a traditional test theory sense. It simply refers to the number of independent parameters ascribed to the persons and items.

associated with the numbers at this stage, then the generalization of (11) is given by‡

$$
\begin{aligned}
p\{X_{vi}=0\,|\,\boldsymbol{\beta}_v,\boldsymbol{\delta}_i,\,m\} &= 1/\psi_{vi} \\
p\{X_{vi}=x_{vi}\,|\,\boldsymbol{\beta}_v,\boldsymbol{\delta}_i,\,m\} &= [\exp(\beta_{vx}-\delta_{ix})]/\psi_{vi}
\end{aligned}
\qquad (18)
$$

where

$$\psi_{vi} = 1 + \sum_{k=1}^{m} \exp(\beta_{vk}-\delta_{ik}).$$

Equation (18) clearly specializes to (11) for $m=1$.

The rationale for developing this generalization for polychotomous responses is based again on the requirement that the parameters be separable and both Rasch (1968) and Andersen (1977) demonstrate the model's uniqueness in satisfying the requirement. Given that the maximum number of independent parameters for each person and each statement is m, the question arises as to the possibility of reducing the number of parameters. Rasch (1968) provides the equations in which the m person and statement parameters are reducible to any number less than m. The particular one of interest here is when that number is 1 in which case the model becomes unidimensional and the categories reflect an ordering.

When the vectors $\boldsymbol{\beta}_v$ and $\boldsymbol{\delta}_i$ can be expressed as linear functions of a single parameter β_v and δ_i respectively, for example according to

$$\beta_{vx} = \kappa'_x + \phi_x(\beta_v) \text{ and } -\delta_{ix} = \kappa''_x + \phi_x(-\delta_i),$$

(18) reduces to the form

$$p\{X_{vi}=x_{vi}\,|\,\beta_v,\,\delta_i,\,\boldsymbol{\kappa},\,\boldsymbol{\phi},\,m\} = (\exp\{\kappa_x+\phi_x(\beta_v-\delta_i)\}]/\gamma_{vi} \qquad (19)$$

where

$$\kappa_x = \kappa'_x + \kappa''_x$$

and where

$$\gamma_{vi} = \gamma(\beta_v,\,\delta_i,\,\boldsymbol{\kappa},\,\boldsymbol{\phi}) = \sum_{k=0}^{m} \exp\{\kappa_x+\phi_x(\beta_v-\delta_i)\}.$$

As in (17), the $\boldsymbol{\kappa}$ are the category coefficients while the $\boldsymbol{\phi}$ are the scoring functions, where $\kappa_0=\phi_0=0$. The relationships between the $\boldsymbol{\kappa}$'s and $\boldsymbol{\phi}$'s in (17) and (19) are explained shortly.

For analysing data according to (19), the techniques developed involve first estimating the m-dimensional statement parameters and then factoring these parameters according to $\delta_{ix} = -\phi_x\delta_i$ (Andersen, 1973a; Spada and Fischer, 1973; Allerup and Sorber 1977). If a likelihood ratio test shows that the m-dimensional and factored form are not significantly different, then the hypothesis of unidimensionality is accepted. Once the δ_i and ϕ_x have been estimated, the κ_x and β_v can be estimated unconditionally by a generalization from the dichotomous case (Wright and Panchapakesan, 1969; Andersen and Madsen, 1977).

‡ There are a number of different ways of expressing (18), but this one will be most convenient here (Rasch, 1968; Andersen, 1973a, 1977).

There are, however, two interesting and related characteristics of (19) which require comment. First, while the multidimensional parameters $\boldsymbol{\delta}_i$ can be estimated independently of the person parameters $\boldsymbol{\beta}_v$, the parameters ϕ_x and δ_i cannot be estimated independently of each other. Secondly, even with known ϕ_x and δ_i, in general the estimate of β_v for person v permits a complete recovery of his pattern of responses across the categories. Therefore, in a well-defined sense, there is no data reduction in the process of estimating a person's parameter. This implies some kind of pseudo-estimate of a unidimensional parameter which in reality is still multidimensional.

In relation to making the latter observation, Andersen (1977) investigated the model in (19) and established that if there exists a sufficient statistic for a unidimensional parameter β_v which is a function of data only, then the differences $\phi_{x+1} - \phi_x$ for all $x < m$ must be equal. Following on from Andersen's work, I provided (Andrich, 1978a; 1979) an interpretation of the category coefficients and the scoring functions essentially in the form presented in a previous section, except that the response process at each threshold $x = 1, \ldots, m$, was initially parameterised to have possibly a different discrimination α_x. The Birnbaum (1968) response model at each threshold, rather than the Rasch SLM of equation (11), in the form

$$p\{y_k = 1 \mid \beta_v, \delta_i, \boldsymbol{\tau}, \alpha_k\} = [\exp[\alpha_k\{\beta_v - (\delta_i + \tau_k)\}]]/\psi_{vik} \qquad (20)$$

where $\psi_{vik} = 1 + \exp[\alpha_k\{\beta_v - (\delta_i + \tau_k)\}]$ was postulated. Applying the rationale presented earlier to this model gives equation (19) in which $\kappa_0 = \phi_0 = 0$, $\kappa_x = -\sum_{k=1}^{x} \alpha_k \tau_k$, and $\phi_x = \sum_{k=1}^{x} \alpha_k$. $x = 1, 2, \ldots, m$. Then if the discriminations α_x are the same at each threshold, this value can be absorbed into the other parameters with the result that $\kappa_x = -\sum_{k=1}^{x} \tau_k$ and $\phi_x = x$, giving (17).

Given that the model (19) is generated by a model different from the SLM at each threshold, it is not surprising that it does not subscribe fully to the requirement of having a sufficient statistic for β_v. However, it is stressed that the derivation of (17) was not made through a specialization of the Birnbaum model (20). Instead (17), or its algebraic equivalent, was first developed by Andersen without interpretation of the parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\phi}$ given here. In addition, I derived model (17) before realizing its connection with models (19) and (20) in terms of discriminations at thresholds. In the presentation of the model (17) (Andrich, 1978a), I derived model (19) first and then specialized it to (17), but this was for purposes of efficient exposition.

In relation to the development of (19), it is stressed that the search for a sufficient statistic in the Rasch approach is directed primarily by the perspective of *eliminating* parameters. While this has particular implications for estimating parameters, it contrasts with the usual approach to ordered categories, exemplified by Samejima (1969), in which the emphasis is on *estimation*. With only this latter emphasis, the development leading to the rating model may have stopped with algorithms for estimating the scoring functions ϕ_x (Andersen, 1973a; Spada and Fischer, 1973). (The substantive interpretations of the scoring functions $\boldsymbol{\phi}$ in Andersen, and Spada and Fischer are quite consistent with the idea of a discrimination at each threshold. However, no similar prior interpretation of the category coefficients $\boldsymbol{\kappa}$ seems to have been made.)

A consequence of this approach, in relation to that of Samejima's in which the cumulative probability generalization is used as in (10), is interesting to consider. Although the emphasis in the latter approach is on estimation, no simple explicit expression for the response in each category follows and the probability of response in each category is the difference of adjacent cumulative probabilities. Therefore no simple sufficient statistic for estimation of parameters follows. The surprising consequence is that the apparently more straightforward generalization of the dichotomous to the ordered category model, which was used more or less formally by both Thurstone and Likert, does not provide the comprehensive synthesis of those approaches as does the independently derived Rasch generalization of the dichotomous model. This feature is not simply a result of the algebraic formulation because the Rasch rating model and the Samejima model cannot be transformed into each other.

The other important contrast of approaches is that they generate problems which are unique to each approach. In the Rasch approach, estimation ideally is carried out through conditional distributions (Andersen, 1973a, 1973b) and even though the models are simple and estimates have desirable properties, the implementation of algorithms for solving the resultant equations can become complex.

In attempting to find approximations to the estimation which may be more efficient, the consistency of the estimates is always a concern because of the demonstration by Andersen (1973a), that unconditional estimates—that is, joint estimates of the statement and person parameters—in the dichotomous case are not consistent. (For a fixed number of statements, as the number of persons increase without limit, the parameter estimates converge to values which are not the actual parameter values.) In the Samejima approach, while the expression for the probability of each category is more complex, the estimation is carried out by the more straightforward unconditional approach. In such approaches actual convergence of estimates in iterative algorithms for implicit equations may sometimes be at issue, but the actual consistency of estimates in the sense of Rasch is not as explicit a concern. Thus the numerical methods problems are distinctly different.

Parameters Elimination and the Terms 'Population-free' and 'Sample-free'

The focus of Rasch on parameter elimination together with some consequent issues have just been described. The possibility of eliminating parameters explicitiy is considered generally a desirable property for psychometric models and the terms 'population-free', 'sample-free', 'person-free', and 'item-free' have been coined with respect to Rasch models.

However, these terms are not always fully appreciated; sometimes they are taken to imply more and sometimes less than what they actually mean. The confusions often stem from the dual uses of both the terms 'population' and 'sample', and the latter especially in relation to the idea of sampling distributions.

One use of population is associated with the specification of a *class* of objects or people as, say, in the population of 15-year-olds in Australian high schools. The other use is with respect to numbers associated with each of the members of the class with respect to some variable. For example, it might be said that the numbers indicating the degree of achievement on some test are normally distributed. In relation to the latter use, random sampling has the virtue that distributional properties of random samples are well specified, hence the common use of the random sample to represent some population. The confusion, of course, readily arises because to get the random sample of numbers, one selects the members of the class at random. However, conceptually, these are different and it is with respect to the numbers associated with people and their distribution and not with respect to the specification of the class of people, that the Rasch models are population-free. They are free of distributional populations, not of classes of people. Because any member of a population as a class can be selected, there is no need to invoke random sampling and its consequent distributional properties to check on the structure of the scale and the quality of measurement. In this sense the models are also sample-free.

In general, both a class of statements and a class of persons are envisaged when some attempt at measurement is made. To check the conformity of the statements and person parameters, various tests of fit can be applied. Both person-fit and statement-fit (Wright and Stone, 1979; Wright and Mead, 1977; Mead, 1976) can be examined in order to isolate and understand why members of the person and statement classes, considered on a priori grounds to belong to the same class, do not conform with respect to the measurement procedure. Within a conformable class, any members can be selected, and in that sense the models are person-free and statement-free, but the classes have to be defined and confirmed to be conformable.

In addition, by examining different classes of statements and persons, for example from two scales devised separately but on the same issue, or two classes of persons such as all 14-year-olds and all 15-year-olds in a school system, and checking if they conform with each other, the generality of the variables and scales can be extended.

If after the generality of a scale is demonstrated across different sub-classes of people, one wished to compare two sub-classes such as all 14-year-olds and all 15-year-olds with respect to location and dispersion on the trait, then a random sample from each sub-class ought to be selected. But then the aim is to *describe* a distribution of a population of numbers with respect to some class, not to confirm structural properties of the class. This point also demonstrates that it is only the relative statement scale values that are objective, and not the absolute values.

In this connection, it might be stressed that the 'distribution-free' properties of the Rasch models, to use another more general term, is a property of the models, not of data. That is, to demonstrate distribution-free properties of models, one only needs to consider the models. Only if real data conform to the models can the corresponding characteristics be applied to the data. A check if data accords to the model can involve checking that the statement or person scale values are distribution-free. When the relationships among statements are not distribution-free or attitude-free with respect to two classes of people, it may be just as informative as when they are, because then a potentially significant difference between the classes of people has been exposed. In this sense, the information is analogous to that of Thurstone in the pair-comparison approach in which populations are described, not by their respective means and variances, but by the scales they generate.

Another Connection to the Thurstone Tradition

It was observed earlier that Thurstone realized the significance of invariance of relative statement-values across persons with different attitude values. However, he never formalized this feature in his models. Therefore it is opportune to note here, and as shown in detail elsewhere (Andrich, 1978c), that by formalizing Thurstone's own verbal statements regarding the discriminal process of equation (1), and by rearranging the error term to separate the among-person variance from the within-person variance, the law of comparative judgment applied to the pair-comparison design does eliminate the person parameters. Thus distribution-free statement-scaling is met by the pair-comparison design.

This effect is manifested by the correlation ϱ_{ij} in (4) being zero, and although in his data analyses Thurstone consistently assumed $\varrho_{ij}=0$ (c.f. Thurstone, 1927b, 1927c) and although in an empirical study (Thurstone, 1931) he calculated that the observed correlation was indeed virtually zero, it does not appear that he ever related this assumption and evidence to the elimination of person scale-values. Thurstone's specifications must be modified to realize it, but his requirement of random sampling of persons is in fact not necessary. Interestingly, it seems that it is difficult to motivate this particular modification unless one has the perspective of explicit person parameter elimination which is so central to the Rasch tradition. That is, to appreciate this characteristic in Thurstone's model, one must look at it from a Rasch perspective.

Not only are the person parameters eliminated in the pair-comparison design but, if the logistic distribution is substituted for the normal and the discriminations α_i are assumed to be the same for all statements, which is Thurstone's Case *V* specialization of (4), then the statement scale values for the pair-comparison design and the direct-response design (whether dichotomous or polychotomous as in the Likert format) will be the same according to the models. Analogously, when statements are categorized as in the equal-appearing-intervals design, then each statement can be considered to have been rated. Accordingly the rating model again can be applied. Thus the scale values obtained by the equal-appearing-intervals design, the pair-comparison design, and the direct-response design, all provide, according to the models, the same statement scale values and all are free of any attitude of the persons involved in the data collection. Whether or not sets of data show these properties is an empirical question, but to the degree that they do, then to that degree generality of relationships is demonstrated.

Another Connection to the Likert Tradition

The estimation of parameters with sufficient statistics complements the elimination of such parameters. The issue of estimating the attitude β_v for person v on a rational scale by transforming the total score $r_v = \sum_i x_{vi}$ is taken up again here.

If the statement and threshold values are assumed accurately estimated, then the direct maximum likelihood equation

$$r_v = \sum_i x \exp\{\kappa_x + x(\beta_v - \delta_i)\}/\gamma_{vi} \qquad (21)$$

can be used to estimate β_v (Andersen and Madsen, 1977; Andrich, 1978b) and the associated error variance of $\hat{\beta}_v$ is approximated by

$$\hat{\sigma}^2_{\hat{\beta}_v} = 1/\sum_i \{(\sum_x x^2 p_{vxi}) - (\sum_x x p_{vxi})^2\} \qquad (22)$$

where p_{vxi} is given by (17).

Table 1 Transformation of Total Scores to Attitude Estimates for a Conformable Set of 16 Likert-style Statements without the Undecided Category

Total score	Cumulative proportion	Attitude estimate	Standard error
r		$\hat{\beta}$	$\hat{\sigma}_{\beta}$
1	0.00	−3.79	1.00
4	0.00	−2.36	0.51
7	0.00	−1.75	0.40
10	0.00	−1.34	0.35
13	0.01	−1.00	0.32
16	0.03	−0.71	0.30
19	0.07	−0.45	0.29
22	0.17	−0.19	0.29
25	0.34	0.06	0.29
28	0.56	0.31	0.30
31	0.75	0.58	0.31
34	0.87	0.87	0.32
37	0.96	1.21	0.35
40	0.98	1.62	0.39
43	0.99	2.17	0.48
47	1.00	3.89	1.01

Clearly all persons with the same total score will have the same attitude estimate. In addition, the transformation of total scores r_v to attitude estimates $\hat{\beta}_v$ is monotonic. Therefore, if there is reasonable variation among total scores of a set of persons, the total scores will correlate closely with the attitude estimates. A transformation of total scores to attitude estimates for the example of 16 Likert-type statements with four ordered categories described in Andrich (1978b) is shown in Table 1. The same relationship is portrayed graphically in Figure 4. It is apparent from this figure that there is a wide range, approximately 5 to 40, in which the scores r_v and $\hat{\beta}_v$ are virtually linearly related. In the real data approximately 95 per cent of the 284 persons were in the range from 17 to 40. This type of relationship perhaps helps explain the success of Likert's integer scoring and, in an interesting sense, renders unnecessary the concern of people who assumed that integer scoring depended on equal distances between thresholds bordering the categories.

However, an interesting question which might be asked is, Why bother with the transformations if the total scores will suffice? Related questions are, what effect, if any, do the affective values have if the total score is all

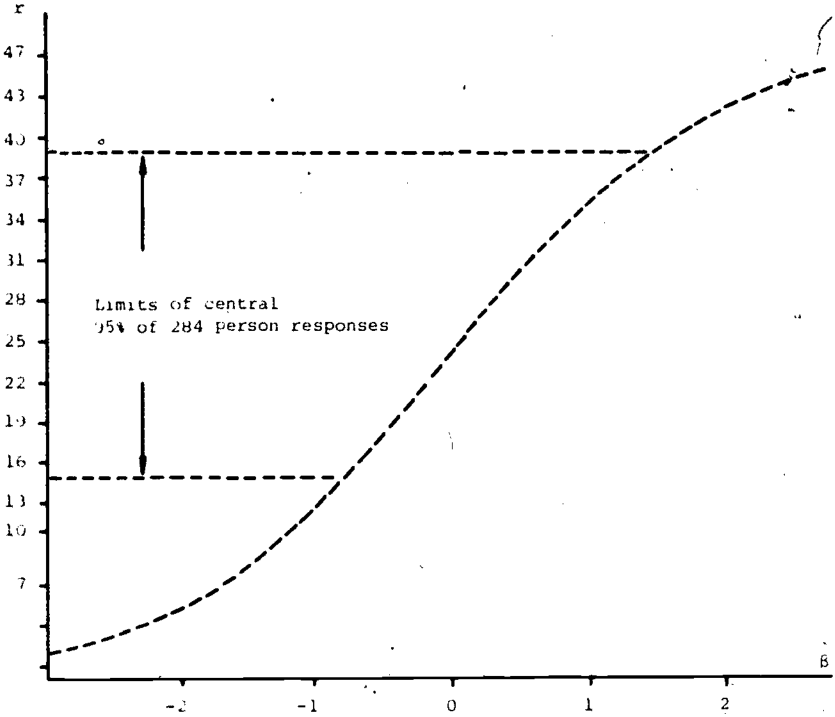


Figure 4 Relationship of Total Scores and Attitude Estimates for a Conformable Set of 16 Likert-style Statements without the Undecided Category

that is needed, and does it not make any difference which statements are endorsed or rejected.

Answers to all of these questions help explain further the aspects of the rating model and how more can be gained by using it than can be obtained by simply using the total score. First, while the monotonic relationship between r_v and β_v is a straightforward algebraic relationship, the meaning of the scores and estimates is only valid if the responses accord to the model. As already mentioned, explicit checks of person-fit and item-fit are available to help define the class of statements and persons which are actually conformable and which indicate which statements and persons need special consideration. In the case of statements, the special consideration may be related to its wording or the like. Presumably, when these statements are constructed and placed in the questionnaire, it is expected that they conform with the other statements. To the degree

Table 2 Response Patterns of Three Persons with the Same Score to Ten Statements

Increasing affective values

→

Persons (v)	Statements (i) 1	2	3	4	5	6	7	8	9	10	r_v
1	3	3	2	2	1	2	0	1	0	0	14
2	3	2	3	2	1	2	1	0	0	0	14
3	0	3	0	1	2	3	0	2	0	3	14

that they do not, to that degree the theory or principle on which the statements are generated is not mastered. Understanding the source of statement-misfit can therefore further help clarify the attitude variable.

The case of misfitting persons indicates that they were not measured as intended. That is, their total score is not 'sufficient' to account for a single attitude and they are not comparable with other persons on the same scale. Such persons too can contribute to the refining of the variable.

Requiring conformity to the model means that only certain patterns with a specified amount of random variation are permissible. If two people have the same total score, they will also have a similar pattern of responses. The total score is sufficient for the parameter estimate and no further information for the parameter can be gleaned from the response pattern, but the pattern can be used for checking the fit. For example, consider a four-category response case of ten statements with increasing affective values from left to right. The response patterns of three people each with the score of 14 are shown in Table 2.

Persons 1 and 2 could readily be conformable and it would be easy to believe that the slight difference in pattern is due to random fluctuation. However, for a total score of 14, the response pattern of person 3 would be considered odd. This person endorses statements of greater or lesser affective values about equally. Such a response pattern would be diagnosed as misfitting the model. Thus the weighting of statements in terms of their scale values plays an important role in recognizing unusual patterns and in confirming that a total score can be used in equation (21) to estimate β_v.†

Another situation where the effects of statement scale values can be seen is when all persons do not respond to the same statements. Since the

† For a contrasting view on this issue, based on a non-Rasch perspective on sufficiency, see Samejima (1969), Chapter 10, entitled, 'Some observations concerning the relationship between formulas for the item characteristic function and the philosophy of scoring'.

person scale values should be independent of the statement values, any sub-set of statements from a conformable set should give statistically equivalent person measures. This is particularly useful in constructing 'parallel forms' of questionnaires when repeated measurements are required and the same statements are avoided to reduce the effects of memory of specific responses to the same questions. But, in this case, the same total score from the two forms will not, in general, have the same attitude estimates. For example, if the first form happened to have statements with somewhat higher affective values than the second form, then a total score on the first form would correspond to a greater attitude value than the same score on the second form.

The Rating Model and the Undecided Category

It was indicated earlier that two concerns with the integer scoring of successive categories still persist. These are the equidistance of intervals or distances between categories in general, and the operating characteristics of the middle or Undecided category in particular. A justification of integer scoring through the rating model, without reference to distances between categories, has already been made. The operating characteristics of the middle category, and its manifestations in the rating model, are now examined.

First consider what might be expected. If the middle category operates consistently with the other categories, then the probability of response in the category for any statement should show an appropriate transition across categories as a function of β_v; in particular, the middle category should neither be over-represented nor under-represented.

The category characteristic curves in Figure 3 reflect such conditions. Explicit probabilities for a 5-category response format from say Strongly Approve (SA) to Strongly Disapprove (SD), which conform to this pattern, are portrayed in Figure 5 for three values of $\lambda_{vi} = \beta_v - \delta_i$ and for threshold values of $\tau_1 = -1.20$, $\tau_2 = -0.40$, $\tau_3 = 0.40$, $\tau_4 = 1.20$.

Because the response depends on $\beta_v - \delta_i$, the three graphs from left to right could represent the response patterns of a single person to three statements of decreasing affective values or of three persons of increasing attitude to a single statement. To illustrate what tends to happen when data including the Undecided category are analysed according to the rating model, some results from a real data set are now described briefly.

These data, which involve the responses of 309 Year 5 school children in Australia who answered 16 statements called 'questions about school' (Western Australia, Education Department, 1974), have been analysed according to the case of the rating model (Andrich, 1978d) having binomial coefficients.

An obvious feature of the threshold estimates, shown in Table 3, is

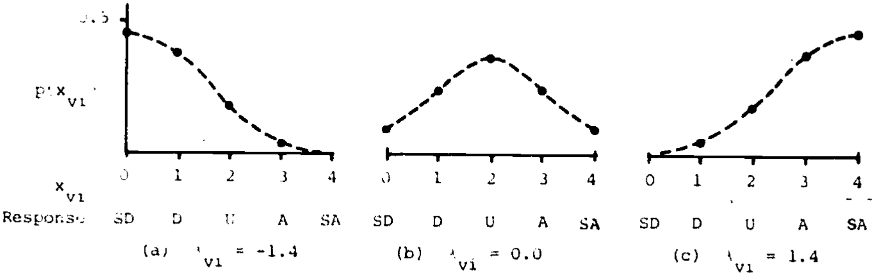


Figure 5 Model Response Probabilities in Ordered Threshold Case for Three Persons with Increasing Attitude to a Single Statement or of a Single Person to Three Statements with Decreasing Affective Values $\lambda_{vi} = \beta_v - \delta_i$; $\tau_1 = -1.20$; $\tau_2 = -0.40$, $\tau_3 = 0.40$, $\tau_4 = 1.20$.

that they are not ordered as expected, in particular $\hat{\tau}_3 < \hat{\tau}_2$. With these values of the thresholds, distributions analogous to those in Figure 5, and with similar values of $\lambda_{vi} = \beta_v - \delta_i$, are displayed in Figure 6. It is evident that the distribution for a central value of λ_{vi} is bimodal. A general principle can be inferred from this illustration, namely, that only if the thresholds are ordered is the rating model distribution strictly unimodal. Ordered thresholds ensure that the coefficients $\varkappa_x$ have the relationship

$$\varkappa_x > \frac{\varkappa_{x-1} + \varkappa_{x+1}}{2} \text{ for all } x = 1, \ldots, m-1$$

which in turn reflects the unimodality. The Poisson and binomial distributions, which are special cases, have this relationship among coefficients.

A bimodal probability distribution, which for any fixed set of parameters $\lambda_{vi} = \beta_v - \delta_i$ is a random error distribution, seems untenable for a unidimensional variable. In general, if an observed distribution is bimodal, it reflects at least two overlapping populations of numbers.

A manifestation of the reversed thresholds can also be seen from the category characteristic curves shown in Figure 7, from which it is evident that no matter what the value of $\lambda_{vi} = \beta_v - \delta_i$, the probability of a response

Table 3 Threshold Estimates from a Real Data Set of 16 Statements Including the Undecided Category

k	1	2	3	4
$\hat{\tau}_k$	−0.40	0.00	−0.38	0.79
$SE(\tau_k)$	0.05	0.04	0.04	0.04

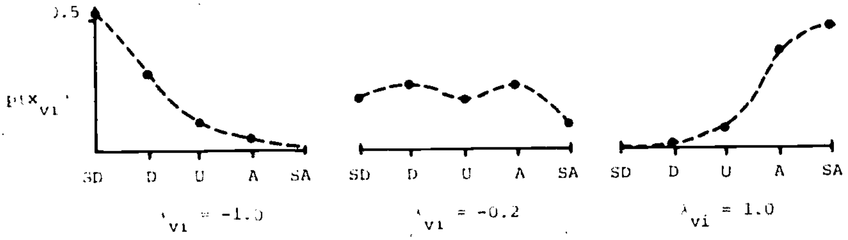


Figure 6 Model Probabilities as a Function of $\lambda_{vi}=\beta_v-\delta_i$ for a Real Data Set in which Threshold Estimates Show Disordering

in the middle category is never greater than the probability of a response in at least one of the other categories. Indeed, if $\lambda_{vi}=0.0$, where effectively the person and statement values cancel each other, then one would expect that the most likely response would be in the middle or Undecided category. However, it is not. The response probability in the Agree category is greater than the probability for the middle category. Note that this has nothing to do with the distribution of people. The distribution of people might be bimodal so that some have a high attitude and some have a low attitude and very few have a middle attitude. But this will not affect the category characteristic curves. The distribution in Figures 6 and 7 pertain to a single individual.

A reversal of threshold estimates can occur if the discriminations at thresholds are not equal but the data are analysed as if they were, that is, *if an incorrect model is applied.* The general perspective taken here is that if threshold estimates are disordered, then the data do not fit the model. It should be noted that the fit or otherwise on this criterion does not invoke a fit statistic with an associated probability. In any case, such a statistic provides only a necessary, and not a sufficient condition, for

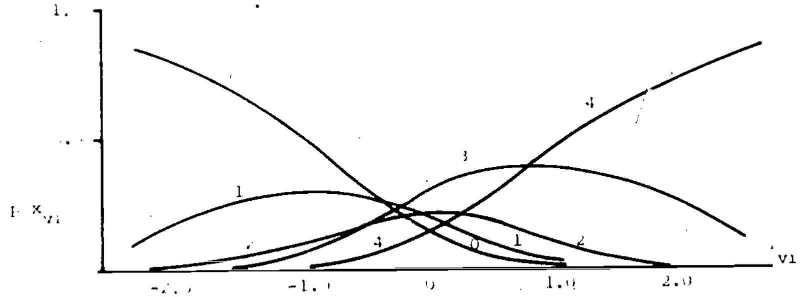


Figure 7 Category Characteristic Curves for a Real Data Set in which Threshold Estimates Show Disordering

Table 4 Threshold Estimates from a Real Data Set of 16 Statements without the Undecided Category

k	1	2	3
$\hat{\tau}_k$	-1.00	-0.02	1.02
$SE(\tau_k)$	0.05	0.04	0.04

deciding the fit. The criterion of threshold order arises directly from the specification of the model.

The question is what to do about this. An answer is available from psychophysics literature in which a similar problem has been encountered. This is to leave the Undecided category out. In the case when the other choices are simply Approve or Disapprove, this category is considered best left out (Bock and Jones, 1968, p. 3). It seems consistent to leave it out also when the intensity of Approval or Disapproval is extended. As a practical recommendation, in the piloting of questions the Undecided category would be included as usual and then any statements which seemed to attract too many responses in this category would be modified or excluded. In the final version of the questionnaire, the Undecided category would be left out and an instruction given to people to make a response even if, sometimes, they were uncertain.

The analysis of a questionnaire constructed under such principles, some results of which are shown in Table 1 and Figure 4, has been reported in Andrich (1978b). For a conformable sub-set of 16 statements from an original set of 20, the resultant threshold estimates are shown in Table 4. As can be seen, the thresholds are in the correct order and the distances are symmetrical.

It must be stressed that the emphasis on the above results is not on the actual values but on their symmetry, and that this symmetry confirms the suitability of the rating model for analysing such data. The threshold estimates were obtained by an unconditional estimation procedure and, therefore, the quality of the estimates in relation to consistency is not known.

Another point that needs mentioning is that, interestingly enough, the objectivity requirements of the model are not destroyed if the natural threshold order is violated. Primarily on the basis of this fact, Wright and Masters (1980) and Masters (1980) are prepared to accept disordered thresholds as not violating the model. Masters also gives an extensive review of category characteristic curves obtained in psychophysics research and discusses the notion of 'response set' with respect to the Undecided category.

Finally on this point, it is noted that the traditional cumulative probability approach would not reveal so vividly the anomaly exposed above. Because the probabilities are accumulated, and the logistic transform applied effectively at each accumulated point, the thresholds' estimates must be strictly ordered. In the Rasch model the logistic transform is applied effectively with respect to each pair of adjacent categories, hence the possible disclosure of disordered thresholds. The corresponding result from the cumulative probabilities approach would be a smaller distance between the two middle thresholds than between the other thresholds. Although the notion of *threshold* is similar in the two approaches in that in both it refers to a cut-off point on a continuum, the thresholds are actually formally defined differently in the two approaches so that different values are obtained from the two models.

Further Aspects of the Rasch Tradition

The elegant features of the rating model, a member of the Rasch models, does not obviate the need for patient, careful, insightful, and sometimes laborious construction of statements. Indeed, because of its explicitly demanding requirements, the care required may sometimes be greater than in traditional approaches. The reward, however, is the generality of the statements constructed with respect to the variables conceptualized.

The rating model itself, as indicated earlier, is relevant beyond the Likert-style questionnaire context; in a sense, that case is only an example of a rating system for classifying ordered data. In the social and biological sciences, a rating system is used usually when formal measurements cannot be made. Application of the rating model to such data can, therefore, provide a check on the quality of rating mechanism and help place the results of ratings on the same level as that of usual measurement. The only differerce between rating and measurement then becomes the degree of accuracy, which in any case can also be estimated from the model. In the physical sciences, the very process of measurement is used to clarify and understand variables. Modifying and improving a rating procedure with respect to some variable according to the rating model should help clarify variables, which cannot be ordinarily measured, in the same way.†

In this context, it is also worth noting that, in the physical sciences, the different variables involved in lawful relationships and the lawful relationships themselves are defined simultaneously (Kuhn, 1972) and that these very definitions often involve measurements (Ramsay, 1975).

† The prevalence of the rating scale in social science research is testified to by Dawes (1972) who states that some 60 per cent of studies involve as dependent variables only rating type variables.

In contrast, there is a tendency in psychology and education, and particularly the latter, to construct various scales for variables in a more or less independent way, and then following their construction, to examine relationships among variables using some form of correlational procedure. Thus measurement is seen to be *prior* to establishing relationships among variables. However, the demands of Rasch's specific objectivity can be seen as demands for general lawful relationships (Rasch, 1977) in which the three aspects—(i) the definition of each variable involved, (ii) the relationships among variables, and (iii) their measurement—are simultaneous.

On this issue which touches some fundamental epistemological questions, but which has only been briefly mentioned in order to indicate further possible developments, three final points may be made. First, the usual study of constructing measured variables as prior to investigating their relationships can still be applied using scales conforming to the Rasch specificatior, and be better because the scales do so conform. Secondly, the emphasis on the specification of lawful relationships, with measurement being simultaneous to it, is exemplified by papers in Spada and Kempf (1977) and Kempf and Repp (1977) and seems to be the direction taken in the German-speaking countries. Finally, conceptualizing rational scale construction or measurement as part of establishing lawful relationships again is consistent with Thurstone's conceptualization for attitude scaling which is derived from the psychophysical framework.

SUMMARY

A Rasch model for ordered response categories is derived and it is shown that it retains the key features, both theoretical and practical, of both the Thurstone and Likert approaches to studying attitude. These key features of the latter approaches are also reviewed.

Characteristics in common with the Thurstone approach are: statements are scaled with respect to their affective values; these values are independent of attitudes of the persons responding; the scales are rational in the sense that they have interval level properties; the scale values, apart from a linear transformation, are the same as in the pair-comparison design and the equal-appearing-interval design; the model for the data, being an explicit probability model, provides formal tests of fit. Characteristics in common with the Likert approach are: no judgment group is required; persons whose attitude is to be quantified respond to statements in the usual way in terms of intensity of Approval or Disapproval; while thresholds or boundaries between categories are estimated, the successive categories are scored with successive integers; though it has to be transformed monotonically, the attitude of a person is characterized by the simple sum of the integral scores across the set of

statements; concerns with Likert's Undecided category are appropriately manifested in the rating model.

Further features of this model which distinguish the Rasch tradition are considered. Thus it is shown that the more conventional latent trait formalizations which were used or broached by Thurstone and Likert for ordered qualitative data do not provide the synthesis of the two approaches that the rating model does, even though the latter model is derived from a very different basis. It is shown that this basis, which generates a completely different set of research questions for estimation from that of the conventional approach, is characterized by identifying sufficient statistics which can be used for eliminating one set of parameters while estimating the others. In this context, further connections to both the Thurstone and Likert traditions are made. Finally, the possibility of orientating the Rasch tradition for studying relationships among variables to one in which measurement is seen as simultaneous to constructing lawful relationships among variables, rather than prior to examining such relationships, is broached briefly.

REFERENCES

Allerup, P. and Sorber, G. *The Rasch model for questionnaires*. Copenhagen: Danish Institute for Educational Research, 1977.

Andersen, E. B. Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973, **26**, 31–44.(a)

Andersen, E. B. *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forskningsinstitut, 1973.(b)

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika*, 1977, **42**, 69–81.

Andersen, E. B. and Madsen, M. Estimating the parameters of the latent population distribution. *Psychometrika*, 1977, **42**, 357–74.

Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 1978, **43**, 561–73.(a)

Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 1978, **2**, 581–94.(b)

Andrich, D. Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 1978, **3**, 449–60.(c)

Andrich, D. A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1978, **31**, 84–98.(d)

Andrich, D. A model for contingency tables having an ordered response classification. *Biometrics*, 1979, **35**, 403–15.

Barnard, G. A. Conditionality, pivotals and robust estimation. In O. Barndorff-Neilson (Ed.). *Proceedings of the Conference on Foundational Questions in Statistics*. Aarhus, Denmark: Department of Theoretical Statistics, University of Aarhus, 1974, 61–80.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Bock, R. D. and Jones, L. V. *The measurement and prediction of judgement and choice.* San Francisco: Holden Day, 1968.

Bradley, R. A. Science, statistics, and paired comparisons. *Biometrics*, 1976, **32**, 213–32.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, 1952, **39**, 324–45.

David, H. A. *The method of paired comparisons.* New York: Hafner, 1963.

Davidson, R. R. and Farquhar, P. H. A bibliography on the method of paired comparisons. *Biometrics*, 1976, **32**, 241–52.

Dawes, R. M. *Fundamentals of attitude measurement.* New York: Wiley, 1972.

Dubois, B. and Burns, J. A. An analysis of meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, 1975, **35**, 869–84.

Edwards, A. L. *Techniques of attitude scale construction.* New York: Appleton Century Crofts, 1957.

Edwards, A. L. and Thurstone, L. L. An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 1952, **17**, 169–80.

Ferguson, L. W. A study of the Likert technique of attitude scale construction. *Journal of Social Psychology*, 1941, **13**, 51–7.

Fischer, G. H. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 1973, **37**, 359–74.

Fischer, G. H. Some probabilistic models for measuring change. In D. N. M. De Gruijter, L. J. Van der Kamp (Eds), *Advances in psychological and educational measurement.* London: Wiley, 1976.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Guttman, L. The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Ill.: The Free Press, 1954.

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

Johnson, N. L. and Kotz, S. *Distributions in statistics: Continuous univariate distributions* (Vol. III). New York: Wiley, 1972.

Kempf, W. F. and Repp, B. H. (Eds), *Mathematical models for social psychology.* Vienna: Hans Huber, 1977.

Kolakowski, D. and Bock, R. D. *A FORTRAN IV program for maximum likelihood item analysis and test scoring: Normal ogive model.* (Research Memorandum No. 12). Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1970.

Kolakowski, D. and Bock, R. D. *A FORTRAN IV program for maximum likelihood item analysis and test scoring: Logistic model for multiple item responses* (Research Memorandum, No. 13). Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1972.

Kuhn, T. S. *The structure of scientific revolutions.* Chicago: University of Chicago Press, 1970.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, No. 140.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, **7**, 1952.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Luce, R. D. *Individual choice behaviour.* New York: Wiley, 1959.

Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, **1**, 477–82.

Masters, G. N. A Rasch model for rating scales. Unpublished Ph.D. dissertation, University of Chicago, 1980.

McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society*, Series B, 1980, **42**(2), 109–42.

Mead, R. J. The assessment of fit of data to the Rasch model through analysis of residuals. Unpublished doctoral dissertation, University of Chicago, 1976.

Oppenheim, A. N. *Questionnaire design and attitude measurement*. London: Heinemann, 1966.

Ramsay, J. O. Review of *Foundations of measurement*, Vol. I, by D. H. Krantz, R. D. Luce, P. Suppes, A Tverskey. *Psychometrika*, 1975, **40**, 257–62.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1961, 321–34.

Rasch, G. A mathematical theory of objectivity and its consequences for model construction. Paper read at the European Meeting on Statistics, Econometrics and Management Science. Amsterdam, 1968.

Rasch, G. On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 1977, **14**, 58–94.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, 1969, **34**, (2, No. 17).

Spada, H. F. and Fischer, G. H. Latent trait models and the problem of measurement in projective techniques (Rorschach, Holtzman). In A. Serrate (Ed.), *VIII International Congress of Rorschach and Projective Methods*. Zaragoza, Spain: 1973.

Spada, H. and Kempf, W. F. (Eds). *Structural models of thinking and learning*. Bern: Huber, 1977.

Thurstone, L. L. The scoring of individual performance. *Journal of Educational Psychology*, 1926, **17**, 446–57.

Thurstone, L. L. A law of comparative judgement. *Psychological Review*, 1927, **34**, 278–86.(a)

Thurstone, L. L. The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, 1927, **21**, 384–400.(b)

Thurstone, L. L. Psychophysical analyses. *American Journal of Psychology*, 1927, **38**, 368–89.(c)

Thurstone, L. L. Attitudes can be measured. *American Journal of Sociology*, 1928, **33**, 529-54.

Thurstone, L. L. Rank order as a psychophysical method. *Journal of Experimental Psychology*, 1931, **14**, 182–201.

Thurstone, L. L. *The measurement of values*. Chicago: University of Chicago Press, 1959.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Western Australia. Education Department. Australian open area schools. Unpublished report, 1974.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, **14**, 97–116.

Wright, B. D. and Douglas, G. A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1977, **1**, 281–94.

Wright, B. D. and Masters, G. N. *Rating scale analysis.* Chicago: MESA Press, 1980.

Wright, B. D. and Mead, R. J. The use of measurement models in the definition and application of social science variables. Report to the U.S. Army Research Institute for the Behavioural Sciences, Virginia, 1977.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, **29**, 23-48.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement.* Chicago: MESA Press, 1979.

REACTANT STATEMENT

Charles Poole

As a teacher interested in measurement I would be decidedly sceptical of any claims that one model of measurement fitted all my various requirements better than another. Much of my measuring activity takes the form of dealing out rough justice to answers submitted in essay form. Admittedly the essay form is not chosen for measurement reasons but because that form of activity fits with the educational aims of the course. However I have found the classical model helpful in warning me of the weaknesses of the technique and in developing weighting schemes to implement policies I have determined for the course. The concerns of the Rasch model developers seem mostly far removed from these activities. For most such purposes the classical model will do just as well.

Andrich has made a good start in his paper towards producing an approach to scaling which suggests that here the Rasch model has considerable advantages over the classical model. Not only do we have a rationale for the Likert-type scales so beloved by designers of questionnaires, but he has also provided some links with the Thurstone scaling techniques. It was of great interest to me to be reminded that Thurstone was aware of the sample-free nature of his scale values and that he recognized the power this gives to obtain measures from incomplete tests.

The finding that the so-called neutral category does not necessarily fall in the centre of the scale neatly demonstrates the validity of the disquiet often expressed about this assumption and gives us good reason for removing this choice from the offered responses. We need a wide variety of studies like this one to make clear the usefulness of the Rasch model in various educational settings. I believe the effort is worthwhile from a teacher's point of view.

Despite my day-to-day involvement with essay examining, it is as a teacher that I appreciate the advantages in adopting the outlook fostered by use of the Rasch model. To think in terms of a child moving up an ability scale as his skill increases better fits with the modern notions of a teacher's task. There is much less stress these days on sorting out the sheep from the goats. No one really wants to know who usually comes top, nor do teachers want to stigmatize children by placing them on a lowly rank in the class. Most teachers would be delighted to be removed from the tyranny of the common examination paper and would appreciate far more the model which suggests that children should be faced with test materials that they find challenging but not impossible.

I share Choppin's uneasiness about rejecting items to meet the assumption of unidimensionality. It would be difficult to accept an attitudinal variable defined by what was left of a set of items after rejection of those not fitting the model. The more appropriate action would seem to be to question the decision rule which allowed these items into the scale. Every effort should be made to revise the decision rules so that unidimensional scales result. At present such revision is better handled by multivariate methods not yet developed within the framework of the Rasch analysis.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

6

Conditional Inference in a Generic Rasch Model

Graham A. Douglas

The aim of this paper is to give an overview of the current state of the *conditional inference argument* as it pertains to a class of statistical models for measuring latent traits which we have come to term, 'Rasch models'. There are a number of reasons why we have chosen this more technical topic for this conference.

In the first place, much of what we in Australia do know about latent trait models, and Rasch models in particular, comes from the American and, to a lesser extent, the English literature. Many of us were nourished on a healthy diet of Wright and others (1968; 1969; 1977) and, while it would be unfair to claim that they tend to ignore the conditional probability arguments expounded by Rasch himself (1960), it is nonetheless true that these arguments and their ramifications are little known or understood in this country. An important factor in this relative ignorance is that much of the subsequent elaboration of Rasch's ideas is to be found in the European literature, and then most of that in the German language.

It is somewhat ironic that, on the one hand, psychometricians working in the field have so readily accepted the unconditional argument and its related computing algorithms, while at the same time embracing the concept of 'specific objectivity' or as Wright (1968) would call it, 'sample-free parameter estimation', because without some form of conditional inference argument it is difficult to demonstrate, at least algebraically, that Rasch models do possess this property of specific objectivity. Conditional arguments alone produce probability expressions which depend on only one set of parameters at a time.

A second reason relates to an increasing interest in and preoccupation with tests-of-fit of data to Rasch models, and in particular to the power of these tests. A debate between Wright (1977) and Whitely (1974; 1977) is an example of this concern. Recent articles by Gustafsson (1977; 1979; 1980) have demonstrated that, whereas an unconditional algorithm may lead very quickly to almost correct parameter estimates in the binary item analysis model, the asymptotic properties of the approximate unconditional tests-of-fit which usually follow such estimation are far from known.

Another reason for looking into this topic pertains to the proliferation of a wide variety of statistical models which we might wish to call Rasch models. The time appears opportune to attempt an initial generalization and synthesis of the principles which underlie the structure of such models. Once again this attempt will borrow extensively from the German literature, and especially from the work of Scheiblechner (1971; 1977), even though we hope to go beyond his developments. I will present the logic and the algebra associated with a general model for measurement, parallel its derivation with that of the binary item analysis model as an illustration, and then highlight significant details of two other models which may be derived from the generic form. Whereas all models derivable from our generalization are legitimately *models for measurement*, we may find the logic and even the language far removed from the familiarity of the binary item analysis model which we usually refer to as *the* Rasch model of educational and psychological measurement. The intention then is to be sufficiently general to encompass a collection of models with common characteristics suitable for the diversity of measurement problems which arise in the behavioural sciences.

A final reason for choosing *conditional inference* is to extend the arguments on conditional versus unconditional algorithms by suggesting, through example, that not all the numerical problems which have in the past been associated with the conditional approach have been solved, and that therefore there is ample scope for major developments in numerical approximations which will still allow us to stay within the rubric of the conditional framework.

The body of the paper is structured into four main sections: a definition of what constitutes a generic Rasch model within the class of latent trait models; an algebraic generalization of this definition and examples of particular cases; an identification of some major problem areas in the implementation of the conditional arguments in practice; and some suggested remedies and directions for the future.

DEFINITION OF A RASCH MODEL

By now the concept of specific objectivity has had sufficient exposure in the literature that we come automatically to equate it with the models of Rasch (1960). It will be valuable for the following, however, to re-state the principle since other ways in which to view the defining characteristics of Rasch models are simply variations of or equivalences to this principle.

By *specifically objective comparisons* among entities, we mean comparisons at a ratio level among a number of entities of a set, such comparisons being uninfluenced in any way by any other entities which may belong either to the same class as those being compared or to completely different classes. For example, the comparison of the difficulty levels of two items in an achievement test is specifically objective if the comparison does not depend on which particular population of subjects is used to arrive at the estimates of item difficulty nor on the difficulties of any other items which might happen to be estimated at the same time. As Wright (1968) re-phrased this principle in an oft-quoted analogy:

> it seemed that . . . my ability depended not only on *which* items I took but on *who* I was and the company I kept . . . we hoped for easy tests so as not to make us look dumb. (p. 85)

Rasch's emphasis on specific objectivity in his writings (1960; 1968; 1977) stemmed from his belief that the principle was not only prevalent but paramount in the rapid development during the last few centuries of laws in physical sciences, even to the extent that one usually takes for granted that one's comparisons of physical entities are of this nature. Such is Rasch's conviction that the principle is all pervading in science that his latest writings (1977) have centred on the common theoretical structure underpinning specific objectivity and its variants in all sciences. The current direction of research in this area is towards a group-theoretic analysis of the concept and some unpublished preliminary work has been completed by Borchsenius (1974; 1977).

Rasch argues that, from the practical point of view, specific objectivity allows one to concentrate attention on analysis, estimation, and fit of one set of parameters at a time in frameworks which usually include potentially many other sets of parameters.

An alternative expression used frequently by Rasch is that of 'separability of parameters'. Separability and specific objectivity may be shown to be synonymous but the former term hints more closely at the logical and algebraic properties inherent in models possessing specific objectivity. By separability of parameters, we mean that, apart from simple and trivial transformations, the pertinent parameters in probability models for measuring must have such a structure that they are capable of separation into disjoint sets. Since, for example, the discrimination parameter, α_i, in Lord's two-parameter logistic model (Lord and Novick, 1968) always acts upon item difficulty δ_i and subject ability β_v, in a multiplicative manner, there is no way to separate difficulty and discrimination parameters in this particular model. This can clearly be seen if we write Lord's model in the following form:

$$P\{X_{vi} = x_{vi}\} = \frac{e^{\alpha_i(\beta_v - \delta_i)x_{vi}}}{1 + e^{\alpha_i(\beta_v - \delta_i)}} \tag{1}$$

where (i) α_i is the item discrimination parameter,
(ii) δ_i is the item difficulty parameter,
(iii) β_v is the subject ability parameter
and (iv) x_{vi} is an indicator variable taking on the value 1 whenever the item is answered correctly, and value zero, otherwise.

In the following we will refer to (1) as the *exponential form* of the model.

It is worth digressing at this stage to emphasize some points concerning the chronological emergence of various models and to set the record straight on comments like 'the Rasch model is just the simple one-parameter case of Birnbaum's two-parameter logistic model'. It would appear that Birnbaum (1968) adopted the logistic model as a mathematical convenience because of some intractable estimation problems associated with the two-parameter normal model; in fact, many expositions of the logistic model (to the base e) contain a multiplicative scaling constant (usually set at approximately 1.7) to bring the logistic ogive more into line with the normal ogive. Certainly the one-parameter logistic was viewed by both Birnbaum and Lord as the special case of a two-parameter model in which all discriminations were set equal to one another.

As far as Rasch was concerned, the *logistic* form of the model arose as a mathematically necessary consequence of his insistence on the principle of specific objectivity in comparisons. The base e is purely a convenience (any base will suffice for a Rasch model), there is no allusion to normal ogives and, although Rasch recognized that lack-of-fit of data to his models was a consequence of more than one parameter operating for each item, it is unlikely that he thought of the second parameter in terms of *discrimination*, and certainly that word does not appear in his 1960 book. Rasch's model was never the consequence of simplifications to a higher-order model but the necessary result of fundamental measurement principles, principles of such generalizability that they could be applied to measurement situations well beyond those rather narrow ones conceived of by many psychometricians working on the other side of the Atlantic; hence the subject matter of this paper, the 'generic Rasch model'.

It is but a short step from the relative looseness of separability to the greater mathematical precision of additivity of parameters, and it is here that Rasch becomes quite explicit about how the parameters in his models must be amalgamated. The additivity pertains to models written in the exponential form; in (1) above, the item difficulty and subject ability parameters appear additively and hence, with all α_i set equal to one another, we have a model in which all parameters appear in the additive form—a model therefore with specific objectivity.

Another term which we use interchangeably with the preceding three is 'sufficient statistic'. Rasch relies heavily on the work of Fisher relating to sufficient estimators, conditional probabilities, and likelihood functions to the extent that much of what is currently known in the mathemetical statistics literature about conditional likelihoods in general arose from Rasch's interest in them with respect to his models for measuring psychological processes. The fact that the conditional probabilities in Rasch's models have known, well-behaved, and potentially useful properties means that the central theorem describing the asymptotic behaviour of unconditional maximum likelihood (u.m.l.) estimates may be extended to conditional maximum likelihood (c.m.l.) estimates. This discovery paved the way for Andersen (1973a) to develop powerful tests of the fit of Rasch models to their respective data sets.

Although it has been demonstrated many times before, the fact that Lord's two-parameter logistic model for item analysis does not exhibit a genuine sufficient statistic for the ability parameter, β_v, bears repetition. The expression

$$T_v = \sum_{i=1}^{k} \alpha_i x_{vi}$$

does not constitute a statistic, let alone a sufficient one and there is little gain in claiming that T_v is sufficient if we know the values of the α's since nearly always we are forced to try to estimate the α's along with the δ's. In fact, were the α's to be known, Andersen (1977) has shown that there can be no *data reduction* in using T_v unless all the α's are equal and so we are forced back onto response patterns—a situation we are trying to avoid since this means that, in order to know something about a subject's ability, we would have to retain the complete set of original data on that subject and not just the summary which is embodied in the sufficient statistic called the raw score. No further information about a subject's ability may be gained beyond a knowledge of the raw score.

It is important in orienting oneself to the concepts involved in Rasch models to point out that in these models the conditional inference argument is used not so much to identify groups of sufficient statistics for the purpose of estimating the parameter sets with which they are associated (as is usually the case in statistical models), but more to *eliminate or condition-out* those sets of unwanted or incidental parameters thus clearing the way for estimation of another set of parameters. Scheiblechner (1977) makes the point abundantly clear when he employs the terms 'general' and 'idiosyncratic' to describe features of the psychological processes which we attempt to model. Conditional inference arguments are designed to remove the idiosyncratic features (i.e. individual differences) which have tended to handicap the development of our understanding of the more general psychological processes.

Furthermore, by extending the conditional inference argument fully, we are always led to probability statements about the data which are completely free of all parameters and hence we set the stage for exact probability (non-parametric) tests of fit. With such tests there is never any doubt about distributional assumptions or the dubious application of asymptotic theory. For Rasch this was the most powerful consequence of adopting the conditional inference stance.

Separation of parameters (and hence sufficient statistics) arises because of the additive (non-interactive) relationships among the parameters. Without this feature the separation does not occur and the conditional argument breaks down. It is not coincidental that, for models of the type considered by Rasch, the necessity for a conditional probability argument could be developed from the work of Neyman and Scott (1948) on *incidental* and *structural* parameters leading to inconsistent estimates. In the framework of binary item analysis where the emphasis is on item estimation, the Neyman and Scott dilemma means that each potentially new subject in the calibrating population carries, in his or her response vector, a certain amount of information about each of the items, plus a new parameter associated with his or her own ability. Thus the number of incidental ability parameters could increase without bound.

We will try to incorporate the various concepts described in previous paragraphs into a general model which exhibits all of these properties in addition to those properties that we usually demand from any latent trait model. We will argue that the generic model represents a probabilistic definition of 'Rasch model' in the sense that all models which appear in the literature under this rubric are derivable from our generic form, and that all models which do not fit into the mould cannot genuinely be called Rasch models.

Our aim in attempting this generalization is not to constrain investigation of probabilistic models in the social sciences just to those which do exactly fit our framework. There may be models which do not conform in various ways to the general expression but which nevertheless display interesting and valuable properties: for example, the dynamic test model of Kempf (1976). Still, within a well-defined set of assumptions based on the concepts under discussion here, a wide variety of models follow to which the label Rasch model may be attached.

THE GENERIC FORM OF A RASCH MODEL

The crucial principle of any Rasch model is the way in which the parameter structure factors into additive components. Hence any general parameter, θ, must be factorable into other undimensional or multidimensional parameters which add together in the exponent. This restriction confirms, for example, that in the binary item analysis model there is no place for either a discrimination parameter α_i, since it necessarily *multiplies* β_v and δ_i, nor is there place for Lumsden's (1977) person sensitivity parameter ψ_v, for exactly the same reason.

Although we are accustomed to thinking in terms of two interacting facets in a Rasch model (the test items and the responding subjects in the binary item analysis model), we must provide in this general framework for any number of facets† interacting simultaneously. Each facet consists of a number of elements and by the term 'interaction' we will mean the simultaneous confrontation of one element from each of the facets. For example, one marker assessing the essay writing ability of one subject on one essay question represents a single individual interaction in a three-facet framework. The totality of observations may be represented in a data 'cube' of as many dimensions as there are facets. Marginal summaries of the data cube, either of one or many dimensions, may be effected by summing the individual responses across various combinations of the facets.

Not all responses are binary. In quantifying attitude questionnaire items, for example, we may allow for multiple category responses scored with the integers 0,1, . . ., m, instead of the usual 0,1, of binary item analysis. Most often the number of response categories is fixed in advance, but there are measurement models in which the number of categories is open-ended. An example of this situation will be discussed in the next section.

While the basic multinomial random variable in our models always represents the response when an individual interaction occurs, we will find it more convenient when expressing the model in its statistical form to use an indicator variable which takes on the value 1 whenever response category h is used and takes on the value of zero, otherwise. This means that the basic unit of observation is a set of $(m+1)$ responses, all of which are zero except for the hth element, which takes the value 1.

With these preliminary comments we are now in a position to give a formal statement of the generic Rasch model for measurement.

(i) A total of t facets of an observational framework are in

† Following Guttman and Cronbach, Rasch's term 'factor' is avoided because of its obvious alternative connotations in psychology.

simultaneous interaction such that each individual response, $x_{i_1 i_2 \ldots i_t}$, may take on a value h in the range $h = 0, 1, \ldots m$.

(ii) $i_1, i_2, \ldots, i_t$ are index sets for the respective facets and there are N_s elements in facet s.

(iii) The data may be arranged in a t-dimensional hypercube such that it is possible to calculate 1-dimensional, 2-dimensional, etc. marginal summaries of the data.

(iv) $\theta_{i_1 i_2 \ldots i_t h}$ is a general function of parameters which is factorable into w additive component parameters as follows,

$$\theta_{i_1 i_2 \ldots i_t h} = \mu_1 + \mu_2 + \ldots + \mu_w$$

where the subscript j in μ_j stands for a sub-set of the indices $i_1, i_2, \ldots, i_t$. μ_j represents a continuous latent trait (unknown property) of facet combination j and is manifested in the observed data.

(v) $x_{i_1 i_2 \ldots i_t h}$ is an indicator variable, taking on the value 1 whenever response category h is utilized by facet combination $i_1 i_2 \ldots i_t$, and taking on the value zero otherwise.

(vi) $X_{i_1 i_2 \ldots i_t}$ is a random variable whose observed value is $x_{i_1 i_2 \ldots i_t}$.

(vii) All interactions are stochastically independent, conditional on the parameters in θ.

With these notations we may write the probability of the random variable $X_{i_1 i_2 \ldots i_t}$ taking on the value $x_{i_1 i_2 \ldots i_t}$ in terms of a general exponential function of the parameter θ and the indicator variable $x_{i_1 i_2 \ldots i_t h}$, as

$$P\{X_{i_1 i_2 \ldots i_t} = x_{i_1 i_2 \ldots i_t}\} = \frac{e^{\sum_{h=0}^{m} \theta_{i_1 i_2 \ldots i_t h} x_{i_1 i_2 \ldots i_t h}}}{\sum_{h=0}^{m} e^{\theta_{i_1 i_2 \ldots i_t h}}} \qquad (2)$$

Example 1: Binary Item Analysis Model

The familiar binary item analysis model arises by making the following changes to symbols in (2).

(i) $i_1 = v$ and $i_2 = i$
with $v = 1, \ldots, N$ subjects and $i = 1, \ldots, k$ items.
$h = 0, 1$.

(ii) Set $\theta_{i_1 i_2 h} = \mu_1 + \mu_2$
with $\mu_1 = h\beta_v$
and $\mu_2 = -h\delta_i$.
β_v is the (latent) ability parameter of subject v
δ_i is the (latent) item difficulty parameter of item i.

(iii) Set $x_{vih} = 1$ whenever subject v gets item i correct, and $= 0$ otherwise. Then the generic model takes the form,

$$P\{X_{vi} = x_{vi}\} = \frac{e^{\sum_{h=0}^{1} h(\beta_v - \delta_i)x_{vih}}}{\sum_{h=0}^{1} e^{h(\beta_v - \delta_i)}} = \frac{e^{(\beta_v - \delta_i)x_{vi1}}}{1 + e^{\beta_v - \delta_i}}$$

We must now demonstrate that the property of specific objectivity follows from this model. In order to do that we must derive the (unconditional) joint probability of all the data, the joint probability of marginal statistics suitably identified, and the conditional probability of one marginal set, conditional on all others. In order to know which marginals to consider in the conditioning, we must select a parameter set which is to be the subject of current investigation; for the binary item analysis model, for example, we usually focus attention first on calibrating a fixed set of items, in which case the subject parameter set, (β), is incidental. On the other hand, when the emphasis is on measuring the abilities of a fixed number of subjects employing items from an item pool, the item parameter set, (δ), is considered to be incidental. A measurement perspective is necessary before adopting the conditional probability argument.

As an appendage to the mainstream argument we will demonstrate, in the manner of Rasch (1960), that completely parameter-free tests of fit follow by extending the conditional argument to its limit.

The conditional probability (likelihood) of the total data set is given by the continued product of the probabilities of the $N_1 N_2 \ldots N_t$ individual responses and may be written as

$$L = P\{X_{11\ldots1} = x_{11\ldots1}, \ldots, X_{v_1 v_2 \ldots v_t} = x_{v_1 v_2 \ldots v_t}\}$$

$$= \frac{e^{\sum_{i_1} \sum_{i_2} \cdots \sum_{i_t} \sum_{h=0}^{m} (\mu_1 + \mu_2 + \ldots + \mu_w) x_{i_1 i_2 \ldots i_t h}}}{\prod_{i_1} \prod_{i_2} \cdots \prod_{i_t} \sum_{h=0}^{m} e^{\mu_1 + \mu_2 + \ldots + \mu_w}} \qquad (3)$$

We have already replaced θ by its appropriate factorization since the structure of θ is determined by the model builder and not by any of the algebra of the derivations.

It will always be possible to re-write the numerator of this expression in such a way that the set of sufficient statistics for the set of parameters, (μ_j), is clearly indicated. We will write (x_{+j}) for that data summary (marginal) which arises from summing $x_{i_1 i_2 \ldots i_t h}$ over h and over all index sets which are *not* included in j. Hence we have

$$L = \frac{e^{\sum_1 \mu_1 x_{+1} + \sum_2 \mu_2 x_{+2} + \ldots + \sum_w \mu_w x_{+w}}}{\prod_1 \prod_2 \ldots \prod_w \sum_{h=0}^{m} e^{\mu_1 + \mu_2 + \ldots + \mu_w}} \tag{4}$$

In the Binary Item Analysis Model,

μ_1 contains the single-dimension parameter β_v,
μ_2 contains the single-dimension parameter δ_i,

and therefore

$$L = \frac{e^{\sum_{v=1}^{N} \beta_v x_{v+} - \sum_{i=1}^{k} \delta_i x_{+i}}}{\prod_{v=1}^{N} \prod_{i=1}^{k} (1 + e^{\beta_v - \delta_i})}$$

where we usually write r_v for x_{v+} and s_i for x_{+i}.

Since (4) is written in terms of all relevant marginals, $x_{+1}, x_{+2}, \ldots, x_{+w}$, and thus does not contain the original data explicitly, the joint probability of all marginal sets is simply C times (4), where C is the number of possible complete data sets which could have produced exactly the observed marginals, $x_{+1}, x_{+2}, \ldots, x_{+w}$. ($C$ is a combinatorial number for which no algorithm, other than listing, has as yet been developed.) The probability is written as

$$\begin{aligned} L_u &= P\{(x_{+1}), (x_{+2}), \ldots, (x_{+w})\} \\ &= CL \end{aligned} \tag{5}$$

As an aside to the main argument on specific objectivity, but of paramount relevance to testing of fit, we may demonstrate another property of the conditional inference procedure. Using (4) and (5) we may write the conditional probability of the observed data set given all the marginals, as

$$\begin{aligned} L_c^* = P\{X_{11\ldots1} &= x_{11\ldots1}, \ldots, X_{v_1 v_2 \ldots v_t} \\ &= x_{v_1 v_2 \ldots v_t} \; (x_{+1}), (x_{+2}), \ldots, (x_{+w})\} = \frac{1}{C} \end{aligned} \tag{6}$$

This probability (likelihood) is free of *all* parameters in the model and its value as a tool in testing fit will be commented upon in a later section.

Returning to the main development, we now focus attention on one component, say (μ_i), and its associated marginal, (x_{+i}), in order to find the marginal probability of all marginal sets *other than* (x_{+i}). We do this by summing the unconditional probability over all possible values of (x_{+i}) which are compatible with the remaining marginals, (x_{+1}), (x_{+2}), . . ., (x_{+i-1}), (x_{+i+1}), . . ., (x_{+w}).

Hence $$L_{\cdot} = P\{(x_{+1}),(x_{+2}), \ldots, (x_{+i-1}),(x_{+i+1}), \ldots, (x_{+w})\}$$
$$= \Sigma C^* L$$
all (x^*_{+i}) such that all other marginals are fixed.

$$= \frac{e^{\sum_1 \mu_1 x_{+1} + \sum_2 \mu_2 x_{+2} + \ldots + \sum_{i-1} \mu_{i-1} x_{+i-1} + \sum_{i+1} \mu_{i+1} x_{+i+1} + \ldots \sum_w \mu_w x_{+w}} \gamma_{(x_{+i})}[\mu_i]}{\prod_1 \prod_2 \ldots, \prod_w \sum_{h=0}^{m} e^{\mu_1 + \mu_2 + \cdots + \mu_w}} \tag{7}$$

where a symmetric function in the parameter set (μ_i) is defined as

$$\gamma_{(x_{+i})}[\mu_i] \equiv \Sigma\, C^* e^{\sum_i \mu_i x^*_{+i}}$$
all (x^*_{+i}) such that all
other marginals are fixed.

In the Binary Item Analysis Model, with attention focused on the estimation of the item parameter set, (δ), we have to determine a symmetric function in the set, (δ), which is defined as the sum over all possible item-count marginal sets, (S^*), which are compatible with the observed raw score set, (r). In practice a number of different binary data matrices, (C^* of them), will result in the same set, (S^*), and this number C^* must be included in the definition, as shown,

$$\gamma_{(r)}[\delta] \equiv \Sigma\, C^* e^{-\sum_{i=1}^{k} \delta_i S^*_i},$$

all (S^*) such that (r) is fixed.

Then $$L_{\cdot m} = \frac{e^{\sum_{v=1}^{N} \beta_v r_v} \gamma_{(r)}[\delta]}{\prod_{v=1}^{N} \prod_{i=1}^{k} (1 + e^{\beta_v - \delta_i})}$$

Finally we need the conditional probability of the marginal set of interest, (x_{+i}), given all other marginal sets which we are attempting to eliminate. This conditional probability is obtained by dividing (5) by (7) to obtain

$$L_c = P\{(x_{+j}) \mid (x_{+1}), (x_{+2}), \ldots, (x_{+j-1}), (x_{+j+1}), \ldots, (x_{+w})\}$$

$$= \frac{C\, e^{\sum_j \mu_j x_{+j}}}{\gamma_{(r_{+j})}[\mu_j]} \tag{8}$$

which is dependent only on the parameter set, (μ_j),and not on any other parameters in the model.

In the Binary Item Analysis Model, upon division of the joint probability by the marginal probability, we have the conditional probability of the set of marginal item-counts, given the set of raw scores, as

$$L_c = \frac{C\, e^{-\sum_{i=1}^{k} \delta_i S_i}}{\gamma_{(r)}[\delta]}$$

This conditional probability depends only on the item parameters and not on the subject ability parameters.

To complete the algebraic story, we next set the vector of first derivatives, with respect to the μ's, of the log-likelihood, equal to a zero vector to obtain a set of conditional maximum likelihood (c.m.l.) equations.

$$\frac{\partial L}{\partial \boldsymbol{\mu}_j} = \mathbf{x}_{+j} - \frac{\partial \ln \gamma_{(r_{+j})}[\mu_j]}{\partial \boldsymbol{\mu}_j} = \mathbf{0} \tag{9}$$

Upon solving these equations, we insert the c.m.l. estimates into a matrix which is the negative inverse of the matrix of second derivatives of the log-likelihood, thus arriving at

$$V[\hat{\boldsymbol{\mu}}_j] = -\left[\frac{\partial^2 \ln \gamma_{(r_{+j})}[\mu_j]}{\partial \boldsymbol{\mu}_j\, \partial \boldsymbol{\mu}_j'}\right]^{-1} \tag{10}$$

which is an $N_j \times N_j$ matrix of estimated error covariances and from which the asymptotic standard errors of the $\hat{\mu}_j$'s are obtained by extracting the square roots of the diagonal elements.

In the Binary Item Analysis Model, the maximum likelihood equations and the error covariance matrix are given by

$$\frac{\partial L}{\partial \boldsymbol{\delta}} = -\mathbf{S} - \frac{\partial \ln \gamma_{(r)}[\delta]}{\partial \boldsymbol{\delta}} = \mathbf{0}$$

and

$$V[\hat{\boldsymbol{\delta}}] = -\left[\frac{\partial^2 \ln \gamma_{(r)}[\delta]}{\partial \boldsymbol{\delta}\, \partial \boldsymbol{\delta}'}\right]^{-1}$$

A perhaps more familiar but algebraically identical version of the c.m.l. equations presented above, as might be found in Andersen (1972) or Wright and Douglas (1977), takes the form,

$$-S_i + \sum_{r=1}^{k-1} n_r e^{-\delta_i} \frac{\gamma_{r-1,i}}{\gamma_r} = 0, \qquad i = 1, k$$

where (i) n_r is the number of subjects with a raw score of r,

(ii) γ_r is the rth order symmetric function in the set (δ), defined as

$$\gamma_r \equiv \Sigma\, e^{-\sum_i \delta_i x^*_{vi}}$$

all response vectors (x^*_{vi})
such that r_v is fixed,

(iii) $\gamma_{r-1,i}$ is the ($r-1$)th symmetric function in which all terms involving δ_i have been removed.

The other two examples which follow are presented in the style of Andersen (1972, 1973a). The difference between his approach and that of Rasch which we have adhered to is that, for the binary item analysis model, Andersen derives the conditional probability of the response vector for a *single* subject (conditional on that subject's raw score) and then arrives at the conditional likelihood of the total data matrix by taking the product of the individual likelihoods over all N subjects. This procedure produces a different likelihood, avoids the use of the numbers C and requires the calculation of separate symmetric functions for each raw score. Despite the different likelihoods, the c.m.l. equations are the same as those of Rasch and thus we are led to the same parameter estimates.

In addition to the binary item analysis model and the two models yet to be considered, some other models which fit into our framework and which have received varied attention in the literature are:

(a) the speed of oral-reading model of Rasch (1960);
(b) sociometric choice models, Scheiblechner (1971);
(c) the multi-dimensional questionnaire model, Andersen (1972);
(d) the measurement-of-change models, Fischer (1976), which introduce the facet of *time*;

(e) the linear logistic model, Fischer (1977);
(f) the multiplicative binomial model, Andrich (1978a), Douglas (1978);
(g) the grader/subject/item model, Malone (1980).

Example 2: The Rating Model, Andrich (1978b)

This model is one of an hierarchy of models derived from the multidimensional model of Andersen (1972). Starting from a general parameter θ_{vih} which may be written as

$$\theta_{vih} = \beta^*_{vh} - \delta^*_{ih}$$

(with v, i and h taking their usual meaning), we further factor this into

$$\beta^*_{vh} = \beta_v \phi_h + \kappa_h$$
$$\delta^*_{ih} = \delta_i \phi_h$$

Andersen identifies the ϕ's as the 'scoring' parameters and the κ's as the 'category' parameters. This model appears pertinent to any rating situation in which a series of items (or more generally, questions), each permit a response on a scale which may be quantified from 0 to m. Andrich (1978a) and in his paper prefers to work with a simple transformation of the κ's:

$$\kappa_h = - \sum_{l=1}^{h} \tau_l \, (h = 1, \ldots, m)$$

He calls the τ's the 'threshold' parameters, after Thurstone.

In order for a genuine Rasch model to eventuate, we must set the scoring parameters equal to consecutive integers, that is

$$\phi_h = h (h = 0, 1, \ldots, m).$$

Only by doing this will we have a genuine sufficient statistic,

$$x_{+i+} = \sum_{v=1}^{N} \sum_{h=0}^{m} h x_{vih}$$

for the structural (item) parameters, thus permitting their elimination from the conditional likelihood. As Andersen (1977) and Andrich (1978b) have frequently pointed out, such a 'restriction' is very much in keeping with the notion of integer scoring advocated by Likert (1932) and does appear appropriate for a wide class of data sets of the rating or attitudinal type.

There still remains the question of the identifiability of the category parameters, κ_h. Some simple algebra will show that since the marginal set $((x_{v+h}))$ – the number of times that subject v used category h – is sufficient for the parameter combination set $((h\beta_v + \kappa_h))$, the category parameters are well defined and are permitted in the model since they will be eliminated (along with the subject parameters) when the conditional likelihood is obtained. With these comments it should be clear that the probability of an individual response may be written as

$$P\{X_{vi}=x_{vi}\}=\frac{e^{\sum_{h=0}^{m}[\kappa_h+h(\beta_v-\delta_i)]x_{vih}}}{\sum_{h=0}^{m}e^{\kappa_h+h(\beta_v-\delta_i)}}$$

and that the conditional likelihood for estimating item parameters follows as

$$P\{X_{11}=x_{11},X_{12}=x_{12},\ldots,X_{NK}=x_{NK}\,|\,((x_{v+h}))\}$$

$$=\frac{e^{-\sum_{i=1}^{k}\delta_i x_{+i+}}}{\prod_{v=1}^{N}\gamma_{x_{v+h}}[\delta]}$$

Although we have as yet made no mention of numerical problems associated with the estimation of the parameters in any of the models, it should be noted that, in applications of this model to real data, Andrich at least has used an estimation algorithm based on the unconditional likelihood, and then has 'corrected' the estimates to bring them more into line with what would arise had the conditional likelihood been correctly used; more on these problems later.

Example 3: The Rasch (1960)/Andrich (1973) Essay Grading Model

A perennial problem in the grading of extended response answers (such as essays), when more than one grader is involved in the marking, is the question of the varying grader harshnesses which result in comparisons among subject essay-writing abilities which are highly suspect. One avenue out of this dilemma has been to train graders to such a level of consistency that we are prepared to accept all graders as virtual replications of one another. Problems of marker reliability are then assumed to have been controlled. This training, however, is never very satisfactory and, even more importantly, by trying to force all graders into the same mould, we lose potentially important information about the psychological processes involved in essay marking as reflected in the very individual idiosyncracies that we are trying to eliminate. It is preferable to control grader differences but still retain information about them than to throw away that information altogether.

Andrich devised a model in which grader harshness was explicitly parameterized; because the model was a Rasch model and the specific objectivity property held, it was possible to estimate subject essay-writing ability independently of the particular grader involved in the marking. In this example we see a straightforward application of the strategy of identifying a sufficient statistic for the fundamental purpose of parameter elimination.

The model adopted by Andrich had formal similarities to the one for errors in oral reading described by Rasch in his 1960 book. In that development and in the thesis work of Andrich, the Poisson distribution was the starting point. We will show now that by means of simple transformations this model arises naturally from our generic form.

In the first place we assume that graders are permitted to detect an unlimited number of errors in a subject's script. This direction to graders to use an open-ended scoring scale is equivalent to letting m tend to infinity in the list, $h = 0, 1, \ldots, m$.

Furthermore, since the basic random variable in this model is the number of *errors* detected, the subject ability parameter β_v should enter the model with a negative sign if we are to make the same interpretation of it as we have done in other models. With these minor variations and with η_g as the grader parameter, we may write the probability of an individual response when grader g assesses the script of subject v as

$$P\{X_{vg} = x_{vg}\} = \frac{e^{\sum_{h=0}^{\infty} [\ln(\frac{1}{h!}) + h(\eta_g - \beta_v)] x_{vgh}}}{\sum_{h=0}^{\infty} e^{\ln(\frac{1}{h!}) + h(\tau_{ig} - \beta_v)}}$$

Let us consider further elaboration of the model from the perspective of estimating the grader harshnesses, η_g. It is not difficult to see that x_{v++} is sufficient for β_v and that x_{+g+} is sufficient for η_g. The conditional likelihood is written as

$$P\{X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{Vg} = x_{Vg} \mid (x_{v++})\}$$

$$\frac{e^{\sum_{h=0}^{\infty} \ln(\frac{1}{h!}) x_{++h} + \sum_{g=1}^{G} \eta_g x_{+g+}}}{\prod_{v=1}^{V} \gamma_{v++}[\eta]}$$

For those of us for whom statistics are quite often a mystery, the relationship between this likelihood and the Poisson distribution appears rather remote. However, if we make the following transformations,

$$A_g = e^{\eta_g}$$

and

$$B_v = e^{\beta_v}$$

and note, through a first-principles definition of the exponential function, the infinite sum in the denominator,

$$\sum_{h=0}^{\infty} e^{\ln(\frac{1}{h!}) + h(\beta_v - \delta_i)}$$

converts to

$$\sum_{h=0}^{\infty} \frac{1}{h!} \left(\frac{A_g}{B_v}\right)^{h}$$

which is just a definition of

$$e^{A_g/B_v}$$

and furthermore that the obtuse expression

$$e^{\sum_{h=0}^{\infty} \ln(\frac{1}{h!}) x_{vgh}}$$

may be re-written more simply in terms of factorials as

$$\frac{1}{x_{vg+}!},$$

we arrive finally at an equivalent form of the model,

$$P\{X_{vg} = x_{vg}\} = \frac{e^{-A_g/B_v} (A_g/B_v)^{x_{vg+}}}{x_{vg+}!}.$$

This distribution is directly Poisson with parameter $\lambda = A_g/B_v$. Despite the fact that it is Poisson rather than logistic we may still apply our marginal and conditional arguments, by which we are led to the conditional likelihood,

$$P\{X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{Vg} = x_{Vg} \, (x_{v++})\} = \frac{\prod_{g=1}^{G} A_g^{x_{+g+}}}{\prod_{g=1}^{G} x_{+g+}! \prod_{v=1}^{N} \gamma_{v++}[A]}$$

Expressed in this manner, even the symmetric functions may be shown to have a more favourable form, in that we may write

$$\gamma_{v++}[A] = \frac{\left(\sum_{g=1}^{G} A_g\right)^{x_{v++}}}{x_{v++}!}$$

Thus the most transparent form of the conditional likelihood is

$$P\{X_{11} = x_{11}, X_{12} = x_{12}, \ldots, X_{vg} = x_{vg} \mid (x_{v++})\}$$

$$= \frac{\prod_{g=1}^{G} A_g^{x_{+g+}} \prod_{v=1}^{N} x_{v++}!}{\left(\sum_{g=1}^{G} A_g\right)^{x_{+++}} \prod_{g=1}^{G} x_{+g+}!}$$

Andrich (1973) gives details of estimation and tests of fit and shows that the conditional and the unconditional likelihoods give identical parameter estimates. One should not, of course, take this as a sign that there are many other models for which this is true.

NUMERICAL ANALYSIS PROBLEMS ASSOCIATED WITH CONDITIONAL INFERENCE

It is one thing to produce a mathematically rigorous statement of probabilistic models involving many parameters, and another thing to devise efficient and accurate numerical methods to answer the kinds of practical questions we pose about the operation of the models. To a certain extent these arithmetic problems have curtailed the widespread use of Rasch models among social scientists; on the other hand, their apparent intractability has led others to adopt approximations the validity of which is frequently unknown.

In the development of new techniques, however, initial caution must eventually give way to guarded extension if the models are to have acceptance in a wider sphere. Often this means that either assumptions are relaxed or approximations are introduced. It is the latter option which has found most favour with respect to Rasch models.

It would be naïve to imply that the only remaining problems of latent trait models are those associated with numerical methods. However, for the current exercise, we will address ourselves to some of these problems in the hope that, if the problems are well identified, their solution will follow a little more easily. Normally we would not concern ourselves with arithmetic problems like these at a seminar of this nature but, since they are intricately related to the conditional inference argument, their discussion is quite relevant, particularly when crude approximations are used to verify the powerful properties of Rasch models. In a characteristic understatement of Rasch (1960):

> the practical applications of the theory present, however, difficulties yet to be removed. As they are 'only' algebraical and computational, we may hope for a satisfactory way out. (p. 122)

We turn now to a brief discussion of four of these problem areas.

The Symmetric Functions

The number of terms in a middle-order symmetric function for binary-item tests of even moderate length, say 50 to 60 items, is astronomical; for example, the symmetric function $\gamma_{25}[\delta]$ in a 50-item test has in excess of 12.6×10^{13} terms. Our computers are fast, but there are limits. Even granting that these calculations could be made, there still remains the problem that there is no closed explicit form for a symmetric function. All algorithms which determine them, or ratios of them, use a recursive expression which builds up each successive function from those determined previously in the list. This obviously reduces the actual number of calculations involved, but does introduce another more damaging complication – rounding error.

The calculation of the exponential function by a computer necessarily means retaining only a finite number of significant figures in each calculation. If only a few calculations are being made, by using double precision arithmetic there is usually no problem about rounding error. When determining symmetric functions of the order we are describing, however, this rounding error can accumulate dramatically to the extent that negative estimates of the functions arise persistently, causing the estimation algorithm to abort. Hence there are no conditional estimates.

Gustafsson (1979) claims to have solved the rounding error problem by utilizing a number of previously unused recursive relationships among the successive symmetric functions. By employing these and a number of other expedient devices, Gustafsson has written a program for which conditional estimates can be determined for binary-item tests of up to about 100 items, as long as not too many of the items are at the extremes of the difficulty range. Unfortunately we do not have prior knowledge of just how many or how extreme are the items in order to predict whether or not the program will abort. More crucially, there will always be the suspicion that, although rounding errors have started to 'set in', they are not yet of such a magnitude to abort the program. What then of our 'estimates'? Even if the estimates are to be believed, we might still question the effort involved.

Standard Errors

Another numerical problem arises in connection with the algorithm used to solve the c.m.l. equations. If the raison d'être of the exercise was the calculation of the c.m.l. estimates only, then there is little doubt that one would prefer variations of the so-called 'switching' method to the multiparameter version of the Newton/Raphson method. Briefly the switching method for binary item analysis involves solving for δ_i from the equation which arises as a re-arrangement of the c.m.l. equation:

$$e^{-\hat{\delta}_i} = \frac{S_i}{\sum_{r=1}^{k-1} n_r \left(\frac{\gamma_{r-1,i}}{\gamma_r}\right)}$$

$$\therefore \quad \hat{\delta}_i = \ln\left[\frac{\sum_{r=1}^{k-1} n_r \left(\frac{\gamma_{r-1,i}}{\gamma_r}\right)}{S_i}\right] \quad i = 1, 2, \ldots, k.$$

Although more iterations are required to achieve convergence with the switching method, at least a $k \times k$ matrix of second derivatives does not have to be inverted at each iteration, as is requiired with the multiparameter algorithm. But the purpose of the exercise is certainly not just to estimate parameters; it should involve also the determination of the standard errors of those estimates as well as tests of fit.

If we are to adhere to the principles of c.m.l. estimation, then the most appropriate standard errors will be given by the square roots of the diagonal elements of the matrix in (10). If we use the switching method and avoid inversion at each iteration, the inversion would still be necessary after convergence in order to extract the standard errors. The implied criticism by Gustafsson (1979) of unconditional procedures for parameter estimation is not levelled consistently since unconditional standard errors are used by Gustafsson in his programs. Elsewhere there is evidence, Douglas (1978), to suggest that conditional and unconditional standard errors are not the same and, unlike the estimation of the parameters themselves, we do not have known correction factors which enable us to change from one form to the other.

More Complex Models

If the numerical problems are not wholly controlled in the binary item analysis model, there is little surprise in finding that for more complex

Rasch models the situation is quite unclear. For polychotomous models for rating, for example, little is known about how to arrive at the c.m.l. estimates (and their standard errors), other than in relatiyely simple cases where the structural parameters are few in number. Unfortunately the identification of this situation as a problem area is camouflaged by the fact that, in the published literature, most examples limit the number of structural parameters and one could be excused for believing that numerical problems are totally non-existent. Yet in practice we are likely to be dealing with large numbers of structural parameters. As we increase the number of response categories—and five categories is certainly not uncommon—the number of terms in each symmetric function also increases beyond what it would be for the binary case. Although correction factors are applied to unconditional item estimates in the polychotomous model of A1drich (1978a), no published studies are available as there are for the binary model (Wright and Douglas, 1977) which detail thoroughly the circumstances in which the corrected unconditional and the conditional estimates are similar. Naturally the problems of matrix inversion and the standard errors are also magnified in these models.

The advisability of trying to find corrected estimates becomes more questionable the greater the number of sets of parameters in the model. For example, in a 3-facet model, the unconditional algorithm must estimate simultaneously three sets of non-linear equations and must keep a check on not only the convergence within each set but also on the overall convergence to ensure that the complete likelihood is maximized. Once correction factors have been applied to one or more sets of parameters, the likelihood is no longer maximized.

Estimating other Sets of Parameters

One advantage of viewing these models in their most general form is that we are less likely to become fixated on item analysis to the detriment of subject analysis. The area of subject ability analysis in a number of models where it is relevant has received virtually no attention in the literature. If we are to follow the spirit of Rasch and our generic form, the subject ability parameters would be seen as structural parameters in the presence of the incidental item parameters when the items for a test have been selected from a bank or pool. In that case, the focus is on person measurement and we are at liberty to vary the number of items administered.

All of the arguments used previously to derive conditional inference statements with respect to item parameters may be employed in a directly parallel manner to derive conditional inference expressions for the subject ability parameters. Even though the number of subjects being measured simultaneously by a test may be thought of as fixed, this number could be quite large. Hence all numerical problems which we have identified with item estimation are often intensified with conditional subject estimation; one only has to note that for items, there are $\binom{k}{r}$ terms in the rth symmetric function but for subjects, there are $\binom{N}{r}$ terms.

While it is rare that we would wish to calibrate one or two items at a time, it is not uncommon to find contexts (mastery learning, criterion-referenced testing, tailored testing) where one person is to be measured at a time. This situation raises a whole series of conceptual as well as numerical problems when conditional inference is employed, since it is patently impossible to estimate the ability of a single subject conditionally without recourse to the ability of a reference subject – and hence we are back to norm-referenced measurement. It is beyond the scope of this paper to delve into these problems but the dilemma does highlight once again that the models of Rasch are models for *comparisons* and that *absolutes* really have no place here. Claims to the contrary are false.

We may identify at least two procedures (in addition to the conditional) which have been used to arrive at subject measurement in the binary item analysis model. According to Andersen and Madsen (1977, p. 359), 'the *logical* implication is to base the inference concerning the β_v's on the remaining part of the likelihood'. Since unconditional, marginal, and conditional likelihoods are connected via the relationship,

$$L_c = \frac{L_u}{L_m}$$

Andersen is advocating that the expression

$$L_m = \frac{e^{\sum_{v=1}^{N} \beta_v r_v} \prod_{v=1}^{N} \gamma_{r_v}[\delta]}{\prod_{v=1}^{N} \prod_{i=1}^{k} (1 + e^{\beta_v - \delta_i})}$$

be used to estimate the ability parameters. Although the symmetric functions contain no β's, the denominator does involve item parameters as well as β's and it is customary to replace the δ_i's by their c.m.l. estimates. On the other hand, Wright and Panchapakesan (1969) and Andrich (1978a) base the inference on the unconditional likelihood, L_u. Although the likelihoods are different, both approaches produce the same m.l. equations for the β's and Wright substitutes the corrected unconditional δ's rather than the conditional ones. Given that for a very wide class of binary item analysis examples the corrected unconditional item estimates are virtually identical to the conditional ones, the approaches of Andersen and Wright should coincide. However, since correction factors are not well established for models other than the binary item analysis one, we should not expect coincidence of the Andersen and Wright methods in other models.

We might note also that the inconsistency of the unconditional item estimates which leads to their values, $\hat{\delta}_i^{(u)}$, being exactly twice those of the conditional estimates, $\hat{\delta}_i^{(c)}$, in the case of two items, is directly duplicated when attention is turned to estimating the abilities of two subjects. Although to our knowledge no confirmatory studies have been carried out, we might surmise that an identical correction factor which converts unconditional estimates to approximate conditional ones for items,

$$\hat{\delta}_i^{(c)} = \frac{k-1}{k}\hat{\delta}_i^{(u)}$$

also operates in converting unconditional estimates to approximate conditional ones for subjects,

$$\hat{\beta}_v^{(c)} = \frac{N-1}{N}\hat{\beta}_v^{(u)}$$

Clearly this correction factor is not insignificant when we are measuring a small number of subjects.

With respect to the standard errors of subject ability (the equivalent of what psychometricians would refer to as the *precision of measurement*), both the Andersen and Wright methods lead to approximate expressions not involving the inversion of matrices and once again we have little information about whether these are under- or over-estimates of the error of measurement.

SOME DIRECTIONS FOR THE FUTURE

There appear to be a number of possible ways out of the dilemmas of numerical analysis as outlined in the previous section. The most straightforward but uncompromising solution is to follow Gustafsson's example and attempt to improve the algorithms for calculating directly the symmetric functions for all models. We tend to doubt the advisability of this action for models other than those on which it currently works since the numerical problems are inordinately complex. It is not uncommon, for example, to find oneself working with attitude questionnaires of the Likert-type (with five response categories) consisting of something like 50 questions. Since in this case raw scores range from zero to two hundred, it is impossible to analyse these data conditionally with the algorithms presently available. Other alternative solutions must be sought.

An alternative which suggested itself to Rasch himself in the early seventies (personal communication) was to find numerical approximations to the symmëtric functions along the lines of the formula known as 'Stirling's approximation' for higher order factorials. These approximations would take the form of explicit expressions for symmetric function ratios of all orders. Other than initial skirmishes with the problem, little development appears to have taken place in this direction.

A potentially promising approach is offered in related work on *exact probability tests* being carried out by Agresti and his colleagues (1977; 1979). The approach offers advantages not only for parameter estimation via the likelihood equations but, more importantly, for carrying out exact tests of fit. Before outlining Agresti's approach, we should say something further about the tests of fit employed in Rasch models.

The applicability of any model derivable from our generic form rests substantially on the assumption that the model fits the data to within acceptable probability limits; in that case, all the properties of the model on which we place so much importance must follow necessarily. Viewed in this manner, the determination of fit precedes in importance the determination of parameter estimates to the extent that an understanding of the psychological processes underlying the interactions, which give rise to our data, comes from our assumption that we have the correct model. To talk of Rasch models as 'providing specific objectivity' is to understand that these properties obtain in the presence of the model fitting the particular data set. Without fit we really have very little to talk about.

Complications occur when we realize that data fail to fit probability models for many reasons and that it is highly unlikely that we will find a statistical test which will detect lack of fit against all possible alternative models. A test which is suitable for detecting unequal item discriminations in the binary item analysis model, for example, may have virtually zero power for detecting other departures from that model (i.e. a model with equal item discriminations but unequal person sensitivities). At the other extreme we have a problem which is constantly with us, that of sample size: if we manage to collect enough data pertinent to our model, any test we use gains sufficient power eventually to reject the model against every alternative hypothesis and we conclude that no data will ever fit the model.

This is not the place for an extended discussion of the question of tests of fit. Gustafsson (1977; 1979) has written extensively on this topic in recent articles, where he raises some fundamental questions about the power of the approximate chi-square tests of fit many of us are accustomed to employing in our Rasch model programs. What concerns us here is that one of the reasons we use these approximate tests (apparently without knowledge of their statistical power) is that, despite our awareness of the existence of the more powerful tests, the operation of the former does not depend on the calculation of the complete likelihood and consequently the symmetric functions.

There is no doubt that we do know the theoretically correct path to follow. According to Andersen (1973b),

> the main result so far on conditional inference is that a uniformly most powerful unbiased (U.M.P.U.) test for a composite hypothesis can be constructed from the conditional likelihood.

In practice this test takes the form of a likelihood ratio test in which the statistic

$$-2[L_c - \sum_{g=1}^{G} L_g]$$

is distributed as chi-square on $(G-1)(k-1)$ degrees of freedom and where

(i) L_c is the log of the complete conditional likelihood as derived in (8),

(ii) L_g is the log of the conditional likelihood for the gth subset of the data, where G is chosen such that the number of observations in each subset is sufficiently large to warrant the assumption of asymptotic theory.

Although the combinatorial number C disappears, the item parameters have to be estimated for $G+1$ data sets and, of course, the conditional estimates must be used.

There is also no doubt that in many instances we are simply not in a position to apply this test, either because we are unable to calculate the symmetric functions (and hence the likelihoods) or because the sample size is so small that asymptotic theory is of dubious validity. The purpose of highlighting this problem is not to exhort researchers to drop their approximate tests of fit, but to induce a healthy scepticism and caution when using these approximate tests with the anticipation that, when the numerical analysis details are worked out, we will be able to operate the conditional tests in all circumstances.

The implementation of *exact probability tests*, on the other hand, requires no assumptions about distributional shape, parameter estimation, or large sample sizes (Fisher, 1934). As we have noted in equation (6), an exact test of fit of data to a Rasch model is theoretically possible since we have a conditional probability statement completely free of all parameters in the model. This enables us to control the model on the basis of the observed quantities alone since no parameters have to be estimated. Ideally we would calculate the probability of the observed data, given the marginals (which we know to be equal to $1/C$) and the probability of each other possible data set with the same marginals (each of which also happens to have probability of $1/C$) which tend to favour a hypothesis which is an alternative to the null one of specific objectivity. These other data sets are said to be 'more extreme'. The sum of all these probabilities would then be compared with the probability of a Type I error and conclusions about fit follow.

What prevents us going ahead as described above is the calculation of the combinatorial number, C, the number of possible $0, 1, \ldots, m$ data matrices. A direct frontal attack on this number seems futile, even though the answer is known when we do not restrict our observations to a pre-determined maximum of m; however, what Agresti recommends, in a related context, is a *sampling* of a relatively small number of the large number of possible matrices. With a high-speed computer the probability of a Type I error could be determined to any degree of accuracy. Sampling of both symmetric functions and matrices appears a possibility so that the technique might be employed for estimation as well as testing. These ideas are in their formative stages only but they do appear to offer one way out of the dilemma and will possibly pay strong dividends for someone interested in starting an investigation along these lines.

By now the reader will have been prompted to ask the question, 'Why use the corrected unconditional approach in both estimation and fit?' If it were possible to find the appropriate correction factors for parameter sets in all models, the problem of estimation would no longer be with us, even though we still see no way of getting around approximate expressions for standard errors. But this still leaves the tests of fit since Andersen's test requires the calculation of the log-symmetric functions.

Whereas we must agree with Gustafsson's (1980) exhortation: 'whenever it is judged important that goodness of fit is evaluated with sound methods, the c.m.l. approach should be used', we see no disadvantages accruing from a strategy which takes the corrected unconditional estimates (involving no calculation of symmetric functions) and using them in the conditional likelihood (involving a single calculation of the symmetric functions). Our stance is to make use of the best of all available methodology to arrive at solutions whose rigour is unquestioned. An increased awareness of the importance of fitting is certainly an encouraging sign in an area prone to ad hoc approximations. Furthermore an emphasis on questions of *person fit* is equally timely and opens up possibilities only previously hinted at (Leunbach, 1976; Wright and Stone, 1979).

CONCLUSION

My aim has been to review the central place of conditional inference in the theoretical and practical operation of a class of latent trait models which we label as Rasch models. The pedagogical stance has been one of recognition of the technically correct procedures to adopt both in parameter estimation and hypothesis testing about fit, combined with a cautious use of numerical approximations where applicable.

A number of avenues have been hinted at for the future direction of numerical analysis problems, all of which approximate the conditional algorithms. I have stressed the crucial aspects of tests of fit. In particular, I hope that those using approximate tests will temper their claims for 'good fit' with statements which acknowledge two fundamental facts: 'fit' is never fully determined by a finite set of tests; and information on the power of tests adds credibility to such claims.

REFERENCES

Agresti, A. and Wackerly, D. Some exact conditional tests of independence for *rxc* cross-classification tables. *Psychometrika,* 1977, **42**(1), 111–25.

Agresti, A., Wackerly, D., and Boyett, J. M. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika,* 1979, **44**(1), 75–83.

Andersen, E. B. The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society* (Series B), 1972, **34**(1), 42–54.

Andersen, E. B. A goodness of fit test for the Rasch model. *Psychometrika,* 1973, **38**(1), 123–40.(a)

Andersen, E. B. *Conditional inference and models for measuring.* Copenhagen: Mentalhygiejnisk Forskningsinstitut, 1973.(b)

Andersen, E. B. Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology,* 1973, **26**, 31–44.

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika,* 1977, **42**(1), 69–81.

Andersen, E. B. and Madsen, M. Estimating the parameters of the latent population distribution. *Psychometrika,* 1977, **42**(3), 357–74.

Andrich, D. Latent trait psychometric theory in the measurement and evaluation of essay writing ability. Unpublished Ph.D. dissertation, Department of Education, University of Chicago, 1973.

Andrich, D. A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology,* 1978, **31**, 84–98.(a)

Andrich, D. A rating formulation for ordered response categories. *Psychometrika,* 1978, **43**(4), 561–73.(b)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Borchsenius, K. A group-theoretical formulation of the concept of objectivity. Unpublished paper, 1974.

Borchsenius, K. On specific objectivity and its role in elementary physics. Unpublished paper, 1977.

Douglas, G. A. Conditional maximum-likelihood estimation for a multiplicative binomial response model. *British Journal of Mathematical and Statistical Psychology,* 1978, **31**, 73–83.

Fischer, G. H. Some probabilistic models for measuring change. In D. de Gruitjer and L. van der Kamp (Eds), *Advances in psychological and educational measurement.* London: Wiley, 1976.

Fischer, G. H. Linear logistic test models. In H. Spada and W. F. Kempf (Eds), *Structural models of thinking and learning.* Bern: Huber, 1977.

Fisher, R. A. *Statistical methods for research workers.* (5th ed.). Edinburgh: Oliver & Boyd, 1934.

Gustafsson, J-E. *The Rasch model for dichotomous items: Theory, applications and a computer program.* (Report No. 63). Göteborg: University of Göteborg, Institute of Education, 1977.

Gustafsson, J-E. Testing and obtaining fit of data to the Rasch model. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Gustafsson, J-E. A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, 1980, **40**, 377-85.

Kempf, W. F. Dynamic models for the measurement of 'traits' in social behaviour. In W. F. Kempf and B. H. Repp (Eds), *Some mathematical models for social psychology.* Bern: Huber, 1976.

Leunbach, G. *A probabilistic measurement model for assessing whether two tests measure the same personal factor.* (Report No. 19). Copenhagen: Danish Institute of Educational Research, 1976.

Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, 1932, **140**, 1-54.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, **1**, 477-82.

Malone, J. et al. Measuring problem solving ability. *National Council of Teachers of Mathematics Yearbook*, 1980.

Neyman, J. and Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrika*, 1948, **16**, 1-32.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut, 1960.

Rasch, G. A mathematical theory of objectivity and its consequences for model construction. Paper read at the European Meeting on Statistics, Econometrics and Management Science at Amsterdam, 1968.

Rasch, G. On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 1977, **14**, 58-94.

Scheiblechner, H. The separation of individual and system influences on behaviour in social contexts. *Acta Psychologica*, 1971, **35**, 442-60.

Scheiblechner, H. Psychological models based on conditional inference. In H. Spada and W. F. Kempf (Eds), *Structural models of thinking and learning.* Bern: Huber, 1977.

Whitely, S. E. Models, meanings and misunderstandings: Some issues on applying Rasch's model. *Journal of Educational Measurement*, 1977, **14**(3), 227-35.

Whitely, S. E. and Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, **11**, 163-78.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Misunderstanding the Rasch model. *Journal of Educational Measurement*, 1977, **14**(3), 219-25.

Wright, B. D. and Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 1977, 37(3), 573–86.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, **29**, 23–48.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement*. Chicago: MESA Press, 1979.

REACTANT STATEMENT

Alan G. Smith

Dr Douglas's paper is a valuable one on several counts. Firstly, it makes a contribution towards a generalization of the model, in bringing together Rasch developments in several areas; and, in doing so, the paper addresses the problems of goodness of fit tests and the power of such tests. Secondly, the paper reminds us of some essential features for 'specific objectivity' in the measurement of attributes and, in particular, reminds us why we should restrict ourselves to only one item parameter. Thirdly, the paper reminds us of the real complexity of modelling psychological concepts such that we have both mathematical completeness and at the same time workable formulae to apply to real data. I wish to take up the last point briefly, and to make one or two other comments.

Despite the great theoretical attractions of the procedure, Dr Douglas has not overstated the difficulties inherent in utilizing symmetric functions with the model. This raises for me the general issue of the potential gulf between theoretical development and real-world applications of a statistic. The Norton studies of the behaviour of the F statistic (Lindquist, 1953) many years ago showed that that statistic often provides good information under conditions where it could be expected to fail; again, colleagues will be very familiar with the robustness of the Pearson product-moment coefficient with data which often greatly abuse its assumptions. Similarly, in the case of the Rasch model applied to binary item analysis as put forward by Wright and Panchapakesan (1969), we find that several of the theoretical problems with the model may not be as significant as Dr Douglas's paper would suggest. Evidence is accumulating that the assumption of uniform item discrimination is not nearly as vital to the performance of the model as writers such as Whitely and Dawis (1974) would have us believe (Dinero and Haertal, 1977; Smith, 1978). It would appear, too, that practical use of the model in person-ability estimation is not significantly affected by the results of goodness of fit tests (Smith, 1978). Again, we find that the model works well with quite small samples (Tinsley and Dawis, 1975) even though it uses a very large sample statistic (Whitely and Dawis, 1974). Finally, although Wright and Douglas (1977) show that the 1969 Wright and Panchapakesan maximum likelihood procedures are biased, they also show that the extent of the bias is of minimal practical importance, especially when compared with the alternative estimation procedure. Thus the merits of new computationally complex procedures which solve theoretical problems of goodness of fit and so on, albeit rather nicely, will have to be very clearly demonstrated; this is especially true given that it has taken 20 years for the Rasch model to reach its current state of limited acceptance.

The work of developing the model and spreading its good measurement news is also hampered by problems of terminology and communication. I am not a mathematician and am uncertain of the correctness of my understanding on this point, but there seems to be a real problem in the use of 'conditional' versus 'unconditional' estimation procedures here. Douglas's use of the term, which is fully explained in the 1977 paper with Wright, is quite different from the earlier use of these terms in the latent trait field. The originators of the terms were Bock and Lieberman (1970) and, if I understand their terms, I would conclude with Baker (1977) and Subkoviak and Baker (1977) that Wright and Panchapakesan's 1969 procedure is conditional, not unconditional, and so we have a terminology problem wherein one might say Douglas has presented us with an unconditional generic model rather than a conditional one.

Although Douglas covers himself well when he says he would not wish to restrict work on other probabilistic models, it is difficult perhaps to see attributes elsewhere from the standpoint of Rasch assumptions. Now it is true that the two-parameter normal ogive item analysis model does not demonstrate 'specific objectivity' as defined mathematically by Rasch, and it would appear to be true that invariance of parameters does not exist for the two-parameter model (Smith, 1975; Baker, 1977). This does not mean it is not useful, nor that it cannot be made to work. Douglas says that in the two-parameter case, we must retain people's original response data in order to know something about their ability, and that is true; but I think we *can* discover more than the limited raw score to which Douglas suggests we are limited. I have been doing some work with the normal ogive model involving iterative conditional estimation of the two-item parameters in the first step, and person-ability in the second step, and there is evidence that useful results emerge. It must also be said that it is a complex expensive process which does not compare well with the Rasch logistic model. There is the further point, however, that the normal ogive model offers through the use of the normal curve a link with psychological theory which is most attractive.

I have two other brief comments. The first pertains to the notion of person-fit to the model. I must confess to some abhorrence of this concept, depending on how it comes to be used. Our main purpose in educational measurement is to be able to make definitive statements about relative person-ability. While the concept of person-fit is statistically nice, given model parameters, people are paramount and the model must accommodate them, within populations. Therein is the problem: the definition of populations must be broad rather than narrow, and researchers must be careful about conclusions drawn from significant person-fit tests. The second comment is related to the first in the notion of person-ability. Classical test theory is substantially concerned with unidimensional abilities, and we spend much of our time establishing reliable sub-scales in major tests. The fact is, of course, that latent trait models are similarly concerned with unidimensional abilities, notwithstanding their other merits. One of the major practical problems which remains with latent trait models is to show how person-ability values derived from several test scales can be related and treated, and whether the models can be made to work with tests which measure complex abilities. Wide use of the Rasch model, for example, will depend on such attributes being clearly demonstrated, and fortunately evidence (e.g. Smith, 1975) is encouraging in this regard.

In conclusion, while I have been provocative about several aspects of Dr Douglas's paper, one does need to note his comments about requirements for Rasch goodness of fit tests and their power. His paper will no doubt prove to be constructive in the further development of the model.

REFERENCES

Baker, F. B. Advances in item analysis. *Review of Educational Research*, 1977, **47**(1), 151-78.

Bock, R. D. and Lieberman, M. Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 1970, **35**(2), 179-97.

Dinero, T. E. and Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1977, **1**(4), 581-92.

Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

Smith, A. G. A study of some invariant item parameters applied to item banking. Unpublished Ph.D. thesis, University of New England, Armidale, NSW, 1975.

Smith, A. G. The stability of Rasch item easiness indices across groups for a teacher-made test. Paper presented at the Australian Association for Research in Education Conference, Perth, 1978.

Subkoviak, N. J. and Baker, F. B. Test theory. *Review of Research in Education*, 1977, **5**, 275-317.

Tinsley, H. E. A. and Dawis, R. V. An investigation of the Rasch simple logistic model: Sample free item and test calibration. *Educational and Psychological Measurement*, 1975, **35**, 325-39.

Whitely, S. E. and Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, **11**(2), 163-78.

Wright, B. D. and Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 1977, **37**, 573-86.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, **29**, 23-48.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

7

The Use of Latent Trait Models in the Development and Analysis of Classroom Tests

John F. Izard and John D. White

INTRODUCTION

Teachers use tests for a number of different but related purposes. These include assessing the success of a relatively short sequence of instruction, indicating how much knowledge or skill has been retained over a period of time, describing or summarizing achievement over an extended period of study, and diagnosing aspects of curriculum which need further instruction. Obviously, if curricula vary, then the supply of tests which mirror each curriculum presents problems.

When using published tests to assess progress through an instructional sequence, teachers may be concerned that some of the questions are of limited value because the content differs from the material presented in their own classes, or that certain important objectives have been given little consideration in the test specification. Such concerns may result in teachers rejecting the use of published tests, making do with inadequate data, or using other questions in an unsystematic way in an attempt to meet deficiencies in the published tests.

In order to meet these concerns, some teachers have been using collections of test questions such as the *Australian Item Bank* (Year 10 Mathematics, Science, and Social Science) and the *New Zealand Item Bank: Mathematics* (Levels 2-7 Mathematics). However, selection of questions on the basis of content alone and without consideration of other characteristics such as difficulty and discrimination makes the interpretation of the results obtained from such questions difficult and of doubtful validity.

Since variations in curricula place different emphases on different objectives, it seems desirable to produce collections of questions for each objective so that tests can be tailored to suit local needs. Provision of such collections will not represent any advance on the existing item banks unless difficulty data are supplied with the items and unless procedures are developed to adjust for difficulty when interpreting scores. Such procedures will need to be easy to apply without reference to computing centres and will need to allow for additional questions to be added to the bank where necessary. In Australia, data collected in the development of each item bank are not readily available to users (except in Tasmania). This may lead to a situation where two sets of five questions may be used and the achievement level represented by a score of four or more on one test may be totally different from that for the same score on the other test.

Test analysis techniques based on the Rasch model (Rasch, 1960; Wright and Stone, 1979) attempt a separation of the respective contributions of person ability and item difficulty to a score. When this latent trait model is assumed to be appropriate, item difficulty information available for each question selected for a test may be used to make ability estimates for various scores on that test.

This paper describes the development of a pool of calibrated items for use by teachers and then presents a number of simplified procedures which may be applied by teachers to construct tests from the item pool, to interpret the results, and to calibrate further items devised by teachers.

DEVELOPMENT OF AN ITEM POOL

A solution to the problem of providing appropriate testing procedures for schools seems to lie in the use of item banking techniques. This is a long-term solution and may take the next decade to be implemented as a working assessment procedure at the school level. An intermediate solution lies in the development of a pool of test items and the production of progress and review tests from this item pool. The essential feature of these tests is that they relate to well-defined objectives and are used to determine whether or not these objectives have been met or the extent to which they have been met.

We do not make any assumption about the sequence in which skills are taught or the curriculum in which the skills are embedded. We do assume that skills can be taught and or learnt. We do assume that a teacher knows what he wants learnt by his students and that he has developed a sequence of learning experiences to enable the skill to be acquired. In other words, we assume that there are some identifiable skills which can be taught, learnt, and tested as part of instructional programs utilized by teachers. In the discussion which follows, we distinguish between progress and review tests and then describe the development of an item pool using *addition of whole numbers* as the specific topic.

A progress test is a small collection of items measuring performance on a specific skill such that a score on this test reflects the mastery status of a student relevant to the skill. For example, one progress test may have a sample of items which involve adding two 2-digit numbers without regrouping (carrying), while another progress test may involve adding two 2-digit numbers with regrouping (carrying) from the units.

A review test is a collection of items measuring a student's performance on a number of skills related by content such that a score on this test identifies areas of strength and weakness possessed by the student in the specified skill areas. For example, a review test may have items involving adding two, three, and four 2-digit numbers without and with regrouping (carrying).

In developing the item pool for the addition of whole numbers, the objectives were first discussed with teachers from several education departments. The collection of items was then trial tested with children in Years 3, 4, 5, and 6. The data presented in this paper are taken from responses by the Victorian children in the sample, and were analysed with Version 3 of the BICAL computer program (Wright, Mead, and Bell, 1979).

DESIGN OF PROGRESS TESTS

A teacher will have some idea of the target population for which a test is selected—it may be that teaching material for the objective has been completed recently or a placement or review test may have been administered. We also know from trial testing of the items that we obtain a number of items for each objective which cluster within a restricted range.

If a learning program based on the specific objective has been designed and implemented, there are probably reasonable expectations for the program's success. These expectations would be reflected in a narrow distribution of scores with a relatively high mean, and are represented in Figure 1.

When comparing the target population's ability distribution and the item difficulty distribution for the test or tests, it is convenient to use the terminology suggested by Wright and Douglas (1975). The average difficulty of the items selected for the test is referred to as the height of the test, *H*. The range of item difficulties is the test width, *W*, and the length of the test is the number of items, *L*. Where these are estimated from samples, lower case letters are used.

The best overall test is the uniform test (Wright and Stone, 1979, p. 134) in which items are evenly spaced from easiest to hardest. This test is appropriate for any target population within the usable range of the test.

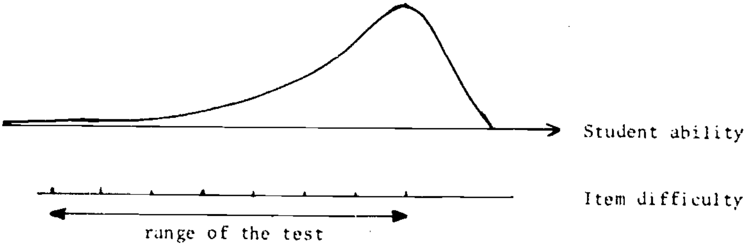


Figure 1 Expected Pattern of Scores after Completion of a Successful Learning Program

The test is described in terms of H, W, and L and is designed with the object of minimizing the standard error of measurement (SEM) subject to certain constraints.

Consider an n-item uniform test as shown in Figure 2 where n is odd.

In the case illustrated in Figure 2, the length of the test is n (5 in this example), the width of the test is $(n - 1)d$, and the height of the test is 0.

Figure 3 shows the corresponding information for an n-item uniform test where n is even. In this example, the length of the test is n (4 in this example), the width of the test is $(n - 1)d$, and the height of the test is 0.

If a student has a raw score of r on an n-item test then b_r, the ability estimate of the student, is related to r by the equation

$$r = \sum_{i=1}^{n} \frac{e^{b_r - d_i}}{1 + e^{b_r - d_i}}$$

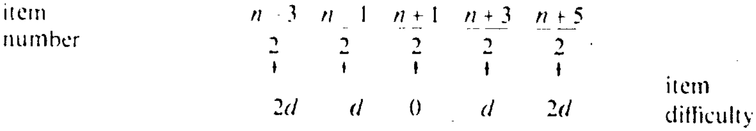


Figure 2 Uniform Test with Odd Number of Items

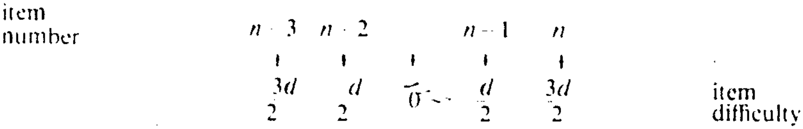


Figure 3 Uniform Test with Even Number of Items

Table 1 Limiting Value for Ability Estimates

			$\lim_{d\to 0} b_r$			
n	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 6$
3	0.693	0.693				
4	1.099	0.000	1.099			
5	1.386	0.405	0.405	1.386		
6	1.609	0.693	0.000	0.693	1.609	
7	1.792	0.916	-0.288	0.288	0.916	1.792

As $d \rightarrow 0$ the uniform test becomes more narrow and $d_i \rightarrow 0$

then
$$r \rightarrow \frac{ne^{b_r}}{1 + e^{b_r}}$$

and $b_r = \ln \frac{r}{n - r}$, as shown for various values of n and r in Table 1.

The standard error of b_r is given by

$$\frac{1}{\sqrt{\sum_{i=1}^{n} \left[\frac{e^{b_r - d_i}}{1 + e^{b_r - d_i}}\right] \left[1 - \frac{e^{b_r - d_i}}{1 + e^{b_r - d_i}}\right]}}$$

As $d \rightarrow 0$, $d_i \rightarrow 0$
and the standard error $s_r \rightarrow$

$$\frac{1}{\sqrt{\sum_{i=1}^{n} \left[\frac{e^{b_r}}{1 + e^{b_r}}\right] \cdot \left[1 - \frac{e^{b_r}}{1 + e^{b_r}}\right]}}$$

that is, $\lim_{d \rightarrow 0} s_r = \sqrt{\frac{n}{r(n - r)}}$, as shown for various values of n and r in Table 2.

Figure 4 shows the ability estimate obtained for raw scores on tests of various lengths; each ability estimate is shown with one standard error bounding either side of the estimate.

If the Rasch model is appropriate, we can specify a probability that a person with a given ability will get a question correct. Similarly we can

Table 2 Standard Errors for Narrow Tests as $d \to 0$

				$\lim_{d \to 0} s_e$		
n	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 6$
3	1.22	1.22				
4	1.15	1.00	1.15			
5	1.12	0.91	0.91	1.12		
6	1.10	0.87	0.82	0.87	1.10	
7	1.08	0.84	0.76	0.76	0.84	1.08

estimate people's ability from the score they receive on a set of questions reflecting a continuum. For a uniform narrow test, this ability estimate can be expressed in terms of the number of standard errors above the mean difficulty of the test. For example from Tables 1 and 2, the ability estimate for a raw score of 4 on a test of length $L = 5$ (where $d \to 0$) is 1.39 with a standard error of 1.12. This estimate is 1.24 standard errors above the mean. By making some normal curve assumptions, we can infer that there is a probability of 0.89 that the student has an ability greater than the test mean. This inference may be used as a definition of mastery for a

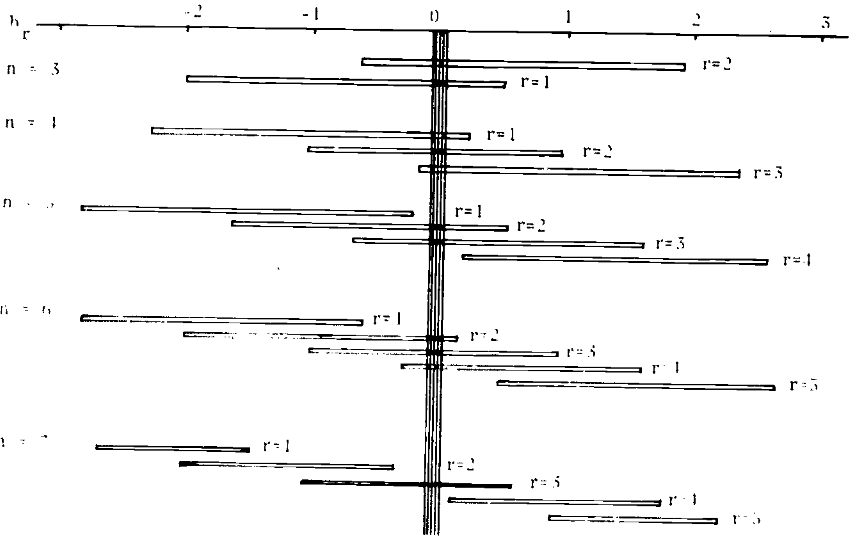


Figure 4 Ability Estimates for Various Raw Scores on a Number of Uniform Narrow Tests

narrow uniform test. In other words, a score corresponding to an ability sufficiently higher than the mean may be regarded as evidence that items from the same domain will be answered successfully.

Our analysis shows that for tests of length le.. than 5 the raw scores other than 0 or n give insufficient indication of mastery or non-mastery as defined. For a narrow uniform test of length 5 we have mastery for $r = 5,4$, non-mastery for $r = 0,1$ and there is no indication of the respondent's mastery status for $r = 2,3$.

If a test of length 5 is considered satisfactory, then tests with $L = 6$ or 7 might be considered wasteful. Given the constraint that only discrete scores are possible, only limited additional information is available when the test is lengthened by one or two questions.

We can now look for values of d for which the uniform test can be described as narrow. In the uniform test design for this section, test width

$$w = (n-1)d, \quad f = \frac{r}{n}$$

where r is the raw score, and ability estimates for given f and w are obtained from the UFORM procedure (Wright and Stone, 1979, p. 144).

That is,
$$b_r = w(f - 0.5) + \ln \frac{1 - e^{-wf}}{1 - e^{-w(1-f)}}$$

Hence

$$b_r = d(n-1)\left(\frac{r}{n} - 0.5\right) + \ln \frac{1 - e^{\frac{-(n-1)dr}{n}}}{1 - e^{\frac{-d(n-1)(n-r)}{n}}}$$

is an ability estimate for $d > 0$.

For example, given that $n = 5$ and $r = 1$, $b_{r(d=0)} = -1.39$, and d is found so that the discrepancy between b_r and $b_{r(d=0)}$ given by $|b_r - b_{r(d=0)}|$ is less than ϵ, where ϵ is a designated accuracy value.

That is,
$$\left|1.2d + \ln \frac{1 - e^{-0.8d}}{1 - e^{-3.2d}} - (-1.39)\right| < \epsilon.$$

Table 3 shows this discrepancy or error term $f(d)$ for $n = 5$, $r = 1$ when $d \neq 0$. Similar values for $f(d)$ will be obtained when $n = 5$ and $r = 4$.

With $d = 0.3$, $n = 5$, $w = 1.2$, the ability estimate changes in magnitude by 0.03 from the ability estimate derived from a narrow test.

This represents a 2.2 per cent change in the ability estimate or 2.7 per

Table 3 Values of the Error Term $f(d)$ for Values of d when $n = 5$, $r = 1$

d	$f(d)$
0.1	0.0003
0.2	0.012
0.3	0.032
0.4	0.059

cent of the standard error. Similarly, with $d = 0.4$ the percentage change is 4.2 per cent or 5.3 per cent of the standard error for a test of zero width.

If we have a cluster of items across a range of about 1.5 logits, then we can select five of these items to construct a narrow uniform test. If these items are specific to an objective then a progress test is constructed with the following properties: (a) a student has mastered the skill if he scores 5 or 4; (b) he has not mastered the skill if he scores 0 or 1; (c) his mastery status is not determined for scores of 2 or 3.

We will summarize the design so far by looking at data from an actual test which consists of 20 items constructed to assess the objective of the addition of two 2-digit addends with a sum less than 100 and regrouping (carrying) from units to tens.

Figure 5 shows the difficulty estimates obtained for each item when these 20 items were calibrated with other addition items. The mean item

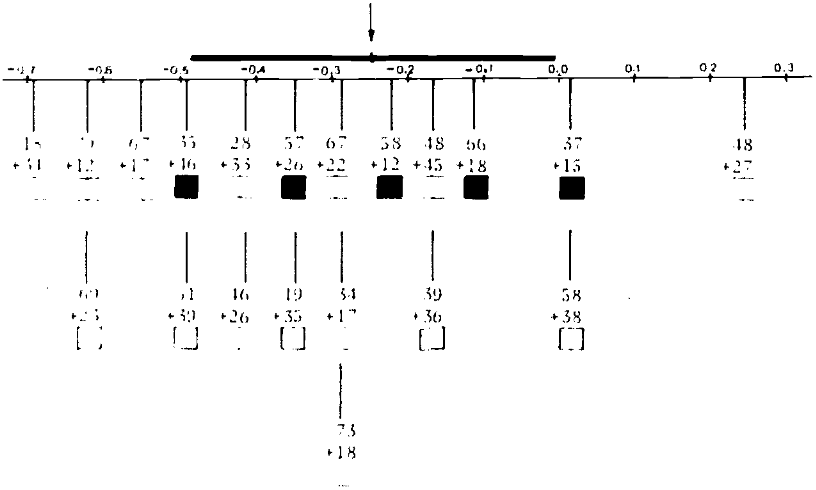


Figure 5 Item Difficulties for each Item on One Addition Test

difficulty for all 20 items is shown by I and the magnitude of the standard deviation of the difficulty estimates is shown as the dark line.

We can construct a number of progress tests by selecting items from the pool, two of which are:

Progress Test A:	35	57	58	66	37
(Coded ■)	+ 46	+ 26	+ 12	+ 18	+ 15
difficulty:	−0.47	−0.35	−0.23	−0.12	0.03
Progress Test B:	69	51	19	39	58
(Coded)	+ 23	+ 39	+ 35	+ 36	+ 38
difficulty:	−0.61	−0.47	−0.35	−0.17	0.03

For Progress Test A

$$h = \frac{\Sigma d_i}{L} = -0.23, \text{ and}$$

$$w = 3.5 \sqrt{(\Sigma d_i^2 - Lh^2)/(L-1)} = 0.68$$

Using Table 1, the ability estimates for $r = 1, 2, 3$, and 4 are:

r	Ability estimate
1	$-0.23 - 1.39 = -1.62$
2	$-0.23 - 0.41 = -0.64$
3	$-0.23 + 0.41 = 0.18$
4	$-0.23 + 1.39 = 1.16$

with standard errors of 1.12, 0.91, 0.91, and 1.12 respectively, using Table 2.

In the progress test model, these values are not so important because the user of the progress test is interested in the mastery status determined from the raw score.

The user knows that $r = 0, 1$ corresponds to non-mastery requiring further teaching, $r = 4,5$ corresponds to mastery, and $r = 2,3$ will require further testing to confirm mastery status.

However, for the purposes of this discussion, the ability associated with a raw score may be estimated by the UFORM procedure (Wright and Stone, 1979, p. 144).

This ability estimate is given by the equation

$$b_r = h + w(f - 0.5) + \ln \frac{1 - e^{-wf}}{1 - e^{-w(1-f)}}$$

Table 4 Comparison of Estimates of Ability Using UFORM and Narrow Uniform Test Assumptions

	Approximate estimate		UFORM estimate	
f	b_f	s_f	b_f	s_f
0.2	1.62	1.12	1.63	1.12
0.4	0.64	0.91	0.64	0.92
0.6	0.18	0.91	0.18	0.92
0.8	1.16	1.12	1.17	1.12

with standard error given by

$$s_f = \left\{ \frac{w}{L} \frac{[1-e^{-w}]}{[1-e^{-wf}][1-e^{-w(1-f)}]} \right\}$$

A comparison of the abilit**y** estimates with the approximations resulting from narrow uniform test assumptions is presented in Table 4.

Figure 6 illustrates the mastery sta.us associated with a raw score of 4 and the non-mastery status for a raw score of 1. It also illustrates that there is insufficient information to determine the mastery status for $r = 2$ or $r = 3$.

For a narrow uniform test we assume that all items have equal difficulty. Where $L = 5$, and $h = -0.23$, the ability estimate for $r = 4$ is $b_4 = 1.17$. Hence

$$p(x = 1, b_4 = 1.17, d = -0.23) = \frac{e^{1.40}}{1 + e^{1.40}} = 0.802$$

is the probability of a success on an item encounter for a student with ability 1.17. Similarly,

$$p(x = 0, b_4 = 1.17, d = -0.23) = \frac{1}{1 + e^{1.40}} = 0.198$$

is the probability of a failure on an item encounter for a student with ability 1.17. The probabilities of the student obtaining various raw scores given that his ability is 1.17 are

$$\begin{aligned}
p(r = 5) &= (0.802)^5 &&= 0.332\\
p(r = 4) &= 5(0.802)^4(0.198) &&= 0.410\\
p(r = 3) &= 10(0.802)^3(0.198)^2 &&= 0.202\\
p(r = 2) &= 10(0.802)^2(0.198)^3 &&= 0.050\\
p(r = 1) &= 5(0.802)\ (0.198)^4 &&= 0.006\\
p(r = 0) &= (0.198)^5 &&= 0.0003
\end{aligned}$$

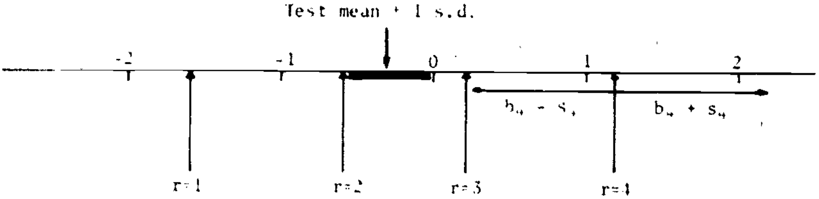


Figure 6 Ability Estimates from Raw Score in Relation to Progress Test A

When a student has an ability of 1.17, the maximum probability of making an incorrect statement of his mastery status is therefore

$$0.202 + 0.050 + 0.006 + 0.0003 = 0.258$$

This estimate is inflated due to the inclusion of the cases where $r = 3$ and r 2.

DESIGN OF REVIEW TESTS

In the case of the progress tests, the items for each test had a relatively lower range of difficulty. However if review tests are to be construeted by selecting items from the various progress test item pools, the difficulty continuum for addition items ranges from −4 to +4. It is possible to design several review tests to span the continuum as shown in Figure 7.

We can consider one such test where $h = -2.5$, $w = 3.0$, and L is to be determined.

For the progress tests we were able to use the characteristics of a narrow test to determine the value of L.

In the review tests we cannot assume 'narrowness' and will have to determine L from an assumption about the magnitude of the standard error of measurement, using a method proposed by Wright and Stone (1979, p. 140) in which

$$L = \frac{C_{fw}}{\text{SEM}^2},$$

where C_{fw} is termed the error coefficient, SEM the standard error of measurement and f the expected relative score. They define the error coefficient as

$$C_{fw} = \frac{w[1 - e^{-w}]}{[1 - e^{-fw}][1 - e^{-(1-f)w}]}$$

For the design above we obtain

$$C_{fw} = \frac{3\ [1 - e^{-3}]}{[1 - e^{-3f}][1 - e^{-(1-f)3}]}$$

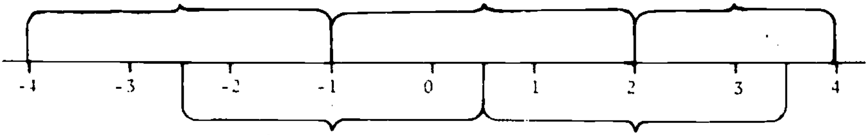


Figure 7 Ability Ranges to be Covered by Review Tests

Table 5 Values of C_{fw} for $w = 3$ and $f = 0.1(0.1)0.9$

f	C_{fw}
0.1	11.8
0.2	6.9
0.3	5.5
0.4	4.9
0.5	4.7
0.6	4.9
0.7	5.5
0.8	6.9
0.9	11.8

Table 6 Test Length Associated with Values of SEM and C_{fw} when $w = 3$

				C_{fw}		
SEM	SEM^2	11.8	6.9	5.5	4.9	4.7
0.1	0.01	1180	690	550	490	470
0.2	0.04	295	173	138	123	118
0.3	0.09	131	77	61	54	52
0.4	0.16	74	43	34	31	29
0.5	0.25	47	28	22	20	19
0.6	0.36	33	19	15	14	13
0.7	0.49	24	14	11	10	10
0.8	0.64	18	11	9	8	7
0.9	0.81	15	9	7	6	6
1.0	1.00	12	7	6	5	5
1.1	1.21	10	6	5	4	4
1.2	1.44	8	5	4	3	3
1.3	1.69	7	4	3	3	3
1.4	1.96	6	4	3	3	2
1.5	2.25	5	3	2	2	2

Table 7 Item Difficulties Generated for Review Test 1

i	δ_i	i	δ_i
1	−3.9	9	−2.3
2	−3.7	10	−2.1
3	−3.5	11	−1.9
4	−3.3	12	−1.7
5	−3.1	13	−1.5
6	−2.9	14	−1.3
7	−2.7	15	−1.1
8	−2.5		

Values for C_{fw} have been tabulated for various values of f as shown in Table 5, and the corresponding test lengths for various values of SEM are shown in Table 6.

A test length of $L = 15$ would not be an unreasonable length for a review test in terms of the time for administration. If we look at $L = 15$ in the body of Table 6 we see that SEM ranges from 0.9 (for $f = 0.1$, and $f = 0.9$) to $\text{SEM} \simeq 0.6$.

The formula $\delta_i = H - (w/2L)(L + 1 - 2i)$ for $i = 1,15$ generates the preferred item difficulties for the 15-item test, as shown in Table 7.

We now have to decide on the items to include in a review test. Table 8 shows the overall review test structure which could completely span the addition continuum by including items from a number of objectives pools. The selection is shown in Table 9; the total deviation from desired difficulties is 0.00.

We are now in a position to calculate the characteristics of review test 1.

Test height is estimated by:

$$h = \frac{\Sigma d_i}{L} = -2.5$$

Table 8 Review Test Design to Cover Addition Continuum

Review Test	Objectives (codes)
1	31, 32, 33
2	33, 34, 35
3	35, 36, 37
4	37, 38, 39
5	39, 40, 41
6	41, 42, 43

Test width is estimated by:

$$w = 3.5\sqrt{(\Sigma d_i^2 - Lh^2)/(L-1)}$$
$$w = 3.13$$

Ability estimates for each raw score are given by:

$$b_f = h + w(f-0.5) + \ln \frac{1-e^{-wf}}{1-e^{-(1-f)w}}$$
$$= -2.50 + 3.13(f-0.5) + \ln \frac{1-e^{-3.13f}}{1-e^{-(1-f)3.13}}$$

with standard error

$$s_f = \left\{ \frac{3.13}{15} \frac{1-e^{-3.13}}{[1-e^{-3.13f}][1-e^{-(1-f)3.13}]} \right\}^{1/2} = \left\{ \frac{0.2}{[1-e^{3.13f}][1-e^{-(1-f)3.13}]} \right\}^{1/2}$$

Table 10 presents the ability estimates and standard errors for various scores on this review test.

If this review test is being used as an instrument to ascertain the position of a student relevant to the objectives after an extended period of instruction related to the objectives, then the student's score can suggest which objectives have been mastered provided that the average difficulty of each objective is known. Further we can argue that:

(i) $r < 3$ indicates that a less difficult review test is necessary;

(ii) $3 \le r \le 12$ indicates that student's ability is in the range of the objectives;

Table 9 Item Selection for Review Test 1

Item number	δ_i Desired difficulty	d_i Selected difficulty	$\delta_i - d_i$	Item number and objective
1	−3.9	−3.85	−0.05	(9, 31)
2	−3.7	−3.66	−0.04	(2, 31)
3	−3.5	−3.48	−0.02	(14, 31)
4	−3.3	−3.33	0.03	(8, 31)
5	−3.1	−3.06	−0.04	(2, 33)
6	−2.9	−2.94	0.04	(5, 33)
7	−2.7	−2.82	0.12	(3, 31)
8	−2.5	−2.52	0.02	(7, 33)
9	−2.3	−2.27	−0.03	(8, 33)
10	−2.1	−2.12	0.02	(7, 32)
11	−1.9	−1.91	0.01	(11, 33)
12	−1.7	−1.72	0.02	(17, 32)
13	−1.5	−1.49	−0.01	(6, 33)
14	−1.3	−1.33	0.03	(13, 32)
15	−1.1	−1.00	−0.10	(18, 32)

Table 10 Ability Estimates and Standard Errors for Raw Score Values on Review Test 1

Raw score r	$f \; \frac{r}{L}$	b_r	s_r
1	0.067	- 5.47	1.06
2	0.133	4.65	0.79
3	0.200	- 4.12	0.68
4	0.267	−3.69	0.63
5	0.333	−3.32	0.59
6	0.400	−2.98	0.57
7	0.466	- 2.66	0.57
8	0.533	- 2.34	0.57
9	0.600	- 2.02	0.57
10	0.667	- 1.68	0.59
11	0.733	−1.31	0.63
12	0.800	- 0.88	0.68
13	0.866	−0.35	0.79
14	0.933	+0.47	1.06

(iii) $r > 12$ suggests a more difficult review test relating to other objectives because the student's ability is beyond the range of these objectives.

USING AN ITEM BANK OF CALIBRATED ITEMS

Once an item bank or pool is established, we can make the questions and associated data available to teachers. However, the majority of teachers in Australian schools do not have ready access to computers and it is necessary to provide simplified procedures which will not need sophisticated computing facilities. By contrast, hand-held calculators are widespread and small programmable calculators are becoming more common. Our experience in lecturing to teacher trainees and graduate teachers indicates that worksheets can assist teachers to collate information and to produce relevant statistics. Accordingly we sought to develop worksheets which would enable teachers to use information from an item bank to construct tests with known characteristics, to check that their group of students perform on such tests in a manner consistent with the performance of the reference group used to set up the item bank, and to scale their own items to the continuum underlying the item bank.

Where the items in an item bank have been scaled on a single continuum, a teacher may construct either a test for relatively precise measurement in a particular part of the continuum or a broader test which will provide estimates of the range of achievement in that classroom. If desired, both types of test could be used. The results on the wide range test could suggest the types of item which might be tested in more detail.

The UFORM procedure referred to earlier in this paper may be used to estimate the ability associated with each raw score on a test using the item bank data for each of the items in the bank. (Ability estimates cannot be made for zero or a perfect score.) This procedure assumes that the items in the bank are calibrated on a single continuum and it is recommended that items are uniformly spaced in difficulty for the intended target population (Wright and Douglas, 1975).

The difficulty of the items selected for the bank is averaged to estimate the test height, and the variance of the item difficulties is used to estimate the width of the test. The estimated ability is

$$b_r = h + w(f - 0.5) + \ln(A/B)$$

where h is the mean difficulty of the items,
w is the estimated test width
f is the proportion of the items correct
A is $1 - \exp(-wf)$, and
B is $1 - \exp[-w(1-f)]$.

The associated standard error is

$$s_r = [(w/L)(C/AB)]$$

where L is the length of the test and
C is $1 - \exp(-w)$.

The worksheet for this task (see Appendix I) is used with a calculator having 'e^x' and 'ln' function keys.

Table 11 shows the results obtained using the worksheet for a test of five items from an item b. .ik with difficulties −0.560, −0.174, +0.012, +0.197, and +0.573.

If required, the ability estimates may be recalculated for the six-item test which results when an item of difficulty +0.975 is added to the five-item test. Table 12 shows the corresponding results.

Instead of using the worksheet, we can obtain this table of ability estimates and associated standard errors from convenient tables presented in Wright and Stone (1979, p. 146). For the example shown above, the estimates from the Wright and Stone tables are compared with the worksheet calculations (all correct to 2 decimal places) in Table 13.

These estimates from item bank data can be compared with actual observations to see whether the predictions provide useful information. However such a comparison requires a procedure to calibrate items with

Table 11 Person Ability Estimates and Associated Standard Errors for Various Scores on a Five-item Test

Raw score	Proportion correct	Ability estimate	Standard error
r	f	b_r	s_r
1	0.2	1.430	1.134
2	0.4	- 0.414	0.932
3	0.6	0.433	0.932
4	0.8	1.450	1.134

Table 12 Person Ability Estimates and Associated Standard Errors for Various Scores on a Six-item Test

Raw score	Proportion correct	Ability estimate	Standard error
r	f	b_r	s_r
1	0.17	- 1.538	1.118
2	0.33	- 0.572	0.894
3	0.50	0.170	0.846
4	0.67	0.913	0.894
5	0.83	1.879	1.118

Table 13 Ability Estimates and Associated Standard Errors for Various Scores

	Worksheet calculations		Wright and Stone[a]	
Score	b_r	s_r	b_r	s_r
1	1.43	1.13	− 1.49	1.16
2	0.41	0.93	− 0.39	0.94
3	0.43	0.93	0.41	0.94
4	1.45	1.13	1.51	1.16

[a] Wright and Stone, 1979, Tables 7.3.1, 7.3.2, p. 146.

another group as well as a procedure to check whether both the original group and the new group react to the items in a consistent fashion. Both types of procedure are now described.

CALIBRATION OF ITEMS USING PROX

Wright and Stone (1979) describe a procedure called PROX which provides a Rasch calibration of test items and 'approximates the results obtained by more elaborate and hence more accurate procedures extremely well' (Wright and Stone, 1979, p. 28).

This procedure assumes that item difficulties and person abilities are more or less normally distributed. Item difficulty is estimated with reasonable accuracy (when compared with more sophisticated computing procedures) where both person abilities and item difficulties are more or less symmetrically distributed around one mode, and the location and spread of person abilities and the item difficulties are similar.

The procedure requires the responses to be listed in a student-by-item matrix and the calculations are carried out on marginal totals of correct and incorrect counts for both items and students. (If there are any students with perfect or zero scores and any items with perfect or zero success rates, these are deleted from the matrix before the calculations.) This listing of student responses may present the classroom teachers with a sizable clerical chore. However a class analysis chart after the style of the CAIM material (ACER, 1976; 1979) allows the teacher to avoid this clerical work by transferring the actual student responses to a chart.

Using PROX it is possible to calibrate items using a hand calculator and paper and pencil. Wright and Stone point out that the PROX procedure has an application in the classroom but, if classroom teachers are to use such a procedure, it is our view that further assistance needs to be provided. This assistance is provided in the form of a worksheet (see Appendix II) which uses marginal totals from a class analysis chart, and an example of its use is presented in Table 14. When the procedure is applied to these data, the results shown in Tables 15 and 16 are obtained.

CALIBRATING ITEMS ON TO THE ITEM BANK SCALE

The procedure for calibrating teacher-made items on to the same continuum defined by an item bank requires that items constructed by the teacher be administered to a group of students together with a set of items from the item bank. The items from the item bank constitute the link, and the quality of this link can be investigated using the procedures advocated by Wright and Stone (1979, p. 96-116). The quality control of the link enables the original results obtained by the reference group (which provided the data on the banked items) to be compared with the results obtained from the sampled items test. We would expect that the observed difficulties for the link items would differ from the difficulties of those items obtained for the reference group to the extent that the group of persons being tested is more or less able than the reference group. After adjusting for the group difference in ability, the remaining discrepancies for each item are expected to have a mean of zero (Wright

Table 14 Data Matrix for Ten Persons and Six Items

Item number	Person number 1	2	3	4	5	6	7	8	9	10	Item score
1	1	1	1	1	1	1	1	0	1	0	8
2	1	1	1	0	1	0	0	0	0	1	5
3	1	1	1	0	0	1	1	0	0	0	5
4	1	0	1	1	1	0	0	1	0	0	5
5	0	1	0	0	0	0	0	0	0	0	1
6	1	0	0	1	0	0	0	0	0	0	2
Person score	5	4	4	3	3	2	2	1	1	1	$N = 10$ $L = 6$

Table 15 Item Calibrations Obtained Using PROX

Item number	Item score	$x = \ln \frac{N - s}{s}$	Initial calibration	Corrected calibration	Standard error
1	8	−1.386	1.752	−2.399	1.082
2	5	0.000	−0.366	−0.501	0.866
3	5	0.000	0.366	−0.501	0.866
4	5	0.000	0.366	−0.501	0.866
5	1	2.197	1.831	2.506	1.443
6	2	1.386	1.020	1.396	1.082

Mean = 0.366
U variance = 1.573

Table 16 Person Measures Obtained Using PROX

Person number	Person score r	Initial measure $b = \ln \frac{r}{L - r}$	Corrected measure	Standard error
1	5	1.609	2.287	1.557
2	4	0.693	0.985	1.231
3	4	0.693	0.985	1.231
4	3	0.000	0.000	1.160
5	3	0.000	0.000	1.160
6	2	0.693	−0.985	1.231
7	2	0.693	−0.985	1.231
8	1	−1.609	−2.287	1.557
9	1	1.609	−2.287	1.557
10	1	1.609	2.287	1.557

Mean = 0.322
V variance = 1.250

and Stone, 1979, p. 114). If both groups react to the test items in a consistent way, then all items are adjusted to the original scale defined by the item bank reference group data.

The worksheet (see Appendix III) devised for this task uses the data from the item bank and the data obtained from the group of students for the link items. (The latter data could be obtained using Worksheet 2 as described above.) The difference in ability between the reference group and the group of students is estimated as the mean of the differences in difficulty for each item. The magnitude of the remaining discrepancy is considered for each item separately as well as for the set of items from the bank.

If it is decided that the discrepancies are small enough to be ignored, then the calibrations of the teacher-made items are adjusted to the same extent. If the discrepancies are too large to be ignored, then it may be necessary to conclude that the group of students reacts to the questions in the item bank in a different way from the group which provided the original item bank data. Further, the differences between the two groups cannot be accounted for by a difference in ability.

In order to illustrate this procedure, results from a calibration of a 55-item mathematics test were used to construct two seven item tests. Results for these tests on another sample from the same population as that sampled for the calibration provided 'observed difficulties'. The difficulties for the reference group and the sample group are summarized for two collections of items in Table 17. The table lists the approximate chi-squared estimates associated with residual discrepancies for each item. The discrepancies are related to the particular group of seven items for which the calibration was performed. Test A illustrates an acceptable linking collection of items since the chi-squared estimate of 7.87 is less than the critical value of 14.07. Test B items do not appear to be consistent for both the reference group and the sample group (chi-squared estimate is 33.88) and therefore Test B is not a satisfactory link. In Test A, item 24 appears to be a poor item for the link; in Test B, items 13, 49, and 54 contribute most to the poor quality of the link.

This section set out to describe the procedure for calibrating teacher-made items for a particular line of inquiry onto a calibrated collection of items along the same line of inquiry. After finding suitable link items, the teacher-made items are calibrated onto the existing scale using a simple translation calculated from the difference between link item difficulties on the reference scale and the scale obtained on the link test. In developing the computational steps for this procedure, it has become clear that we must investigate further those characteristics of an item or groups of items which could indicate to us suitability for the linking process. Another area requiring further investigation comes immediately to mind

Table 17 Observed Difficulties for Two Seven-item Tests

	Item number	Item bank difficulty	Observed difficulty	Residual discrepancy	Chi-squared estimate
Test A	7	−0.928	−1.224	0.115	0.297
	16	−1.854	−2.106	0.071	0.113
	24	−0.314	−0.953	0.458	4.711
	39	1.262	1.127	−0.046	0.048
	46	0.316	0.326	−0.191	0.819
	53	0.912	0.912	−0.181	0.736
	55	1.873	1.918	−0.226	1.147
Test B	11	−0.584	−0.857	0.356	2.624
	13	−2.122	−2.791	0.752	11.711
	19	−0.954	−1.131	0.260	1.400
	38	−0.185	−0.209	0.107	0.237
	45	0.482	0.754	−0.189	0.740
	49	1.002	1.696	−0.611	7.731
	54	1.779	2.537	−0.675	9.435

when we consider, if two seven-item tests produce such different link qualities, whether they also produce different ability estimates for the group to which the items are exposed. The answer to this investigation may have implications for item banking and although it would be the substance of another paper a preliminary analysis is reported in Table 18. Table 18 shows the estimated abilities for each raw score and the cor-

Table 18 Ability Estimates and Associated Standard Errors for Two Seven-item Tests

	Raw score	Estimated from calibration data		Observed	
		b_t	s_t	b_t	s_t
Test A	1	−2.16	1.17	−1.99	1.18
	2	−1.07	0.96	−1.06	0.98
	3	0.22	0.90	−0.32	0.92
	4	0.58	0.90	0.37	0.91
	5	1.44	0.96	1.07	0.95
	6	2.53	1.17	1.96	1.16
Test B	1	−2.43	1.17	−2.19	1.24
	2	−1.34	0.96	−1.17	1.02
	3	0.48	0.90	−0.37	0.95
	4	0.32	0.90	0.38	0.96
	5	1.17	0.96	1.19	1.02
	6	2.26	1.17	2.20	1.22

responding observed abilities for the second sample on both the calibrated and observed data. The difference between ability estimates from the four sources are well within acceptable tolerances when we examine the respective standard errors. When this area is investigated completely we would expect to develop instructions for classroom use of established calibrations of items.

SUMMARY

It has been shown that a pool of items which has been calibrated onto an ability scale using Rasch analysis can be used to assist classroom assessment in two ways. The first application involves the production of progress and review tests by a test development group. The users of these tests do not necessarily have to understand the underlying theoretical structure of the tests, but they must know the simple rules to use in interpreting raw scores of students on the tests.

The second application involves the user with decision-making associated directly with the pool of items. Although many easy-to-follow worksheets were developed for calibration of items, estimation of abilities, and use of the established item pool, it is anticipated that the user would need to be aware of the assumptions and concepts of Rasch measurement if the sheets were to be used. In either case the use of Rasch analysis has been directed towards the provision and development of objective measuring instruments in which the teacher has a great deal of flexibility in choosing the individual questions that match the teaching intention.

REFERENCES

Australian Council for Educational Research. *CATIM: Class Achievement Test in Mathematics Year 6/7.* Hawthorn, Vic.: ACER, 1976.

Australian Council for Educational Research. *CATIM: Class Achievement Test in Mathematics Year 4/5.* Hawthorn, Vic.: ACER, 1979.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960.

Wright, B. D. and Douglas, G. A: *Best test design and self-tailored testing.* (Research Memorandum No. 19). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1975.

Wright, B. D., Mead, R. J., and Bell, S. R. *BICAL: Calibrating items with the Rasch model.* (Research Memorandum No. 23B). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1979.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement.* Chicago: MESA Press, 1979.

APPENDIX I

Worksheet 1: Calculation of Ability Estimates from Item Bank Data

Item code (1)	Scaled difficulty (2)	{Scaled difficulty}² (3)	r Score (4)	$f = \frac{r}{L}$ (5)	wf (6)	A $= 1 - \exp(-wf)$ (7)	$1 - f$ (8)	M $= w(1 - f)$ (9)	B $= 1 - \exp(-M)$ (10)	P $= \ln \frac{A}{B}$ (11)	Q $= w(f - 0.5)$ (12)	$b_r = P + Q + h$ (13)	$S_r = \sqrt{\frac{D}{AB}}$ (14)
1			1										
2			2										
3			3										
4			4										
5			5										
6			6										
7			7										
8			8										
9			9										
10			10										
11			11										
12			12										
13			13										
14			14										
15			15										
16													
Total													

l □ $d.$ $h = \frac{\square}{l}$; $S_d^2 = [\square - \square] \div (L - 1)$ □

h □

h^2 □ ; $l\,h^2$ □

$w = 3.5 \times \sqrt{\square}$

$w =$ □ ; $\frac{w}{L}$ □ ; $1 - \exp(-w) =$ □ $= C$ $D = \frac{Cw}{L} =$ □

L = Test length
h = Test height
w = Test width

Worksheet 2: Calibration of a Test Using PROX (Part 2: Persons)

(8)	(9)	(10)	(11)	(11a)	(12)	(13)
			Initial measure			Standard error
Person number	Person score	$\frac{r}{L-r}$	$b = \ln \frac{r}{L-r}$	b^2	Corrected measure bX	$X\sqrt{\frac{L}{r(L-r)}}$
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						
32						
33						
34						
35						
36						
37						
38						
39						
40						
41						
42						
43						
44						
45						

Worksheet 2: Calibration of a Test Using PROX (Part 3: Calculations)

From Part 1

$$\text{Mean}_i = \frac{\text{Column (4) Total}}{\text{No. of items}} = \frac{\boxed{}}{\boxed{}} = \boxed{}\ ;\ (\text{mean}_i)^2 = \boxed{}$$

$$\text{Variance}_i = U = \frac{\text{Column (4a) Total} - L \times (\text{mean}_i)^2}{L - 1} = \frac{\boxed{} - \boxed{} \times \boxed{}}{\boxed{}}$$

$$U = \boxed{}$$

From Part 2

$$\text{Mean}_b = \frac{\text{Column (11) Total}}{\text{No. of persons}} = \frac{\boxed{}}{\boxed{}} = \boxed{}\ ;\ (\text{mean}_b)^2 = \boxed{}$$

$$\text{Variance}_b = V = \frac{\text{Column (11a) Total} - N \times (\text{mean}_b)^2}{N - 1} = \frac{\boxed{} - \boxed{} \times \boxed{}}{\boxed{}}$$

$$V = \boxed{}$$

X = Person ability expansion factor due to test width

$$= \sqrt{\frac{1 + \frac{U}{2.89}}{1 - \frac{UV}{8.35}}} = \sqrt{\frac{1 + \frac{\boxed{}}{2.89}}{1 - \frac{\boxed{} \times \boxed{}}{8.35}}}$$

$$X = \boxed{}$$

Y = Item difficulty expansion factor due to sample spread

$$= \sqrt{\frac{1 + \frac{V}{2.89}}{1 - \frac{UV}{8.35}}} = \sqrt{\frac{1 + \frac{\boxed{}}{2.89}}{1 - \frac{\boxed{} \times \boxed{}}{8.35}}}$$

$$Y = \boxed{}$$

APPENDIX III

Worksheet 3: Comparing a Sample Group with a Reference Group

(1) Item number	(2) Item bank difficulty	(3) Observed difficulty	(4) (2)-(3)	(5) (4)-Mean	(6) (5)²	(7) (6) × *Q*	(8) Is (7) < 3.84? (Yes/No)
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
31							
32							
33							
34							
35							
36							
37							
38							
39							
40							

K = no. of items ___ Total (4) ___ Total (7) ___

N = no. of persons ___ Mean ___ Is Total (7) < *R* from Table A?

$Q = \frac{NK}{12(K-1)}$ ___ Yes No

Table A

K	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R	5.99	7.82	9.49	11.07	12.59	14.07	15.51	16.92	18.31	19.68	21.03	22.36	23.68	25.00

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

8

The Use of the Rasch Latent Trait Measurement Model in the Equating of Scholastic Aptitude Tests

George Morgan

This paper reports the results of an exploratory investigation which attempted to assess the capabilities of Rasch's Simple Logistic Model in the calibration and equating of final and trial forms of the Australian Scholastic Aptitude Test (ASAT). The investigation had two main aims: (i) to determine to what extent the items in the final and trial forms of the ASAT can be successfully fitted to latent variables of general scholastic aptitude determined by calibrations of items in whole tests and various sub-tests, based primarily on content, and (ii) to determine whether equatings of ASAT forms can be undertaken successfully at the whole test or sub-test levels.

The items in the ASAT† are grouped into units, each unit being concerned with a particular theme. A unit begins with stimulus material, presented in a variety of forms, drawn from the four broad subject (content) areas of humanities, social science, mathematics, and science, and is followed by a group of binary-scored, multiple-choice items related to the stimulus material. The items in the ASAT are designed to measure a wide range of abilities and skills, such as those concerned with the interpretation and comprehension of scholastic materials, that are relevant to academic courses at the Year 12 level of secondary education and at the tertiary level. When the tests are constructed, care is taken to avoid using materials directly related to Year 12 syllabuses.

The ASAT is an omnibus test of scholastic aptitude but it does not relate to any particular theoretical model. Broadly the test's structure is determined to a large extent by the pool of abilities and skills underlying

† The ASAT is a secure test, but a booklet containing a sample collection of items may be obtained for inspection.

the particular items which happen to be incorporated in the test, when a form of the test is constructed.

Although a considerable literature exists on the ASAT (Lees, 1978), for the most part it is concerned with the test's power to predict success in tertiary courses and its use as a scaling instrument.

It appears that the earliest investigation into the psychometric properties of the ASAT was undertaken by McGaw and Greddon (1973). A few years later a comprehensive study of the psychometric properties of the 1973 version of the test, ASAT-B, was carried out by Bell (1977) and, more recently, Bell (1979) factor analysed the 1977 version of the test, ASAT-F. He found that the first principal component of ASAT-F accounted for 10 per cent of the test variance. These studies indicated that the ASAT is factorially complex, and that at a global level the test can be characterized by a general ability factor, and more specifically by factors representing quantitative and verbal abilities.

In his study of ASAT-B, Bell (1977) analysed the test using traditional item analysis procedures as well as those based on the Rasch Simple Logistic Model. He found that about two-thirds of the ASAT-B items conformed to the Rasch model. More recently, Bond (1978) applied the Rasch Simple Logistic Model in the multiplicative binomial framework in an analysis of ASAT-F. He suggested that Rasch measurement of the ASAT should be based on the units rather than on the items, because the items tend to lose the part played by the stimulus material of each unit. Eleven of the eighteen units in the test were calibrated by him to a unidimensional latent trait of general ability.

With a factorially complex test like the ASAT, it is not clear which group of items in a form should be calibrated together in order to permit satisfactory equatings between forms. Obviously, basing the equatings on the estimates from a Rasch multiplicative binomial analysis of the units in forms is impracticable, because the number of link units required for an adequate analysis would require the construction of inordinately long tests.

Even so, if equatings between ASAT forms are to be based on the individual items in the test, it is not clear which items should form the links. Reckase (1979) showed that for factorially complex tests, the Rasch Simple Logistic Model estimates the sum of the factors when there is more than one independent factor, and estimates the first dominant factor when it exists. In the latter situation, he found that stable item calibrations can be obtained even if the first factor accounts for less than 10 per cent of the variance. These findings suggest that stable ASAT equatings might be obtained at the whole test level and, if so, the various test forms could be equated within the existing framework of test development and application.

Currently the test's main function is to act as a uni-valued variable in the scaling of Year 12 public examination results or teacher assessments of student achievements where such examinations do not exist. If the Rasch Simple Logistic Model could be used to equate the ASAT forms at the whole test level, equated scores could be derived from the different forms of the ASAT and then be applied in the scaling process, thus bringing the effects of the scaling to a common base. Otherwise equatings may need to be undertaken at a sub-test level determined by criteria like content homogeneity or pure factor structure.

PROCEDURE

The ASAT program of test development has evolved over a number of years and is now well established. It provides a somewhat routine schedule to be followed in the construction and trial testing of each form of the test. Thus, beginning with a pool of units in each of the subject areas of mathematics, science, humanities, and social science, the units are processed and, from these, units are selected for inclusion in the trial forms. The trial forms are then administered to a sample of Year 12 students. Subsequently, using classical test theory principles of test construction in conjunction with expert considerations about the content and kinds of abilities and skills measured by the items, units are selected to make up the final form of the test.

In formulating the course of the investigation, the intention was to allow work to proceed within the framework of the existing program of test development outlined above. This seemed a profitable course to follow, since preliminary calibrations of the items in the whole test and some sub-tests of the ASAT-G showed that appreciable numbers of the available pool of items were satisfactorily fitted to the Rasch latent ability continuum associated with the whole test and the sub-tests based on the four broad subject areas.

An alternative course would have involved the creation of Rasch-like forms from the outset, perhaps aiming to have forms of equal length and containing sufficient numbers of link items to ensure satisfactory equatings of the forms. However, such an approach would have entailed going beyond the current practice of test development as outlined in the test's specification (ACER, 1978), and this would need to be agreed to by the ASAT users. Nevertheless it seemed at the time that an exploratory investigation within the existing test development framework would be able to shed some light on the kinds of results that might be expected when Rasch measurement techniques are applied to the ASAT.

Whole Tests and Sub-Tests

The investigation was concerned with equating two final forms, ASAT-G

Table 1 Brief Statistical Description of ASAT Forms

Form	Number of items	Number[a] of students	Mean score	Standard deviation	KR 20 reliability
ASAT-G	100	2345	64	16.0	0.93
ASAT-H	100	2422	60	14.8	0.91
V	71	246	32.2	8.2	0.79
W	72	252	29.7	8.2	0.78
Y	72	249	34.6	8.6	0.80
Z	72	248	31.4	8.6	0.80

[a] Data for ASAT-G and ASAT-H came from students in the Australian Capital Territory; data for forms V, W, Y, and Z from students in Tasmania and South Australia who took part in the trial testing of ASAT-H.

and ASAT-H, through four trial forms of ASAT-H. Equatings of forms were analysed using various combinations of the items which were independently calibrated. Table 1 provides a brief statistical description of these forms.

Table 2 shows the distribution of items in each form across the four subject areas. The items in each of the six forms were grouped into a whole test, consisting of all the items in the form, and eight sub-tests which were: Mathematics/Science, Humanities/Social Science, Humanities, Social Science, Mathematics, Science, Quantitative, and Verbal. Except for the Quantitative and Verbal sub-tests, each sub-test contained all items in the relevant subject area(s) that were available.

In constructing the Quantitative and Verbal sub-tests, the following arbitrary criteria were used. Humanities items were assigned to the Verbal sub-test and mathematics items to the Quantitative sub-test. Of the science and social science items, if the point-biserial correlation between

Table 2 Distribution of Items in ASAT Forms According to Subject Area

	Percentage in subject area				
Form	Humanities	Social science	Mathematics	Science	Total number of items
ASAT-G	30	20	30	20	100
ASAT-H	30	20	20	30	100
V	48	0	24	28	71
W	32	22	18	28	72
Y	26	25	24	25	72
Z	32	26	28	14	72

an item and the Mathematics/Science sub-test was greater by 0.03 than the point-biserial correlation between the item and the Humanities/Social Science sub-test, the item was assigned to the Quantitative sub-test. Conversely items were assigned to the Verbal sub-test. In cases where a decision could not be made on the basis of correlations, items were classified on the basis of their face validity. A few items could not be classified using these criteria and hence were omitted from these sub-tests.

Link Structure

The following list gives the length of each form, the subject area of the link items in the form, and the name (in parenthesis) of the form(s) to which it was linked. Forms V, W, Y, and Z are the trial forms of ASAT-H used in this study.

Form V	71 items	16 humanities items, 10 science, and 5 mathematics items (Form H) 10 humanities items (Form Z)
Form W	72 items	14 humanities, 6 social science, and 9 science items (Form H) 5 science and 5 mathematics items (Form Y)
Form Y	72 items	10 social science, 6 science, and 4 mathematics items (Form G) 5 science and 3 mathematics items (Form H) 10 humanities items (Form V)
Form Z	72 items	10 humanities, 5 science, and 5 mathematics items (Form G) 4 science and 6 mathematics items (Form G) 5 science and 5 mathematics items (Form W)
Form G (ASAT-G)	100 items	10 social science, 6 science, and 4 mathematics items (Form Y) 10 humanities, 5 science, and 5 mathematics items (Form Z)
Form H (ASAT-H)	100 items	16 humanities, 10 science, and 5 mathematics items (Form V). 14 humanities, 6 social science, and 9 science items (Form W) 5 science and 3 mathematics items (Form Y) 5 science and 5 mathematics items (Form Z)

This arrangement of link units among the forms allowed the investigation of form equating at the whole test level and at the Mathematics, Science, Quantitative, and Verbal sub-test levels. The link structure is shown schematically in Figure 1.

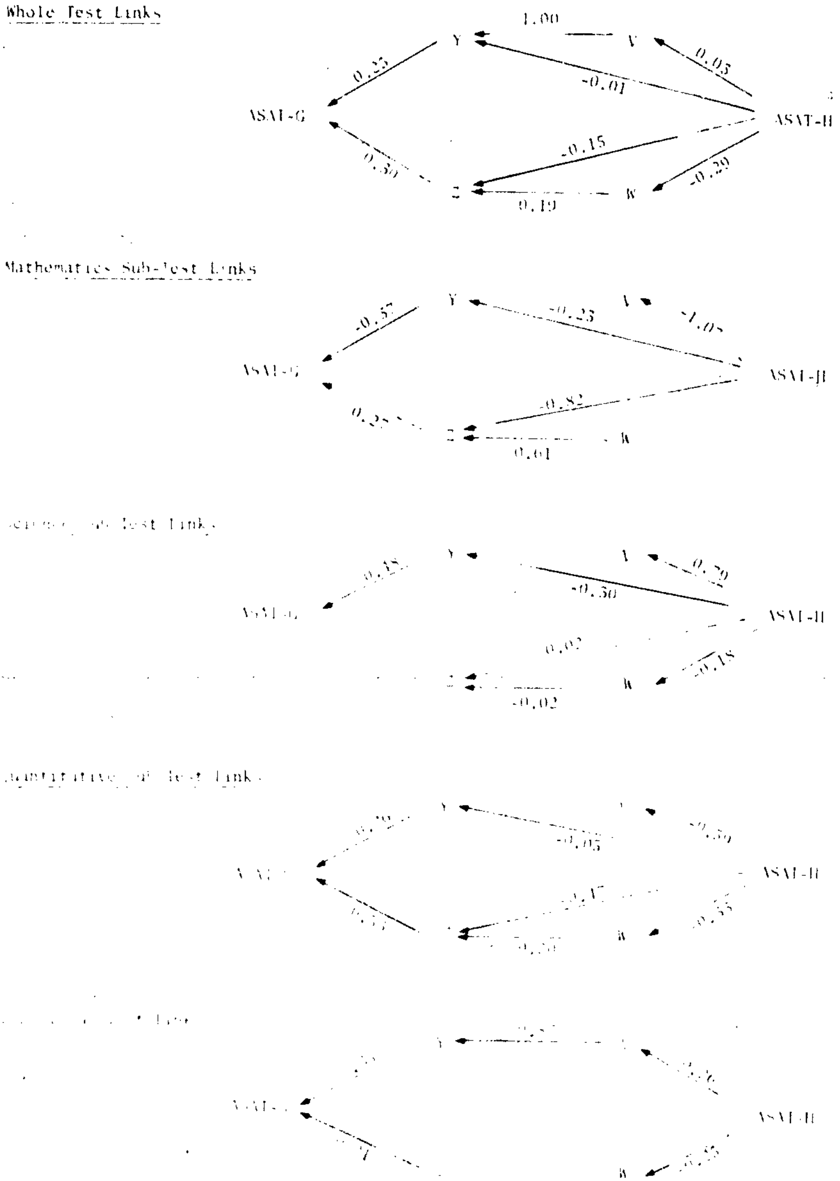


Figure 1 Link Structure for ASAT Equatings with Translation Constants

In order to preserve the unit structure of the ASAT when equating forms, entire units were initially selected to provide the links between forms rather than individual items. In all cases the link units had average item facilities and average item point-biserial discrimination values which were neither too large nor too small.

Given the amount of material to be trial tested, the samples of students available for analysing the total test data, and the availability of a relatively short testing time (of 1 ¾ hours), it was not possible to include more link items within the trial forms.

Calibration Samples

For the ASAT-G and ASAT-H, the calibration samples were two separate groups of randomly selected Year 12 students who sat the tests in the Australian Capital Territory in September of 1978 and 1979, respectively; for Forms V, W, Y, and Z the calibration samples were Year 12 students who participated in the trial testing of the ASAT-H in Tasmania and South Australia in March 1979.

Computer Program for Rasch Measurement

The program used was CALFIT-3, a computer program adapted by R. Wines and D. Keunemann from one designed by B. Wright and R. Mead (Cornish, 1976).

CALFIT-3 estimates item difficulties and person abilities of the Rasch Simple Logistic Model using the corrected unconditional maximum likelihood statistical procedure (Wright and Panchapakesan, 1969). In addition it estimates how well an item conforms to the Rasch model. CALFIT-3 also estimates a probability of sub-test fit which indicates how well a group of items conforms to the model, as items are accumulated one by one into a sub-test, starting with the best fitted item.

The program performs its calculations in two cycles. First it calibrates all the items, omitting items which everyone answers correctly or everyone answers incorrectly, and then estimates person abilities after deleting persons with zero or possible maximum raw score. In the next cycle it gathers the best-fitted items, according to a probability of sub-test fit cut-off provided by the user, recalibrates this group of items, and produces revised estimates of person abilities.

Method of Equating or Linking

This was the Rasch common item method of equating tests which is described in detail by Wright (1977) and Wright and Stone (1979).

Suppose Test *a* and Test *b* share a common set of *K* items, called the link items. In the Rasch common item method of equating two tests, the scale of the latent variable of one of the tests, say Test *b*, is adjusted to the scaie of the latent variable of the other test, Test a, using the difference in average estimated difficulties of the common items from the two separate calibrations to translate the difficultv estimates of Test b to the scale of Test a. Proyiding the common items and the other items in both tests conform to the Rasch model, and are calibrated to the same latent variable, this method yields a pool of calibrated items whose estimated difficulties are on a common scale.

A summary of the main elements of this method, proposed by Wright (1977) and Wright and Stone (1979) follows.

1 Begin by separately calibrating the items in Test a and Test b, which give two independent sets of estimated item difficulties for the link items. Let d_{ia} and d_{ib} represent the estimated item difficulties of the ith item in the link, in Test a and Test b respectively.

2 Calculate the translation constant which effectively translates all item difficulty estimates from the calibration of Test b to the calibration scale of Test a, using the formula

$$t_{ab} = \sum_{i}^{K} (d_{ia} - d_{ib})/K.$$

This translation constant is the difference in average estimated item difficulties of the common items in the two calibrations. The standard error of the estimated translation constant, SE(t_{ab}) is approximately $3.5/(NK)$ where N is the calibration sample size of the link items and K is the number of items in the link. Unfortunately this expression for the standard error of the translation constant applies to the situation where the link items are calibrated in a separate test, taken by N examinees. In this investigation the link items were placed in two separately calibrated forms and hence the formula did not apply because the calibration sample sizes differed for the two forms. However, so as to obtain an approximation for this error the value of N in the expression was arbitrarily taken to be the smaller of the two calibration sample sizes.

3 The validity of the link between Test a and Test b may be tested using the statistic

$$\frac{N}{12} \frac{K}{(K-1)} \sum_{i}^{k} (d_{ia} - d_{ib} - t_{ab})^2,$$

which is distributed approximately as a chi-square with K degrees of freedom. Alternatively the validity of the link may be tested by determining the mean and standard deviation of the standardized residuals

$$\frac{d_{ia} - d_{ib} - t_{ab}}{S_D}$$

where $S_D = (SE(d_{ia})^2 + SE(d_{ib})^2)$, to see if these estimate the expected mean equal to zero and expected standard deviation equal to 1.

4 The validity of an item in the link may be tested using the statistic

$$\frac{N}{12}\frac{K}{K-1}(d_{ia}-d_{ib}-t_{ab})^2$$

which is distributed approximately as chi-square with one degree of freedom.

5 Alternatively the validity of the link and the items in the link may be ascertained visually by plotting the estimated difficulty estimates of the common items from the two calibrations, and observing the extent to which the points are scattered about the line of perfect agreement.

6 If three or more tests are linked so as to form a closed loop, the consistency of the links may be tested by summing the corresponding translation constants around the loop and examining whether this sum estimates zero within one or two standard errors of this sum. For example, if Test *a*, Test *b*, and Test *c* form a loop, then

$$t_{ab}+t_{bc}+t_{ca}\approx 0$$

The standard error of the sum may be estimated using the expression

$$3.5\left\{\frac{1}{N_{ab}K_{ab}}+\frac{1}{N_{bc}K_{bc}}+\frac{1}{N_{ca}K_{ca}}\right\},$$

where, in this study N_{ab}. etc. were taken to be the smaller of the two calibration sample sizes, and K_{ab}. etc. are the number of common items in the links.

RESULTS AND DISCUSSION

Only the calibration results based on the whole test, and the Mathem..ics, Science, Quantitative, and Verbal sub-tests are presented here. The results for the other sub-tests were similar, and have consequently been omitted.

In all the item calibrations undertaken, the major reason for some items not conforming to the Rasch model was the item discrimination – the observed discriminations were either too large or they were too small, their values departing markedly from the model value.

Table 3 reports the percentages of fitted items of tests and sub-tests according to the subject area of the items. Considering the results for whole test calibrations, and all items in the form, greater percentages of items in the trial forms were fitted than in the final forms, ASAT-G and ASAT-H. This result is not unexpected for it reflects the sample size sensitivity of the chi-squared method of assessing item and sub-test fit. The larger the calibration sample size, the more likely will small discrepancies between the observed and estimated item characteristic curves be found

Table 3 Percentages of Fitted Items[a] of Tests and Sub-tests According to the Subject Area of the Items

Test/ Sub-test	Form	Humanities	Social science	Mathematics	Science	All
		Percentage fitted/Number of items in subject area				
Whole test	ASAT-G	57/30	75/20	53/30	70/20	62/100
	ASAT-H	73/30	70/20	50/20	97/30	75/100
	V	94/34		76/17	90/20	89/71
	W	87/23	94/16	77/13	90/20	88/72
	Y	95/19	83/18	100/17	94/18	93/72
	Z	83/23	89/19	80/20	80/10	83/72
Mathematics sub-test	ASAT-G			63/30		
	ASAT-H			80/20		
	V			76/17		
	W			77/13		
	Y			88/17		
	Z			90/20		
Science sub-test	ASAT-G				90/20	
	ASAT-H				97/30	
	V				90/20	
	W				100/20	
	Y				89/18	
	Z				100/10	
Quantitative sub-test	ASAT-G		79/14	57/30	62/13	63/57
	ASAT-H		100/4	65/20	92/26	82/50
	V			59/17	82/11	68/28
	W		83/6	92/13	89/18	89/37
	Y		100/3	94/17	100/13	97/33
	Z		67/6	90/20	50/8	76/34
Verbal sub-test	ASAT-G	70/30	83/6		83/6	74/42
	ASAT-H	83/30	85/13		100/4	85/47
	V	94/34			100/9	95/43
	W	91/23	100/10		100/2	94/35
	Y	95/19	100/12		100/5	95/36
	Z	78/23	100/2		100/2	86/37

Cut off for probability of sub-test fit was 0.01.

to be significant. Actually the calibration sample sizes of ASAT-G and ASAT-H were appreciably greater than those of the trial forms. The percentages of items fitted in the trial forms were about the same. Moreover the ASAT-G items fitted less well as a group than the ASAT-H items, because fewer humanities and science items in the test conformed to the model. In terms of the subject area, the group of items which fitted worst of all were the mathematics items. On the basis of these results it seems, perhaps with the exception of the mathematics items, that acceptable percentages of items were fitted from the pool of items available in each form. Obviously, had the pools of items been sufficiently large it would have been easy to calibrate and fit Rasch-like items to give calibrated whole tests of any desired length.

The results in Table 3, for the calibrations of the items in the Quantitative and Verbal sub-tests, indicate that generally slightly more items were fitted than in calibrations based on the whole test. However, the pattern in the percentages of fitted items, according to the subject area of the items, was not entirely consistent across the forms. In some cases a greater percentage of items was fitted in a subject area than was the case with calibrations based on the whole test, and in some cases the situation was reversed. In the Quantitative and Verbal sub-tests there was no obvious pattern in the subject areas of the better fitted items. That is, the rank order of the better fitted items in both sub-tests did not show a pattern of preferences for any of the subject areas from which the items were drawn.

Calibrations of items in the Mathematics and Science sub-tests in general fitted greater percentages of the available items than did calibrations based on the whole tests. For example, 10 per cent more mathematics items in ASAT-G were fitted in calibrations based on the Mathematics sub-test, and 30 per cent more were fitted in ASAT-H than in the calibrations based on the whole tests.

A tentative generalization is that a greater percentage of ASAT items will conform to the Rasch Simple Logistic Model, if the items in each subject area of the ASAT are calibrated independently. Apparently the items in each subject are more closely represented in terms of a unidimensional latent variable, from the point of view of the Rasch Simple Logistic Model, than are the items in the 'impure' sub-tests which contain items from two or more different subject areas. The problem of dimensionality is not simply a matter that deals with the subject area of the items, but rather one of identifying those abilities and skills, forming the latent variable, that are common to the group of items which must explain consistent examinee performance on the test.

Table 4 presents statistics of item difficulty estimates of the fitted items for the whole and sub-tests calibrated. The mean of the item difficulty estimates is zero in each case, and fixes the origin of the calibration scale.

At each test/sub-test calibration level, the ranges of the estimated item difficulties, for most forms, are quite similar to each other. This, together with the fact that the standard deviations of the estimated item difficulty estimates are much larger than the average standard errors of these estimates, suggests that the items in the tests/sub-tests were sufficiently scattered on the calibration scales to give the latent variables direction. It

Table 4 Calibration Results for Fitted Items[a] of ASAT Forms According to Test/Sub-test Calibrated

Test/ Sub-test	Form	Percentage of fitted items	Statistics of item difficulty estimates[b]					Calibration sample size
			d_{min}	d_{max}	Range	SD	Average SE(d)	
Whole test	ASAT-G	62	−2.26	2.36	4.62	0.91	0.13	312
	ASAT-H	75	−2.05	1.56	3.61	0.71	0.12	308
	V	89	−2.98	2.11	5.09	0.96	0.14	240
	W	88	−2.07	2.00	4.07	0.87	0.14	247
	Y	93	−1.64	1.83	3.47	0.86	0.14	241
	Z	83	−2.92	2.18	5.10	0.94	0.14	238
Mathematics sub-test	ASAT-G	63	−1.45	1.80	3.25	1.04	0.13	336
	ASAT-H	80	−1.22	1.75	2.97	0.85	0.14	309
	V	76	−1.53	1.33	2.86	0.98	0.19	192
	W	77	−1.83	1.60	3.43	1.21	0.21	147
	Y	88	−1.72	1.04	2.76	0.80	0.15	225
	Z	90	−3.42	2.24	5.66	1.62	0.18	231
Science sub-test	ASAT-G	90	−2.04	1.41	3.45	1.03	0.13	303
	ASAT-H	97	−1.88	1.60	3.48	0.74	0.13	288
	V	90	−2.69	2.52	5.21	1.14	0.15	232
	W	100	−1.41	0.98	2.39	0.64	0.15	239
	Y	89	−1.80	1.69	3.49	0.88	0.15	238
	Z	100	−0.74	0.81	1.55	0.54	0.15	232
Quantitative sub-test	ASAT-G	63	−1.95	2.35	4.30	0.97	0.13	304
	ASAT-H	82	−2.16	1.60	3.76	0.70	0.13	343
	V	68	−1.77	1.71	3.48	1.01	0.17	225
	W	89	−2.29	1.92	4.21	0.95	0.15	236
	Y	97	−1.58	1.65	3.23	0.76	0.15	236
	Z	76	−3.30	2.18	5.48	1.34	0.15	225
Verbal sub-test	ASAT-G	74	−2.12	1.38	3.50	0.91	0.13	280
	ASAT-H	85	−1.43	1.28	2.71	0.66	0.12	345
	V	95	−2.73	1.65	4.38	0.89	0.14	235
	W	94	−1.53	1.62	3.15	0.80	0.14	247
	Y	95	−1.50	1.48	2.98	0.88	0.15	237
	Z	86	−1.31	1.83	3.14	0.64	0.14	238

[a] Cut-off for probability of sub-test fit was 0.01. [b] Measured in logits.

appears that each form was somewhat successful in providing enough items for each test sub-test, providing useful yardsticks against which the abilities and skills of examinees could be assessed in the continuum of the relevant latent variable. However, the effective ranges of the estimated item difficulties are not as large as were expected for an omnibus test like the ASAT. It is not surprising that the ASAT test items measure scholastic aptitude along a narrow range of the potential scholastic continuum, because the test construction procedures currently in use select items with facilities centred around 50 per cent and generally exclude items whose facilities are below 20 per cent or greater than 80 per cent.

Table 5 shows the range of examinee abilities for ASAT-G and ASAT-H at the test sub-test calibration levels. Without exception, the range of estimated abilities was greater than the range of estimated item difficulties. The match between estimated abilities and estimated item difficulties, measured in terms of their overlap, was not entirely satisfactory for efficient measurement practice. As can be seen in the examples of ASAT-G and ASAT-H (Tables 4 and 5), the whole tests and sub-tests, with the exception of the Mathematics sub-tests, were somewhat too easy for the calibration sample. In the case of the Mathematics sub-tests, they were too difficult for some students, matched to the abilities of some, and far too easy for the rest of the calibration samples. Similar results were obtained with the trial forms.

Equating Analyses

Unfortunately, as it turned out, insufficient numbers of link items were fitted in some forms, and this undoubtedly affected the validity of the subsequent equating analyses.

The results of the equating analyses, for the links illustrated in Figure 1, are reported in Table 6. The forms have been statistically linked at the whole test level, and the Mathematics, Science, Quantitative and Verbal sub-test levels. Further details of the results of equating are given in Tables A.1 to A.4 in the Appendix to this paper.

In Table 6 are reported estimates of the translation constant t_{ab}, and its estimated standard error SE(t_{ab}), for the situation where Form a is linked to the calibration scale determined by Form b. The standard deviation of the difference in the estimated item difficulties of the link items in the two independent calibrations, SD($d_a - d_b$), provides a measure of the coherence of the two sets of estimated item difficulties. The smaller the SD($d_a - d_b$), the less 'noise' there is in the link. Links with a lot of noise might result from calibrations which have defined two different latent variables, perhaps to the extent that one or both calibrations were based in part on extraneous variables.

Table 5 Estimates[a] of Range of Examinees' Abilities for Fitted Items of Tests/Sub-tests of ASAT-G and ASAT-H

Form	Test Sub-test	Minimum		Maximum		Min	Estimated range of examinees' abilities
		b	Raw score	*b*	Raw score	SE(*b*)	
ASAT-G	Whole test	-1.43	14	3.78	60	0.28	5.21
	Mathematics sub-test	-2.49	2	3.35	18	0.51	5.84
	Science sub-test	-1.93	3	3.20	17	0.52	5.13
	Quantitative sub-test	-1.66	7	4.05	35	0.36	5.71
	Verbal sub-test	-1.91	5	3.71	30	0.39	5.62
ASAT-H	Whole test	-1.11	20	2.64	69	0.24	3.75
	Mathematics sub-test	-2.17	2	3.02	15	0.54	5.19
	Science sub-test	-1.73	5	2.35	26	0.39	4.08
	Quantitative sub-test	-1.55	8	3.18	35	0.33	4.73
	Verbal sub-test	-1.06	11	2.09	18	0.33	3.15

[a] Measured in logits.

Table 6 Results of the Equating Analyses

Test Sub-test	Forms linked b a	Fraction of link items used	SD($d_a - d_b$)	t_{ab}	SE(t_{ab})	Smaller calibration sample size
Whole test	G–Y[a]	12/20[b]	0.18	0.25	0.06	241
	G–Z	10/20	0.19	0.30	0.07	238
	Y–V	8/10	0.25	1.00	0.08	240
	Y–H	3/8	0.30	−0.01	0.13	241
	Z–W	7/10	0.20	0.19	0.08	247
	Z–H	3/10	0.50	−0.15	0.13	238
	V–H	22/31	0.23	0.03	0.05	240
	W–H	19/29	0.27	−0.29	0.05	247
Mathematics sub-test	G–Y	4/12	0.31	−0.37	0.12	225
	G–Z	3/5	0.31	0.27	0.13	231
	Y–H	2/3	0.04	−0.23	0.16	225
	Z–W	3/5	0.08	0.61	0.17	147
	Z–H	5/6	0.45	−0.82	0.10	231
	V–H	4/5	0.32	−1.07	0.13	192
Science sub-test	G–Y	4/6	0.18	0.48	0.11	238
	Y–H	4/5	0.26	−0.30	0.11	238
	Z–W	4/6	0.21	−0.02	0.11	232
	Z–H	4/4	0.53	0.02	0.11	232
	V–H	7/10	0.63	0.79	0.09	232
	W–H	7/9	0.41	−0.18	0.09	239
Quantitative sub-test	G–Y	9/12	0.24	0.26	0.08	236
	G–Z	3/6	0.19	0.53	0.13	225
	Y–H	3/7	0.15	−0.05	0.13	236
	Z–W	7/10	0.19	0.20	0.09	225
	Z–H	5/10	0.44	−0.47	0.10	225
	V–H	6/12	0.24	−0.59	0.10	225
	W–H	7/10	0.37	−0.33	0.09	236
Verbal sub-test	G–Y	4/7	0.09	0.35	0.11	237
	G–Z	7/12	0.22	0.21	0.09	238
	Y–V	8/10	0.25	0.87	0.08	235
	V–H	13/19	0.25	0.26	0.06	235
	W–H	14/18	0.25	−0.25	0.06	247

[a] Form Y linked to the scale determined by Form G.
[b] Of the 20 link items originally calibrated, the 12 best fitted items were used in the link analysis.

Only the better fitted link items were selected for calculating the translation constants. Some of the links proved to be very tenuous, consisting of only two or three items, but the estimated translation constants for these links were in general agreement with other translation constants, as was shown in the assessment of link coherence.

Since the standard error of a translation constant is approximately inversely proportional to the product of the number of items in the link and the size of the smaller calibration sample, a link with a few items could still give a respectable standard error providing the sample was large enough. But it seems too much to hope that links consisting of a few items can be stable in most practical situations, especially if the links contain much noise.

An interesting result of this investigation is that the translation constants at the test sub-test levels are mostly small, and in many of the equatings closer than about three standard errors from zero.

Examples of Equating Analyses

To illustrate the process of equating ASAT forms, results are presented for the equating at the whole test level of Form W to the scale determined by Form Z, and for the equating at the whole test level of Form V to the scale determined by Form Y.

Separate calibrations of items, using test data from the calibration samples, were carried out for the four forms, producing item difficulty estimates and estimates of their standard error. The number of students in the calibrations were: 247 for Form W, 238 for Form Z, 241 for Form Y, and 240 for Form V. Sixty-two of the items of Form W and 59 of the items of Form Z were successfully calibrated using a probaoility of subtest fit cut-off equal to 0.01, while for Forms Z and V the number of items successfully calibrated were 60 and 63, respectively.

Equating Form W to Form Z

Items 5Z and 72Z of Form Z and items 72W and 5W of Form W were omitted from the equating process because they failed to fit the Rasch model in one of the separate calibrations. Item 68Z of Form Z (item 1W of Form W) was omitted because it fell outside the 95 per cent confidence region (see Table A.1).

Consequently, of the original ten items in the link, seven were used to calculate the equating constant. Table A.1 sets out the stages in the calculation of the equating constant ($t_{wz} = 0.19$), and the test of the validity of the link. The link is statistically valid since the standardized residuals in Table A.1 are distributed with approximately zero mean and approximately unit standard deviation. The obtained mean (– 0.05) and standard deviation (0.82) do not differ appreciably from the expected values.

The two estimated difficulties of the link items were transformed to a common scale determined by Form Z, first by adding the translation constant to the estimate d_w of Form W, and then finding the average of this new value and the estimates d_z. These averages (indicated by superscript^a) are shown in Table A.2 under d_i^a and d_j^a. In transforming the rest of the items in the two forms to a common scale, difficulty estimates of items in Form Z not in the link remained unchanged, while difficulty estimates of items in Form W were increased in value by 0.19 (the translation constant).

Equating Form V to Form Y

Item 68Y of Form Y and item 6V of Form V were omitted from the equating process because they failed to fit the Rasch model in both calibrations. Item 5V of Form V (item 67Y of Form Y) was omitted because it fell outside the 95 per cent confidence region (see Table A.3). Only eight items were used to calculate the equating constant and to test the validity of the link. Tables A.3 and A.4 show the results of the linking exercise for Forms V and Y.

Consistency of the Links

Figure 1 shows a number of closed loops joining three or more of the forms at the whole test and Mathematics, Science and Quantitative sub-test levels. If the sum of estimated translation constants for a certain loop should estimate zero within one or two standard errors of the sum, the links in the loop are said to be statistically consistent. In essence this kind of information supplies additional support for the validity of the links making up the loop.

Table 7 shows this sum and its associated standard error for eleven loops. In all cases, except loops containing the link Y → V at the whole test level and Loop 11, this sum is within one standard error of zero. In Loop 11 the sum is approximately one standard error away from zero. It appears that the links in these loops, at the test/sub-test calibration levels, are consistent at least in terms of the criteria proposed by Wright and Stone (1979).

The loops containing the link Y → V appear to be inconsistent. Since the link items in Y → V, which comprise a unit of 10 humanities items, are the first unit in Form V and the last unit in Form Y, it is possible that the position of the unit in the two forms affected the estimation of the translation constant. Indeed the average facility of the link items was 40 per cent in Form Y and 57 per cent in Form V; and for the eight best fitted items it was 43 per cent and 63 per cent, respectively. The difference in individual item facilities was almost constant between the two forms, ranging between 15 and 20 per cent. Moreover both calibration samples were comparable in terms of their composition. The relatively large translation constant reflects the fact that the link items in Form Y were estimated to be more difficult than in Form V. These results suggest that,

Table 7 Evaluation of the Consistency of the Links in the Loops Displayed in Figure 1

Test/Sub-test	Loop number	Loop	Sum of the translation constants	Standard error of the sum of the translation constants	
Whole test	1	G→Y→V→H→W→Z→G	−1.08	0.39	(inconsistent)[a]
	2	G→Y→V→H→Z→G	−1.13	0.39	(inconsistent)
	3	G→Y→H→W→Z→G	−0.04	0.39	
	4	G→Y→H→Z→G	−0.09	0.39	
	5	Y→V→H→Y	−1.02	0.26	(inconsistent)
	6	Z→W→H→Z	−0.05	0.26	
Mathematics sub-test	7	G→Y→H→Z→G	−0.02	0.51	
Science sub-test	8	Z→W→H→Z	0.22	0.31	
Quantitative sub-test	9	G→Y→H→W→Z→G	0.19	0.52	
	10	G→Y→H→Z→G	−0.15	0.44	
	11	Z→W→H→Z	−0.34	0.28	

[a] The links in the loop are consistent if the sum of the translation constants estimates zero within one or two standard errors.

in this case, the item calibrations and form equatings based on the trial test data may be somewhat unreliable.

CONCLUSION

The results obtained in this investigation indicate that it is feasible to calibrate the ASAT and to equate its forms using the Rasch Simple Logistic Model. However, if the ASAT test is to be prepared on the basis of Rasch measurement principles, the existing program of test development, as exemplified in the test's list of specifications, will need to be modified to allow the preparation of Rasch-like tests. From the percentages of fitted items at the whole test and sub-test levels, it is clear that larger pools of items in each of the subject areas would be required than are now available, if test lengths of 100 items are to be achieved.

A crucial and important aspect of Rasch measurement is the assessment of item fit to a 'unidimensional' latent variable. The Rasch model assumes that only one latent variable exists, but it might be reasonably argued that with factorially complex tests like the ASAT, more than one latent variable is really needed to explain the complex pattern of test responses. Perhaps the way round this problem is to break up the test into homogeneous parcels or sub-tests each of which can be safely characterized by a single latent variable. But this action might not guarantee unidimensional sub-tests, because a single item may measure many kinds of abilities and skills. This is a real dilemma for the test developer who has to construct and arrange test items into meaningful and useful parcels.

The results seem to indicate that the Rasch Simple Logistic Model will attempt to fit to a common latent variable any group of items that cohere in some fashion. It will attempt to do this on statistical grounds and will pick up as the latent variable a kind of lowest common denominator. This observation is in agreement with the findings of Reckase (1979).

Finally the limited results of the equating analyses suggest that the ASAT may be equated at the whole test level and various sub-test levels. Unfortunately the stability of some of the links is questionable because they consist of very few items. Nevertheless a general picture has emerged which should lend some support to those interested in applying Rasch measurement principles to factorially complex scholastic aptitude tests like the ASAT.

REFERENCES

Australian Council for Educational Research. *Research papers relating to the Australian Scholastic Aptitude Test.* Hawthorn, Vic.: ACER, 1978.

Bell, R. C. *A psychometric study of the Australian Scholastic Aptitude Test (Series B).* Nedlands, WA: University of Western Australia, Research Unit in University Education, 1977.

Bell, R. C. The structure of ASAT-F: A radial parcel double factor solution. In Australian Council for Educational Research, *Research papers relating to the Australian Scholastic Aptitude Test.* Hawthorn, Vic.: 1979, 71–8.

Bond, M. W. An analysis of the Australian Scholastic Aptitude Test (Series F). Unpublished report for the Board of Secondary Education, Western Australia, 1978.

Cornish, G. B. *Calfit 3: A program for the Rasch item analysis technique.* Hawthorn, Vic.: Australian Council for Educational Research, 1976.

Lees, L. *Research relating to the Australian Scholastic Aptitude Test: A select annotated bibliography.* Hawthorn, Vic.: Australian Council for Educational Research, 1978.

McGaw, B. and Gredden, G. Factor structure of ability and achievement tests for science and humanities students. Paper presented at the annual conference of the Australian Association for Research in Education, Sydney, 1973.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 1979, **4**, 207–30.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, **14**, 97–116.

Wright, B. D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, **29**, 23–48.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement.* Chicago: MESA Press, 1979.

APPENDIX

Table A.1 Calculation of the Translation Constant (t_{wz}) and Test of the Validity of the Link Z←W: Whole Test Equating (Excluding Items 5Z, 68Z, 72Z, 72W, 1W, 5W)

Form Z			FormW					Item link fit		
Item	d_z	$SE(d_z)$	Item	d_w	$SE(d_w)$	$D = d_z - d_w$	$D - t_{wz}$	CHI[a]	S_D	Z_i
1Z	0.87	0.16	68W	0.91	0.16	- 0.04	0.23	1.23	0.23	1.00
2Z	1.98	0.22	69W	1.62	0.19	0.36	0.17	0.67	0.29	0.59
3Z	2.18	0.24	70W	1.86	0.21	0.32	0.13	0.39	0.32	0.41
4Z	2.18	0.24	71W	2.00	0.22	0.18	-0.01	0.0	0.33	0.03
5Z	1.04	0.16	72W	[b]	[b]					
68Z	- 0.40	0.14	1W	- 1.26	0.14	0.86				
69Z	0.13	0.14	2W	- 0.22	0.13	0.09	0.10	0.0	0.19	0.53
70Z	1.14	0.17	3W	0.68	0.15	0.46	0.27	1.70	0.23	1.17
71Z	1.09	0.16	4W	1.12	0.17	- 0.03	-0.22	1.13	0.23	0.96
72Z	[b]	[b]	5W	0.55	0.15					
Mean	1.33			1.14		0.19	0.00			0.05
SD	0.85			0.77		0.20	0.20			0.82

Translation constant $t_{wz} = \sum_{1}^{7} D_i/7 = 0.19$

[a] CHI is distributed approximately as χ^2 with 1 degree of freedom (Wright and Stone, 1979, p. 96).
[b] item not fitted when calibrated.

Table A.2 Item Difficulty Estimates for Form W and Form Z from Initial Calibrations and upon Translation to a Common Scale Determined by Form Z: Whole Test Equating
(Estimates for first 20 items of each form are shown.)

Item	Form W			Item	Form Z		
	d_w	SE(d_w)	d_w^t		d_z	SE(d_z)	d_z^t
1W	−1.26	0.14	−0.74	1Z	0.87	0.16	0.99[a]
2W	−0.22	0.1[illegible]	−0.08[a]	2Z	1.98	0.22	1.90[a]
3W	0.68	0.[illegible]	1.01[a]	3Z	2.18	0.24	2.12[a]
4W	1.12	0.17	1.20[a]	4Z	2.18	0.24	2.19[a]
5W	0.55	0.15	omit	5Z	1.04	0.16	omit
6W	2.07	0.17	1.88	6Z	−1.54	0.16	−1.54
7W	0.76	0.13	−0.57	7Z	−0.84	0.14	−0.84
8W	−1.20	0.14	−1.01	8Z	−0.69	0.14	−0.69
9W	−0.67	0.13	−0.48	9Z	−1.25	0.15	−1.25
10W	1.06	0.16	1.25	10Z	0.84	0.15	0.84
11W	1.20	0.14	1.01	11Z	−0.29	0.15	−0.29
12W	[b]	[b]	[b]	12Z	−0.60	0.14	−0.60
13W	0.34	0.14	0.53	13Z	−0.48	0.14	−0.48
14W	[b]	[b]	[b]	14Z	0.39	0.14	0.39
15W	−0.17	0.13	0.02	15Z	−0.60	0.14	−0.60
16W	0.37	0.14	0.56	16Z	−0.42	0.14	−0.42
17W	0.20	0.14	0.39	17Z	−0.73	0.14	−0.73
18W	0.18	0.14	0.37	18Z	[b]	[b]	[b]
19W	−0.15	0.14	0.04	19Z	0.41	0.14	0.41
20W	1.14	0.17	1.33	20Z	[b]	[b]	[b]

d_w, d_z item difficulty estimates in logits
SE(d_w), SE(d_z) standard errors of estimates in logits
d_w^t, d_z^t average difficulty estimates on common scale in logits
omit item omitted from common scale because it was not fitted in both calibrations

[a] used to calculate the translation constant
[b] item not fitted when calibrated

Table A.3 Calculation of the Translation Constant (t_{yv}) and Test of the Validity of the Link Y – V: Whole Test Equating (Based upon the First Four and Last Four Items in the Table)

Form Y			Form V					Item Link Fit,		
Item	d_y	SE(d_y)	Item	d_v	SE(d_v)	$D = d_y - d_v$	$D - t_{yv}$	CHI[a]	S_D	Z_i
63Y	0.03	0.13	1V	-1.21	0.15	1.24	0.24	1.34	0.20	1.20
64Y	0.77	0.14	2V	0.00	0.14	0.77	-0.23	1.21	0.20	-1.15
65Y	-0.02	0.13	3V	-0.81	0.14	0.79	-0.21	1.01	0.19	-1.11
66Y	-0.79	0.14	4V	-2.17	0.20	1.38	0.38	3.30	0.24	1.58
67Y	1.07	0.15	5V	0.98	0.15	0.09				
68Y	1.05	0.15	6V	[b]	[b]					
69Y	0.03	0.13	7V	-1.08	0.15	1.11	0.11	0.28	0.20	0.55
70Y	1.03	0.15	8V	0.17	0.14	0.86	-0.14	0.45	0.21	-0.67
71Y	0.21	0.14	9V	-0.93	0.14	1.14	0.14	0.45	0.20	0.70
72Y	0.92	0.15	10V	0.24	0.14	0.68	-0.32	2.34	0.21	-1.52
Mean	0.42			-0.52		1.00	-0.00			-0.05
SD	0.46			0.63		0.25	0.25			1.12

Translation constant $t_{yv} = \sum_{1}^{8} D_i/8 = 1.00$

[a] CHI is distributed approximately as χ^2 with 1 degree of freedom (Wright and Stone, 1979, p. 96).
[b] Item not fitted when calibrated.

Table A.4 Item Difficulty Estimates for Form V and Form Y from Initial Calibrations and upon Translation to a Common Scale Determined by Form Y: Whole Test Equating (Estimates for first ten and last ten items of each form are shown.)

Form V				Form Y			
Item	d_v	SF(d_v)	d^*_v	Item	d_y	SE(d_y)	d^*_y
1V	1.21	0.15	−0.09[a]	1Y	−0.71	0.14	−0.71
2V	0.00	0.14	0.88[a]	2Y	0.03	0.13	0.03
3V	0.81	0.14	0.08[a]	3Y	−0.87	0.14	−0.87
4V	2.17	0.20	0.98[a]	4Y	1.64	0.17	−1.64
5V	0.98	0.15	1.96	5Y	0.98	0.15	0.98
6V	[b]	[b]	[b]	6Y	−1.29	0.16	−1.29
7V	1.08	0.15	−0.03[a]	7Y	−0.54	0.14	−0.54
8V	0.17	0.14	1.10[a]	8Y	[b]	[b]	[b]
9V	0.93	0.14	0.14[a]	9Y	−0.23	0.14	−0.23
10V	0.24	0.14	1.08[a]	10Y	−1.37	0.16	−1.37
62V	0.39	0.14	1.39				
63V	0.09	0.14	1.09	63Y	0.03	0.13	−0.19[a]
64V	0.12	0.14	0.88	64Y	0.77	0.14	0.88[a]
65V	0.37	0.14	0.63	65Y	−0.02	0.13	0.08[a]
66V	0.58	0.14	0.42	66Y	−0.79	0.14	−0.98[a]
67V	0.37	0.14	1.33	67Y	1.07	0.15	1.07
68V	0.41	0.14	0.59	68Y	1.05	0.15	omit
69V	0.76	0.15	1.76	69Y	0.03	0.13	−0.03[a]
70V	0.07	0.14	0.93	70Y	1.03	0.15	1.10[a]
71V	0.33	0.14	1.33	71Y	0.21	0.14	0.14[a]
				72Y	0.92	0.15	1.08[a]

d_v, d_y item difficulty estimates in logits
SE(d_v), SE(d_y) standard errors of estimates in logits
d^*_v, d^*_y average difficulty estimates on common scale in logits
omit item omitted from common scale because it was not fitted in both calibrations
[a] used to calculate the translation constant
[b] item not fitted when calibrated

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

9

Some Alternative Approaches to the Improvement of Measurement in Education and Psychology: Fitting Latent Trait Models

Roderick P. McDonald

After some 70 years of research, the virtues and limitations of the linear common factor model, as a tool for the structural analysis of a battery of mental tests, are now reasonably well understood. In the well-known case that Spearman originally treated, we explain the covariation of a set of tests by supposing that they have linear regressions on a single variable—a 'common factor' or 'latent trait'—with residuals that are uncorrelated. The Spearman case is free from those problems of rotational and interpretational indeterminacy that have made some social scientists suspicious of factor analysis, and the model gives us a reasonable definition of unidimensionality or homogeneity for a set of quantitatively scored tests. That is, if the tests fit the single-factor model 'satisfactorily', we say that the battery is 'unidimensional' or 'homogeneous' in the clear sense of these terms that the common factor model provides. Given estimation of the parameters of the model by the method of maximum likelihood, we can obtain a statistical test for the unidimensionality hypothesis. At the same time, the residual covariance matrix supplies a nonstatistical but very reasonable basis for judging the extent of the misfit of the model to the data. In practice, the residuals are, we might argue, more important than the test of significance, since the unidimensionality hypothesis, like all restrictive hypotheses, must be false, and will be proved so by the chi-square test on a sufficiently large sample. If the residuals are small, the fit of the hypothesis can still be judged to be satisfactory.

It is possible to show, but the demonstration would take us too far afield, that there is no psychometric distinction to be made between a

Spearman common factor and a generic true-score as treated in Lord and Novick (1968). A Spearman factor analysis can therefore be made the basis for a considerable amount of test-theoretic analysis, including the assessment of generic reliability; that is, of generalizability in the sense of Cronbach et al. (1972). (See McDonald, 1978a.)

If by construction or by chance, the factor loadings of a set of factorially homogeneous quantitative tests are equal, the tests are essentially tau-equivalent in the sense of Lord and Novick (1968), and the common factor differs only trivially from a specific true score. The factor analysis then supplies a basis for the assessment of reliability in the classical sense of measurement error (whatever that really means).

In its origins, latent trait theory (latent structure analysis) was motivated by the recognition that linear common factor analysis could not be carried over from the quantitative test to the qualitative test-item (Lazarsfeld, 1950; Guttman, 1950). The central reason for this is that the regression curve of a binary item on any independent variable (observed or unobserved) represents the conditional probability of passing the item. (Here we use the word 'passing' for whatever is scored as the positive response, without intending any loss of generality.) Since the regression curve is a curve of conditional probabilities, it must therefore be bounded by zero and unity, and cannot be linear. To the extent that item characteristic curves — the regressions of the items in a test upon a latent trait — can be approximated by straight lines over the interval containing most of the examinees, we can justify the simple process of fitting the latent linear model, which is just the Spearman common factor model (Lazarsfeld, 1950; Torgerson, 1958; McDonald, 1967a), and we can tolerate the continuing practice of factor analysing product-moment correlations of binary items, the so-called phi coefficients. Although there is some evidence that this approximation is not nearly as bad in practice as we might expect from theory, concern about difficulty factors (see McDonald, 1965, 1967a; McDonald and Ahlawat, 1974), as well as the admitted theoretical inappropriateness of the linear model, has led to the introduction of appropriate models for binary data that are essentially counterparts of the Spearman case for quantitative variables.

In the Spearman case, given n variables y_j, $j = 1, \ldots, n$, we assume that there exists a single common factor or latent trait x, such that the regression curve is given by

$$E\{y_j \mid x = x_i\} = m_j + f_j x_i, \qquad (1)$$

and such that for any fixed value x_i of x, the variables are uncorrelated. It is reasonable to suppose that all users of this model, if questioned, would say that they intend the stronger assumption that for fixed x the variables are distributed independently. That is, users probably intend to assume the principle of local independence (Lazarsfeld, 1950; Anderson, 1959; McDonald, 1962). In practice it is both convenient and sufficient in general to test the weak implication that, when the common factor is partialled out, the residual covariances are zero; that is, to test the familiar implication that

$$\mathbf{C} = [E\{y_j - \mu_j)(y_k - \mu_k)] = \mathbf{ff}' + \mathbf{U}^2 \qquad (2)$$

where $\mathbf{f}' = [f_1, \ldots, f_n]$ and $\mathbf{U}^2$ is a diagonal matrix of residual variances. This is tantamount to ignoring possible information in the higher moments of the distribution of $\{y_1, \ldots, y_n\}$.

Lazarsfeld (1950), seeking a suitable counterpart of factor analysis for binary data, first explored the class of polynomial item characteristic curves (which in principle suffer the same difficulties as the linear item characteristic curve). To bypass technical difficulties, he then substituted the latent class model for the polynomial model. With this step the central idea of a distribution of the latent trait or traits over a continuum of any dimensionality is given up altogether. (See McDonald, 1967a.) Lawley (1943) and Finney (1952) independently introduced curves that are actually appropriate for the regression of a binary item upon an observed independent variable, such as, in Lawley's case, the total test score. Lord (1968) attributes the basis of modern item characteristic curve theory to Lawley (1943) but, on one interpretation, it seems to be Lord (1952) himself who first combined the probit curve (the normal ogive) with the principle of local independence to yield the normal ogive latent trait model: Birnbaum, 1957–1958, (see Lord and Novick, 1968) gave theory for the equivalent logistic model.

We can reasonably consider the normal ogive and logistic models as nonlinear counterparts of the Spearman model. This is immediately seen on writing the two-parameter versions of these models as

$$E\{y_j \mid x = x_i\} = N(m_j + f_j x_i) \qquad (3)$$

and

$$E\{y_j \mid x = x_i\} = \Psi[D(m_j + f_j x_i)] \qquad (4)$$

where

$$N(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-z^2/2} dz, \qquad (5)$$

the normal distribution function,

$$\Psi(t) = 1/(1 + e^{-t}), \qquad (6)$$

the logistic function, and D is a known constant, remembering that if y_j is a binary variable, coded unity for 'pass' and zero for 'fail', then

$$E\{y_j | x = x_i\} = P\{y_j = 1 | x = x_i\}. \tag{7}$$

Equations (3) and (4) are nonlinear counterparts and indeed nonlinear transformations of the Spearman model (1), with the transformations chosen so as to satisfy the bounds we require on the regression curves when they represent probabilities. As a consequence of these transformations, the combination of the assumed item characteristic curves with the principle of local independence no longer yields a simple covariance structure such as (2) for the relations between the variables.

Not surprisingly, fitting a nonlinear latent trait model proves more difficult than fitting the counterpart linear common factor model. In the common factor model—except for Lawley (1942) and McDonald (1979)—we treat the common factors as random independent variables, that is, random regressors. The test covariances yield all the information needed for estimating the parameters. If we are interested in factor scores (values of the latent traits of individual examinees), we estimate these for any examinee, whether from the original sample used to estimate the parameters of the model or not, quite independently of the estimation of the model parameters. In contrast, most proposals for fitting the normal ogive or logistic models treat the N latent trait values x_i, $i = 1, \ldots, N$, of the examinees in a sample, as parameters to be estimated simultaneously with the item parameters—usually m_j, f_j in (3) and (4). That is, we treat the latent trait as a fixed independent variable—a fixed regressor. The main exception (Bock and Lieberman, 1970) uses extremely costly numerical procedures and is not recommended by the authors for practical applications, admirable though it may be as a theoretical tour de force.

Before going on to an examination of the problem of fitting a latent trait model by conventional methods, we should note the special case of the logistic model (4) in which we write

$$E\{y_j | x = x_i\} = \Psi[D(m_j + fx_i)], \tag{8}$$

that is, we set every f_j equal to a common value f. This transformation of the case, that we have noted earlier to be that of essentially tau-equivalent tests, was proposed by Rasch and has been popularized recently by Wright and others. For certain purposes we will regard the normal ogive model (3) with equal f_j values as a version of the Rasch model also. It is of course indistinguishable from it. For many applications for which these models seem to have been intended, we must substitute

$$E\{y_j | x = x_i\} = g_j + (1 - g_j)N(m_j + f_j x_i), \tag{9}$$

and

$$E\{y_j | x = x_i\} = g_j + (1 - g_j)\Psi[D(m_j + f_j x_i)], \tag{10}$$

so that the models may be employed on multiple-choice items intended to measure abilities. The guessing parameters g_i could in principle be estimated along with the μ_i and the f_i, (if we are not using the simple Rasch model) together with the x_i values or, as in Lord (1968), they could be estimated independently, perhaps as the chance level, that is, the reciprocal of the number of options.

The literature on fitting latent trait models seems to be in a rather unsatisfactory state. It is a simple matter to write down the likelihood function and its first and second derivatives with respect to the parameters of a fixed regressors model. The second derivative matrix is very strongly patterned, allowing in principle minimization of (minus-log-times) the likelihood function by blocks of iterative steps, one for each set of parameters, that are essentially simple Newton-Raphson steps. From what is stated, and from what is not explained, by Lord (1968), Kolakowski and Bock (1970), and Wingersky and Lord (1973), it appears that investigators who have attempted to program what might seem to be an unusually simple minimization algorithm have had to deal with a large number of problems by trial and error, to the point where the reader cannot be sure just what has been programmed. I hope to be corrected, but there does not seem to be any published demonstration by Monte Carlo study that any of the programs for fitting the two-parameter model recovers the true values of the parameters within reasonable tolerance. Lord (1968) states that his method does not converge unless both the number of items and the number of examinees is large, and that otherwise values of f_i tend to increase without limit for some items. Wright (1977) conjectures that this must happen, and concludes that the two-parameter model therefore cannot be fitted to data. It does indeed seem that the simultaneous estimation of the item parameters f_i and person parameters x_i may strongly tend to run into difficulties of the kind noted by Lord and commented upon by Wright. (A similar problem in Lawley's (1942) fixed-regressors factor model is solved by the choice of a loss function in the form of a more appropriate f: "ction of likelihood—McDonald (1979)—but the present problem does not seem to yield an analogous treatment.) The introduction of the guessing parameters possibly makes the situation worse, especially if we attempt to estimate them rather than supply them as constants. We might also question the claims that have been made in favour of the Rasch model as free from difficulties in the methods used to fit its parameters, at least if the model is applied to multiple-choice items, since in the usual estimation procedures there is no provision for estimating the guessing parameters, and there is no reason to believe that the estimates of the other parameters of the model are unaffected by guessing.

Actually, the case for using maximum likelihood estimation in the two-parameter model begins to look less interesting when we note that a test of fit of the model does not seem to have been given, to go with the likelihood estimates of its parameters, and that 'good' properties of the maximum likelihood estimators have not actually been demonstrated for these models. Although it has been thought otherwise, this last remark may well apply to the Rasch model as treated by Wright and others. Measures of fit have been suggested for the Rasch model and conjectured to have a chi-square distribution (Wright and Panchapakesan, 1969). According to Wright and Mead (1977), however, simulation studies have shown that this distribution is 'not exactly correct'. Our own simulation studies suggest that it is not even approximately correct for the measures of fit in the OISE version of a program originally by Wright and Panchapakesan. It does seem that current methods for fitting these latent trait models lack a properly established statistical criterion for rejecting the model. Perhaps more importantly, they certainly lack criteria for regarding the fit as satisfactory, criteria analogous to the sizes of the residual covariances after fitting the linear common factor model. It is partly for this reason that some writers have stressed the need to test the unidimensionality of a set of items by some means, before actually fitting a model with a single latent trait. Hambledon et al. (1978) state that testing the assumption of unidimensionality takes precedence over other goodness-of-fit tests of a latent trait model, and that further research is needed to establish a proper procedure to test the dimensionality. Crude devices have been suggested, such as the examination of the eigenvalues of the item covariance or correlation matrix, but such procedures are not well founded. (See McDonald, 1981.)

If the analysis just given is correct, latent trait theory is in a problematic state, and one that is not without some historical irony. It was introduced because linear common factor analysis was recognized to be inadequate to supply a dimensional analysis for binary items. It has reached the point where, given the values of the parameters of a latent trait model, we know how to use them for a wide variety of test-theoretic purposes. Yet we still have to resort to a form of linear factor analysis for a crude test of unidimensionality, we still have reason to doubt the estimation procedures that have been proposed, and we still have no satisfactory statistical criterion for rejecting the model and no satisfactory criterion for regarding its fit as adequate.

McDonald (1967a, b) gave theory for nonlinear factor analysis, and numerical methods for fitting nonlinear regressions, in the form of polynomial functions, of quantitative tests or of binary items, on factors, that is, on latent traits. In contrast to item characteristic curves such as the normal ogive and logistic functions, which are nonlinear functions both of the latent traits and the item parameters, the polynomial item characteristic curves are nonlinear in the latent traits but linear in their coefficients (the item parameters). The immediate consequence is that the nonlinear factor model shares many of the simple algebraic properties of the linear common factor model. In particular it allows us to assess the adequacy of the fit of the model by examining residual covariances. In principle, therefore, nonlinear factor analysis supplies a general test of the unidimensionality or homogeneity of a set of binary items without the strong and false assumption of linear item characteristic curves that is implicit in the usual attempts to assess dimensionality prior to fitting a latent trait model. However, polynomial item characteristic curves share the defect of linear item characteristic curves that they are not bounded as required for probabilities. It might therefore seem unlikely that we could use nonlinear factor analysis in practice for this purpose, but shortly we will see that we can in fact do so under some conditions.

McDonald (1967a) sought to show that latent trait models such as the normal ogive model can be treated as special cases of nonlinear factor analysis by expressing the normal ogive curve as an infinite series whose terms are polynomials that are mutually orthogonal under the assumption that the latent trait has a normal distribution. If x has a normal distribution with mean zero and variance unity, then the normalized Hermite-Tchebycheff polynomials given by

$$h_p(x) = \frac{1}{\sqrt{p!}} (-1)^p e^{x^2/2} \frac{d^p}{dx^p} e^{-x^2/2}, \; p = 0, 1, 2, \ldots, \tag{11}$$

have mean zero, variance unity, and covariances zero. That is,

$$E\{h_p(x)\} = 0 \tag{12}$$

and

$$\begin{aligned} E\{h_p(x)h_q(x)\} &= 1, \; p = q, \\ &= 0, \text{ otherwise.} \end{aligned} \tag{13}$$

The first four orthogonal polynomials are given by

$$\begin{aligned} h_0 &= 1 \\ h_1(x) &= x \\ h_2(x) &= (x^2 - 1)/\sqrt{2} \\ h_3(x) &= (x^3 - 3x)/\sqrt{6}. \end{aligned}$$

Recalling that the first six moments of the normal distribution with mean zero and variance unity are $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 0$, $\mu_4 = 3$, $\mu_5 = 0$, $\mu_6 = 15$, we easily verify for example that

$$\begin{aligned} E\{h_1(x)h_2(x)\} &= E\{(x^3 - x)/\sqrt{2}\} \\ &= (\mu_3 - \mu_1)/\sqrt{2} \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} E\{[h_3(x)]^2\} &= E\{(x^6 - 6x^4 + 9x^2)/6\} \\ &= (\mu_6 - 6\mu_4 + 9\mu_2)/6 \\ &= 1, \end{aligned}$$

and similarly for the remainder. (These polynomials serve as a good classroom demonstration of the fact that, if two random variables are uncorrelated, they are not necessarily statistically independent, and indeed one can be a curvilinear function of the other.) Orthogonal polynomials such as these provide the building blocks for a curvilinear regression, in which the uncorrelated components supply additive variance. That is, instead of fitting a polynomial regression of some y on some x as

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + e \tag{14}$$

it is usually better to fit

$$y = b_0 + b_1h_1(x) + b_2h_2(x) + b_3h_3(x) + \ldots + e \tag{15}$$

because the terms in (15) are uncorrelated and supply contributions to the variance of y whose magnitude and significance can be assessed separately. The series can be terminated when all systematic variance has been captured.

Given orthogonal polynomials appropriate to the distribution of x (not necessarily the Hermite-Tchebycheff series (11)), the method of polynomial factor analysis introduced by McDonald (1967a, 1967b) amounts to recognizing that if we write nonlinear common factor models

$$y_j = a_{j0} + a_{j1}x + a_{j2}x^2 + \ldots + e_j, \tag{16}$$

or

$$y_j = b_{j0} + b_{j1}h_1(x) + b_{j2}h_2(x) + \ldots + e_j, j = 1, \ldots, n, \tag{17}$$

with uncorrelated residuals, then in the second version the polynomials behave just like orthogonal common factors. The technical problem of nonlinear factor analysis (which need not concern us here) is to discriminate between a model such as (17) and the parallel linear model

$$y_j = b_{j0} + b_{j1}x_1 + b_{j2}x_2 + \ldots + e_j, \tag{18}$$

both of which imply the covariance structure

$$\text{Cov}\{y_j, y_k\} = b_{j1}b_{k1} + b_{j2}b_{k2} + \ldots, j \neq k. \tag{19}$$

This is done by studying the distribution of factor scores in common factor space, to see if these lie within curved subspaces. (See McDonald, 1967b.)

We can use a kind of harmonic or Fourier analysis to approximate any

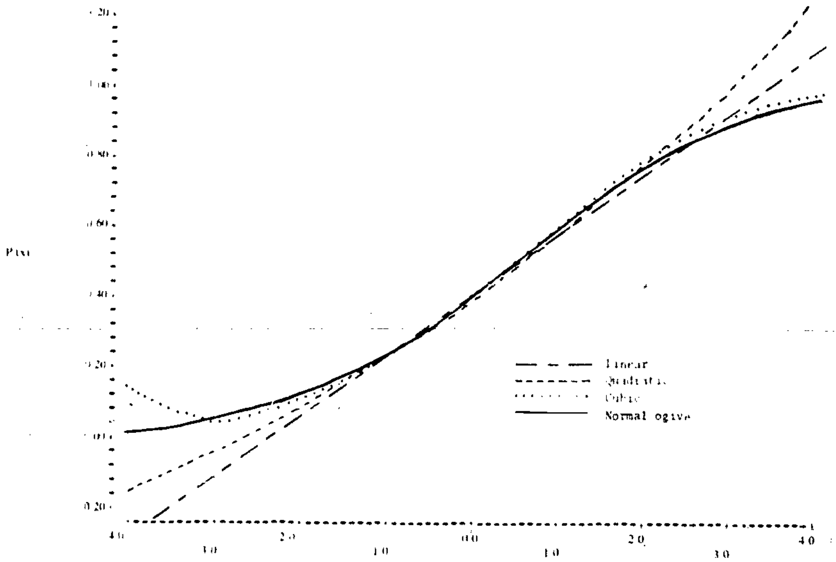


Figure 1 Normal Ogive ($\mu = 0.5$, $\sigma = 2.0$) with Polynomial Approximations

prescribed curve by a polynomial series. By making the series long enough, we can make the approximation as precise as we please over a finite range. For the remainder of this paper, it will be convenient to describe the normal ogive characteristic curve in the traditional way as $N(x; \mu_j, \sigma_j)$ where μ_j and σ_j are the mean and standard deviation of the cumulative distribution function $N(.)$, so that in (3)

$$f_j = 1/\sigma_j \tag{20}$$

and

$$m_j = -\mu_j/\sigma_j. \tag{21}$$

Figure 1 shows a normal ogive with $\mu_j = 0.5$, $\sigma_j = 2.0$. Superimposed upon it are the best-approximating linear, quadratic, and cubic curves, obtained by stopping at the second, third, and fourth term of the series

$$\hat{y}_j \doteq b_{j0} + b_{j1}x + b_{j2}(x^2 - 1)/\sqrt{2} + b_{j3}(x^3 - 3x)/\sqrt{6}. \tag{22}$$

The coefficients b_{j0}, b_{j1}, b_{j2}, b_{j3} are chosen to give a least-squares best fit of the polynomial curve to the normal ogive, weighted by the normal density function. That is, the coefficients are chosen to minimize

$$\Psi = E[(N(x; \mu_j, \sigma_j) - \sum_{p=0}^{r} b_{jp}h_p(x))^2],\ r = 0, 1, \ldots \tag{23}$$

McDonald (1967a) showed that if we express $N(x; \mu_j, \sigma_j)$ as the infinite series

$$N(x; \mu_j, \sigma_j) = \sum_{p=0}^{\infty} b_{jp} h_p(x) \tag{24}$$

where

$$b_{j0} = N(-\mu_j/\alpha_j) \tag{25}$$

and

$$b_{jp} = p^{-1/2} \alpha_j^{-p} h_{p-1}(\mu_j/\alpha_j) n(\mu_j/\alpha_j), \; p = 1, \ldots, \tag{26}$$

where $\alpha_j = (1 + \sigma_j^2)^{1/2}$ and $N(.)$ and $n(.)$ are the normal distribution and normal density functions, then every finite segment of the infinite series has the property that it minimizes (23) for the chosen number of terms. In particular, (26) yields

$$b_{j1} = \alpha_j^{-1} n(\mu_j/\alpha_j) \tag{27}$$

$$b_{j2} = \frac{1}{\sqrt{2}} \alpha_j^{-2} \frac{\mu_j}{\alpha_j} n(\mu_j/\alpha_j) \tag{28}$$

$$b_{j3} = \frac{1}{\sqrt{3}} \alpha_j^{-3} h_2(\mu_j/\alpha_j) n(\mu_j/\alpha_j). \tag{29}$$

The intended application of this theory was to fit the nonlinear factor model to a set of binary data by methods described in McDonald (1967b), up to the second or third degree, say. If the single-factor nonlinear model gave a reasonable account of the data we would then examine the distribution of the factor scores to see if it is normal, and examine the coefficients b_{jp} to see if their relationships were consistent with those required by (25) and (26). The equations (25) and (26) could then be solved for μ_j and σ_j. Unpublished work by McDonald and Ahlawat showed that reasonably precise estimates of the parameters of the normal ogive model could be obtained using this technique. (See also McDonald and Ahlawat (1974) for an account of difficulty factors in terms of this theory.)

Recent developments in the analysis of covariance structures (McDonald, 1978b; 1980) have made possible a more direct application of this theory, and the rest of this paper will focus upon the new method.

Because the representation (24) of the normal ogive is a linear combination of (random) orthogonal functions of the random variable x, it follows by a weak implication of the principle of local independence that

$$p_j = p\{y_j = 1\} = E\{y_j\} = E\{y_j^2\} = b_{j0}, \tag{30}$$

$$p_{jk} = p\{y_j = 1, y_k = 1\} = E\{y_j y_k\} = \sum_{p=0}^{\infty} b_{jp} b_{kp}, \; j \neq k, \tag{31}$$

where b_{j0} and b_{jp} are the functions of μ_j and σ_j given by (25) and (26). We can rewrite (30) and (31) in a familiar matrix form by defining $\mathbf{y}' = [1, y_1, \ldots, y_n]$, an $(n+1)$-component vector whose first component is unity, $\mathbf{B} = [b_{jp}]$, $j = 1, \ldots, n$; $p = 1, \ldots, \infty$, $\mathbf{b}' = [b_{10}, \ldots, b_{n0}]$, $\mathbf{p}' = [p_1, \ldots, p_n]$, and $\mathbf{P} = [p_{jk}]$. Equations (30) and (31) can then be expressed in the form

$$\mathbf{P}^* = \begin{bmatrix} 1 & \vdots & \mathbf{p}' \\ \cdots & \cdots & \cdots \\ \mathbf{p} & \vdots & \mathbf{P} \end{bmatrix} = \begin{bmatrix} 1 & \vdots & \\ \cdots & \cdots & \cdots \\ \mathbf{b} & \vdots & \mathbf{B} \end{bmatrix} \begin{bmatrix} 1 & \vdots & \mathbf{b}' \\ \cdots & \cdots & \cdots \\ & \vdots & \mathbf{B}' \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \\ & \mathbf{U}^2 \end{bmatrix} \qquad (32)$$

where $\mathbf{U}^2$ is the diagonal matrix whose jth diagonal element is

$$u_{jj}^2 = b_{j0} - \sum_{p=0}^{\infty} b_{jp}^2, \; j = 1, \ldots, n. \qquad (33)$$

The right member of (32) is formally the same as the structure implied by the orthogonal common factor model. In this case the column-order of **B**—the number of 'common factors'—is infinite, but the infinitely many elements of **B** are all functions of the $2n$ parameters μ_j, σ_j, $j = 1, \ldots, n$. That is, we have expressed the normal ogive model, which is nonlinear in its parameters and in the latent trait x, as a linear combination of infinitely many nonlinear functions of the latent trait, with coefficients that are nonlinear functions of the parameters of the model. Consequently the normal ogive model becomes a special case of the common factor model.

McDonald (1978b) has described a model for the analysis of covariance structures which allows higher order factor analysis of any order, with residual matrices of any prescribed structure. For the present application, the important property of the model is that the user can impose constraints on the matrices in it by making each element of each matrix a prescribed function of one or more 'fundamental' parameters—the parameters, that is, with respect to which the model is actually fitted. A program COSAN has been written for the model, and some applications are described in McDonald (1980). Program COSAN minimizes one of several loss functions with respect to the parameters of a given model, using a quasi-Newton method. For many purposes, the constraints on the model consist in setting certain elements of a factor loading matrix, or a residual or correlation matrix, equal to a constant (usually zero for simple structure ᴗ. orthogonality, and unity for a self-correlation) or constraining two or more elements to be equal, as in the work of Jöreskog (1970). In addition to these standard provisions of program COSAN, the user can write sub-routines of his own—usually very short and simple—if he wishes to prescribe special constraints upon the elements of the matrices in the model. It is therefore very easy to use program COSAN to fit the parameters of the normal ogive model by fitting the version of the orthogonal factor model in (32) to a sample counterpart of $\mathbf{P}^*$, where the elements of $\mathbf{b}$ and $\mathbf{B}$ are the prescribed functions of μ_j and σ_j, $j = 1, \ldots, n$. In practice, of course, we must truncate the matrix $\mathbf{B}$ to be of some finite column-order, and therefore we are in a sense fitting an approximate version of the normal ogive model. However, the coefficients b_{jp} rapidly diminish as p increases, and trial suggests both that terms beyond the cubic are negligible in magnitude and that including them would not improve the precision of estimation of the fundamental parameters of the model at all, even if it were to improve the fit slightly.

With $\mathbf{B}$ truncated to an $n \times 3$ matrix, we fit the model (24) to a sample matrix

$$\mathbf{S}^* = \begin{bmatrix} 1 & \mathbf{s}' \\ \mathbf{s} & \mathbf{S} \end{bmatrix}, \tag{34}$$

in which the jth component of $\mathbf{s}$, $\Sigma y_{ij}/N$, is the proportion of examinees in the sample passing item j, and the (j, k)th element of $\mathbf{S}$, $\Sigma y_{ij} y_{ik}/N$, is the proportion of examinees passing items j and k. We minimize the usual least-squares function

$$\theta = tr\{(\mathbf{P}^* - \mathbf{S}^*)^2\} \tag{35}$$

with respect to the $2n$ parameters μ_j, σ_j. By fixing the σ_j values to be equal to a common value σ, we may fit an equivalent of the Rasch model, minimizing (35) with respect to the μ_j and σ. If we fix $\sigma = 0$, we seek to fit the perfect scale, estimating the μ_j only. (See McDonald, 1967a.) We can introduce a guessing parameter by replacing the model with

$$P\{y_j = 1 \mid x = x_i\} = g_j + (1 - g_j)N(x_i; \mu_j, \sigma_j). \tag{36}$$

Correspondingly, (25) and (26) become

$$b_{j0} = g_j + (1 - g_j)N(-\mu_j/\alpha_j) \tag{37}$$

and

$$b_{jp} = (1 - g_j)p^{-1/2}\alpha_j^{-p}h_{p-1}(\mu_j/\alpha_j)n(\mu_j/\alpha_j),\ p = 1, 2, \ldots \tag{38}$$

We could then read in guessing parameters, possibly as the reciprocal of the number of options in multiple-choice items, or estimate them, in combination with any of the options for the σ_j values. To apply COSAN to these purposes, one library sub-routine of seven executable statements, to evaluate the normal density and distribution functions, and two special sub-routines, of simple logical structure, of 34 and 68 executable statements are needed. A program has also been written to generate normal ogive data on which to test the method. A program MESAMAX, the OISE version of a program by B. Wright and N. Panchapakesan for fitting the Rasch model was the only program available

Table 1 : Example 1: True $\sigma = 1.33$

True μ	10σ's estimated cubic fitted $N = 50\,000$	10σ's estimated cubic fitted $N = 3000$	10σ's estimated cubic fitted $N = 500$	1σ estimated cubic fitted $N = 500$	1σ estimated straight line fitted $N = 500$	RASCH program $N = 500$
1.00	1.000	−1.006	−0.976	−0.998	−0.989	−0.8997
0.80	0.816	−0.803	−0.604	−0.746	−0.740	−0.5805
0.60	0.593	0.641	−0.538	−0.582	−0.578	−0.5352
0.40	0.407	0.372	−0.424	−0.419	−0.416	−0.3576
0.20	0.197	0.169	−0.173	−0.210	−0.208	−0.0509
0.00	0.012	0.032	0.045	0.049	0.049	0.0250
0.20	0.198	0.164	0.366	0.355	0.354	0.3098
0.40	0.398	0.438	0.407	0.396	0.395	0.4718
0.60	0.582	0.655	0.587	0.591	0.588	0.7010
0.80	0.809	0.820	0.841	0.828	0.825	0.9162
min $\hat{\sigma}$	1.312	1.259	1.032			
max $\hat{\sigma}$	1.359	1.490	1.578	1.318	1.297	

for comparison with conventional methods for fitting latent trait models that gave believable results. This sets limits upon comparisons that can be made in two-parameter cases.

A large number of constructed examples have been run. Of these, three will be described, and other observations of the behaviour of the method will be briefly summarized.

Example 1: Three data-sets were generated, with sample sizes 50 000, 3000, and 500 respectively, whose true μ_i values are listed in the first column of Table 1. A common value of $\sigma = 1.33$ was employed for all items to enable a reasonable comparison with the available program for fitting the Rasch model. Each of the three resulting 11×11 raw product moment matrices (**S*** in (34)) was analysed four times by COSAN: (a) fitting ten μ_i values and a common σ value versus fitting ten μ_i values and ten σ_i values; (b) using the cubic approximation to the normal ogive model versus using the linear approximation, that is, deleting the columns of **B** containing coefficients b_{i2}, b_{i3}. Table 1 gives the estimates of the μ parameters for five of these analyses, as well as the estimates obtained by pro am MESAMAX applied to the raw data for sample size 500 and transformed for compatibility with the normal ogive representation.

These results illustrate observations that have been made from a wider range of analyses. The estimates of μ_i are not noticeably more precise when one σ is fitted than when they are fitted individually. That is, it is no

Table 2 Example 1: Raw Product Moments ($N = 3000$)

	1	2	3	4	5	6	7	8	9	10	11
1	1.000										
2	0.722	0.722									
3	0.689	0.544	0.689								
4	0.651	0.510	0.502	0.651							
5	0.589	0.482	0.458	0.439	0.589						
6	0.541	0.435	0.420	0.407	0.376	0.541					
7	0.493	0.403	0.393	0.375	0.344	0.320	0.493				
8	0.460	0.379	0.369	0.355	0.330	0.311	0.286	0.460			
9	0.402	0.333	0.327	0.309	0.295	0.266	0.254	0.240	0.402		
10	0.355	0.297	0.292	0.281	0.266	0.247	0.224	0.214	0.184	0.355	
11	0.320	0.267	0.264	0.248	0.232	0.221	0.211	0.201	0.175	0.157	0.320

Table 3 Example 1: Residuals, Cubic Approximation

	1	2	3	4	5	6	7	8	9	10	11
1	0.000										
2	0.000	0.000									
3	0.000	0.001	0.000								
4	0.001	−0.005	0.003	−0.000							
5	−0.002	0.007	−0.003	−0.000	0.000						
6	0.001	−0.001	−0.004	0.003	−0.000	−0.000					
7	0.000	0.001	0.001	0.001	−0.005	−0.003	−0.000				
8	0.001	−0.001	−0.002	0.000	−0.001	0.004	−0.000	−0.000			
9	−0.001	0.000	0.002	−0.001	0.005	−0.003	0.003	0.000	−0.000		
10	−0.002	−0.000	0.001	0.003	0.005	0.005	−0.002	−0.002	−0.007	0.000	
11	0.002	−0.001	0.001	−0.004	−0.005	0.001	0.005	0.003	0.001	−0.002	−0.000

Table 4 Example 1: Residuals, Linear Approximation

	1	2	3	4	5	6	7	8	9	10	11
1	0.000										
2	−0.001	0.000									
3	−0.001	0.002	0.000								
4	−0.000	−0.004	0.004	0.000							
5	−0.002	0.008	−0.002	−0.000	0.000						
6	0.001	−0.001	−0.004	0.003	−0.000	0.000					
7	0.001	0.000	0.001	0.001	−0.005	−0.003	0.000				
8	0.001	−0.002	−0.003	0.000	−0.001	0.003	−0.000	−0.000			
9	−0.000	−0.001	0.002	−0.002	0.004	−0.004	0.003	0.000	0.000		
10	−0.001	−0.001	0.000	0.002	0.004	0.005	−0.002	−0.002	−0.006	−0.000	
11	0.002	−0.002	0.000	−0.005	−0.006	0.001	0.005	0.004	0.002	−0.000	0.000

Table 5 Example 1: Coefficients of Cubic

b_0	b_1	b_2	b_3
0.722	0.196	−0.048	−0.018
0.689	0.217	−0.047	−0.025
0.650	0.223	−0.037	−0.028
0.591	0.239	−0.024	−0.035
0.540	0.235	−0.010	−0.033
0.492	0.238	0.002	−0.034
0.459	0.247	0.011	−0.039
0.403	0.216	0.021	−0.026
0.357	0.208	0.030	−0.023
0.318	0.205	0.040	−0.022

more difficult to fit the two-parameter model than to fit the one-parameter model, using this method. The estimates are not noticeably more precise when the cubic approximation is used than when the linear approximation is employed. Table 1 gives the minimum and maximum values of $\hat{\sigma}_r$ when these are estimated as ten distinct parameters, and the value of $\hat{\sigma}$ when a common parameter is estimated.

Table 2 gives the 11×11 raw product-moment matrix for sample size 3000. Table 3 gives the corresponding residual matrix using the cubic approximation, while Table 4 gives the residual matrix using the linear approximation (with ten σ_i values estimated in both cases). Table 5 gives the estimated values of the coefficients b_{j0} in **b** and b_{j1}, b_{j2}, b_{j3} in **B**, corresponding to Table 3. (These are estimated parametric functions of $\hat{\mu}_i$ and $\hat{\sigma}_i$.) Since they are coefficients of orthonormal polynomials, they behave like loadings on orthogonal common factors, showing directly that the data are actually accounted for to a good approximation by a linear model. The approximation is only slightly improved by the addition of the quadratic and cubic terms. Terms beyond the cubic would almost certainly be quite negligible. At the same time, fitting the cubic approximation can be recommended on the basis of a general observation, not illustrated by the comparison of Table 3 and Table 4, that usually the residuals from the cubic approximation are just sufficiently smaller to constitute slightly better evidence that the data are unidimensional and adequately described by the normal ogive model.

Example 2: A data-set, of sample size 3000, was generated, to consist of 50 items, combinations of five σ_i values and ten μ_i values, as shown in the margins of Tables 6 and 7. These contain the estimates by COSAN (using the cubic approximation) of the μ_i values and the σ_i values respectively. Inspection suggests that the precision of the estimates of the μ_i values is not noticeably affected by the size of μ_i itself, or the size of σ_i, and that the precision of the estimates of the σ_i values, while approximately proportional to σ_i, is not noticeably affected by the size of μ_i.

Example 3: Again with a sample size of 3000, and with a common σ value of 1.7, 20 binary items were simulated in ten pairs, with μ_i values, repeated, as in the previous example, but with the parameter g_i in (36) introduced and set to 0.2 for the first member of each pair, and 0.5 for the second member. It is as though each odd-numbered item is a multiple-choice item with five options, while the following even-numbered item is an otherwise equivalent true/false item. The first column of Table 8 gives the true μ_i values. The second contains the estimates by COSAN, with the cubic approximation, and reading g_i values alternately of 0.2 and 0.5 and holding them fixed, as we might do from knowledge of the item formats. The third column contains COSAN estimates assuming that there is no effect of guessing on the data, that is, setting each g_i value to zero. The

Table 6 Example 2: μ_i Estimates ($N = 3000$)

			True σ_i		
True μ_i	2.27	2.00	1.72	1.52	1.33
-1.00	-1.042	-0.997	-1.056	-1.056	-1.033
-0.80	0.800	-0.865	-0.756	-0.756	-0.757
-0.60	-0.584	-0.525	-0.529	-0.595	-0.541
-0.40	-0.456	-0.358	-0.398	-0.366	-0.381
0.20	-0.316	-0.236	-0.207	-0.277	-0.245
0.00	0.050	0.040	-0.038	0.013	-0.040
0.20	0.207	0.236	0.180	0.239	0.252
0.40	0.378	0.478	0.405	0.319	0.484
0.60	0.489	0.631	0.596	0.593	0.605
0.80	0.830	0.759	0.847	0.879	0.817

Table 7 Example 2: σ_i Estimates ($N = 3000$)

			True σ_i		
True μ_i	2.27	2.00	1.72	1.52	1.33
1.00	2.257	2.062	1.768	1.598	1.315
0.80	2.289	2.057	1.680	1.569	1.348
0.60	2.332	2.059	1.606	1.471	1.316
0.40	2.231	1.845	1.818	1.585	1.321
0.20	2.504	2.050	1.687	1.747	1.348
0.00	2.620	1.903	1.995	1.452	1.338
0.20	2.123	2.141	1.617	1.579	1.458
0.40	2.078	1.990	1.954	1.534	1.502
0.60	2.241	2.405	1.668	1.589	1.339
0.80	2.419	1.928	1.793	1.652	1.539

fourth column contains the estimates of the μ_i values obtained from MESAMAX, which of course makes no provision for guessing. It is clear that the effects of guessing in multiple-choice items must be allowed for in the analysis. The method of Wright and Panchapakesan yields quite unacceptable estimates of the difficulty parameters in the presence of guessing, as does the COSAN method when the guessing parameters are assumed to be zero. When the guessing parameters are treated as known (as in Lord, 1968), the COSAN method gives good estimates of the other parameters. Attempts to use COSAN to estimate guessing parameters, as well as the other parameters of the model, have run into difficulties requiring further research. It seems likely that the nonlinear factor model (17) will have to be fitted directly to raw data if the present method is to yield estimates of the three-parameter model.

Table 8 Example 3: Estimates

True μ_r	COSAN estimate with true g_r	COSAN estimate with $g_r = 0$	RASCH estimate
−1.00	−1.073	−1.962	−0.392
−1.00	−0.934	−2.648	−0.889
−0.80	−0.887	−1.734	−0.220
−0.80	−0.768	−2.480	−0.771
−0.60	−0.593	−1.380	0.016
−0.60	−0.602	−2.315	−0.648
−0.40	−0.518	−1.291	0.085
0.40	−0.434	−2.153	−0.538
−0.20	−0.208	−0.927	0.332
−0.20	−0.196	−1.932	−0.388
0.00	−0.008	−0.697	0.481
0.00	0.012	−1.745	−0.267
0.20	0.160	−0.507	0.606
0.20	0.137	−1.637	−0.203
0.40	0.431	−0.209	0.802
0.40	0.222	−1.565	−0.143
0.60	0.557	−0.073	0.885
0.60	0.585	−1.273	0.034
0.80	0.800	0.180	1.052
0.80	0.853	−1.075	0.178

While a more systematic Monte Carlo study is desirable, the examples serve to show that we can indeed fit the normal ogive model, in a one-, two-, or three-parameter version (the latter with known guessing parameters), by a program for the analysis of covariance structures, with reasonably satisfactory results. The fact that we can do this at all illustrates an essential unity of psychometric theory for 'quantitative' tests and 'qualitative' items that we might easily lose sight of while concentrating on the details of fitting the models by the conventional statistical procedures. The fact that we can do this reasonably well suggests that the technique deserves further exploration, as possibly a useful one at least for some data sets. Already it is clear that reasonably precise estimates can be obtained over a range of sample sizes, from about the smallest we should ever use for such work to indefinitely large. The obvious advantage of the method is that it supplies a measure of the goodness of fit of the model in the familiar form of a residual matrix, and the sum of squares of its elements, and it does not require a prior examination of the dimensionality of the data. An obvious limitation of the method is its assumption that the latent trait has a normal distribution. There seems to be less willingness on the part of investigators to assume normality of the latent trait underlying a set of items than to make the same assumption for the common factor underlying a set of tests. This would be because it is easier to exert control over the distribution in the former case than in the latter, by the design of the items. (We can skew the distribution of a factor by choosing tests that are too difficult or too easy, or flatten its distribution by choosing a wide range of test difficulties. Since tests are item sums, the cases cannot be very different.) If the method suggested seems worth using, the user can in principle design his item set to have a close-to-normal distribution of the latent trait. Since users may not want to do this, further research will include an investigation of the robustness of the method against violations of the normality assumption. If it is not sufficiently robust for general use, then it will become worthwhile to repeat the theoretical work leading to equations (25) and (26), using a more general distribution of the latent trait. Because this work had been done originally for the normal ogive model, discussion in this paper has been confined to that case, just to save the rethinking that would be necessary, even to state the corresponding results for the equivalent logistic model.

REFERENCES

Anderson, T. W. Some scaling models and estimation procedures in the latent class model. In O. Grenander (Ed.), *Probability and statistics.* (The Harold Cramer Volume). New York: Wiley, 1959, 9–38.

Bock, R. D. and Lieberman, M. Fitting a response model for *n* dichotomously scored items. *Psychometrika*, 1970, **35**, 179–97.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioural measurements: Theory of generalizability for scores and profiles.* New York: Wiley, 1972.

Finney, D. J. *Probit analysis.* Cambridge: Cambridge University Press, 1952.

Guttman, L. Chapters 2, 3, 6, 8, 9. In S. A. Stouffer et al., *Measurement and prediction.* Princeton, NJ: Princeton University Press, 1950.

Hambledon, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R. and Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, **48**, 467–510.

Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika*, 1970, **57**, 239–51.

Kolakowski, D. and Bock, R. D. *A Fortran-IV program for maximum likelihood item analysis and test scoring: Normal ogive model.* (Research Memorandum No. 12, 1970). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1970.

Lawley, D. N. Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh*, 1942, **62**, 176–85.

Lawley, D. N. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, **61**, 273–87.

Lazarsfeld, P. F. Chapters 10, 11. In S. A. Stouffer et al., *Measurement and prediction.* Princeton, NJ: Princeton University Press, 1950.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, 1952, **7**.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, **28**, 989–1020.

Lord, F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

McDonald, R. P. A note on the derivation of the general latent class model. *Psychometrika*, 1962, **27**, 203–6.

McDonald, R. P. Difficulty factors and nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology*, 1965, **18**, 11–23.

McDonald, R. P. Nonlinear factor analysis. *Psychometric Monographs*, 1967, **15**. (a)

McDonald, R. P. Numerical methods for polynomial models in nonlinear factor analysis. *Psychometrika*, 1967, **32**, 77–112. (b)

McDonald, R. P. Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement*, 1978, **38**, 75–79. (a)

McDonald, R. P. A simple comprehensive model for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 1978, **31**, 59–72. (b)

McDonald, R. P. The simultaneous estimation of factor loadings and scores. *British Journal of Mathematical and Statistical Psychology*, 1979, **32**, 212–28.

McDonald, R. P. A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. *British Journal of Mathematical and Statistical Psychology*, 1980, **33**, 161–83.

McDonald, R. P. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1981, **34**, 100–17.

McDonald, R. P. and Ahlawat, K. S. Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, 1974, **27**, 82–99.

McDonald, R. P. and Burr, E. J. A comparison of four methods of constructing factor scores. *Psychometrika*, 1967, **32**, 381–401.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Wingersky, Marilyn S. and Lord, F. M. *A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses*. (RM-73-2). Princeton, NJ: Educational Testing Service, 1973.

Wright, B. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, **14**, 97–116.

Wright, B. and Mead, R. J. *BICAL: Calibrating items and scales with a Rasch measurement model*. (Research Memorandum No. 13). Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1977.

Wright, B. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, 29, 23–48.

ACKNOWLEDGMENT

The author would like to thank C. Fraser for the programming and numerical work in this research.

APPENDIX

Scoring an examinee

A preliminary investigation has not revealed any advantages of the polynomial approximation over the conventional treatment when it comes to scoring an examinee, i.e. obtaining an estimate of his latent trait given his item scores. However, for completeness, some remarks can be made about interesting parallelisms between equations for the estimation of a latent trait from binary data and equations for the estimation of a common factor from test scores. The one possibly useful result to emerge so far from the examination of these parallelisms is an expression that contains an estimate in closed form of the latent trait, based upon the linear approximation. Whether it is close enough to the conventional estimate will need to be investigated by a Monte Carlo study.

In the usual treatment of the common factor model, with random common factors, we first fit the parameters of the model (factor loadings and uniquenesses) to the covariance matrix of a 'calibration' sample. For any examinee in the population that the model purports to describe, we estimate his factor scores as linear combinations of his test scores. In the Spearman case (1), the Weighted Least Squares (WLS) formula of Bartlett,

$$\tilde{x} = \left[\sum_{i=1}^{n} \frac{f_i^2}{1-f_i^2}\right]^{-1} \sum_{i=1}^{n} \frac{f_i}{1-f_i^2}(y_i - m_i), \tag{A1}$$

minimizes the sum of squares of the given examinee's n residuals

$$e_i = y_i - m_i - f_i x, \tag{A2}$$

weighted by the reciprocal of the variances of these residuals in the population. That is, it minimizes

$$\phi_1 = \sum_{i=1}^{n} \frac{e_i^2}{\mathrm{Var}\{e_i\}} = \sum_{i=1}^{n} \frac{(y_i - m_i - f_i x)^2}{1-f_i^2}. \tag{A3}$$

If we assume that each residual has a normal distribution, then (A1) is also the maximum likelihood estimator of x. The corresponding Unweighted Least Squares (ULS) estimator

$$\hat{x} = \left[\sum_{i=1}^{n} f_i^2\right]^{-1} \sum_{i=1}^{n} f_i(y_i - m_i), \tag{A4}$$

minimises the sum of the squares of the given examinee's residuals without taking account of the residual variances. That is, it minimizes

$$\phi_2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - m_i - f_i x)^2. \tag{A5}$$

(Scoring formulae (A1) and (A4) are respectively Method 2 and Method 1 as discussed by McDonald and Burr, 1967.)

The condition for a minimum of ϕ_1 in (A3) may be written as

$$\sum_{j=1}^{n}\left\{\left(\frac{f_j}{1-f_j^2}\right)(y_j-m_j-f_jx)\right\}=0, \qquad \text{(A6)}$$

which on rearrangement gives (A1).

Writing (1) as

$$\hat{y}_j=m_j+f_jx, \qquad \text{(A7)}$$

whence

$$\frac{\partial\,\hat{y}_j}{\partial x}=f_j, \qquad \text{(A8)}$$

and noting that

$$\text{Var}\{e_j\}=\text{Var}\{y_j\,|\,x\}, \qquad \text{(A9)}$$

we can rewrite (A6) as

$$\sum_{j=1}^{n}\left[\frac{\partial\hat{y}_j}{\partial x}\Big/\ \text{Var}\{y_j\,|\,x\}\right](y_j-m_j)=0, \qquad \text{(A10)}$$

or as

$$\sum_{j=1}^{n} w_je_j=0, \qquad \text{(A11)}$$

where

$$w_j=\frac{\partial\hat{y}_j}{\partial x}\Big/\ \text{Var}\{y_j\,|\,x\}. \qquad \text{(A12)}$$

That is, we choose x such that the weighted sum of the residuals becomes zero, with weights that consist of the slope of the regression of each y_j on x divided by the conditional variance. In the linear common factor model both the terms in the weight w_j are independent of x, one because the model is linear, and the other because (in the usual treatment of the model) we assume that the residuals are homoscedastic.

Now suppose we have any latent trait model with a single latent trait x and known parameters in the item characteristic curve $P_j(m_j+f_jx)$. We write $Q_j=1-P_j$. We wish to estimate x for a given examinee with binary item scores $y_1, \ldots, y_n$. By well-known theory, the likelihood equation is given by

$$\sum_{j=1}^{n}\left[\frac{\partial P_j}{\partial x}\Big/\ P_jQ_j\right](y_j-P_j)=0. \qquad \text{(A13)}$$

Since, in the case of binary data,

$$\mathrm{Var}\{y_i \mid x\} = P_i Q_i, \tag{A14}$$

equation (A13) is formally the same as equation (A10). That is, in any latent trait model with a single latent trait and known item parameters, to obtain the maximum likelihood estimate of an examinee's score we choose a score such that the weighted sum of his residuals becomes zero, if weighted by the slope of the regression of each y_i on x divided by the conditional variance. In contrast to the linear case, however, both the regression slope and the conditional variance are in general functions of x.

In the special case where the n item characteristic curves $P_i(x)$ are identical, equal to P, say, and P has an inverse function P^{-1}, it is well known that (A13) can be solved for x in closed form. (See Birnbaum in Lord and Novick, 1968, pp. 458-9.) This follows because we may then write (A13) as

$$\left[\frac{\partial P}{\partial x} / PQ\right] \sum_{i=1}^{n} (y_i - P) = 0 \tag{A15}$$

whence

$$\sum_{i=1}^{n} y_i = nP(x) \tag{A16}$$

so that

$$x = P^{-1}\left[\frac{1}{n} \sum_{i=1}^{n} y_i\right]. \tag{A17}$$

Again following Birnbaum we note that the logistic function $\Psi(t)$ in (6) uniquely has the property that

$$\frac{\partial \Psi}{\partial t} = \Psi(t)[1 - \Psi(t)] \tag{A18}$$

whence it follows that

$$\frac{\partial \Psi[D(m_i + f_i x)]}{\partial x} = \{(\Psi[D(m_i + f_i x)])(1 - \Psi[D(m_i + f_i x)])\} Df_i, \tag{A19}$$

independent of x, so that in this case (A13) becomes

$$\sum_{i=1}^{n} f_i(y_i - \Psi[D(m_i + f_i x)]) = 0 \tag{A20}$$

or

$$\sum_{i=1}^{n} f_i \Psi[D(m_i + f_i x)] = \sum_{i=1}^{n} f_i y_i, \tag{A21}$$

which is a way of expressing the fact that the weighted item sum

$$\sum_{i=1}^{n} f_i y_i$$

is a sufficient statistic for x for a given examinee in the two-parameter logistic model if the values of the f_i are known. In the special case of the Rasch model, in which the coefficients f_i have a common value, it follows that the unit-weighted sum

$$\sum_{i=1}^{n} y_i$$

is then a sufficient statistic for x. This latter property might be considered useful in some practical applications. However, it is primarily if we choose to fit the fixed regressors version of the model, simultaneously estimating the item parameters and the latent traits in the calibration sample, that this fact gives the one-parameter model an advantage over those models that realistically allow for guessing and for the fact that, in general, items measuring a given trait will not measure it equally well.

If we wish to apply the Rasch model to the measurement of ability by means of multiple-choice items, it would seem by Example 3 given earlier that we must introduce a guessing parameter. If we do so, the likelihood equation does not yield a counterpart of (A19), whether or not the f_i values are supposed equal. That is, the attempt to apply the model to multiple-choice items by the introduction of a guessing parameter destroys the sufficiency that has been regarded as an important property of the logistic model (and destroys other properties of the Rasch model that have been regarded as its important special characteristics). Perhaps more importantly, whether or not a single function of the examinee's responses is a sufficient statistic for his x, we cannot in general solve the likelihood equation in closed form.

The remarks in this appendix arose out of a tentative exploration of the problem of scoring the examinee in terms of the polynomial approximation model that was introduced in the body of the paper. The hope was that the nice properties of the Spearman linear case, including closed form, might carry over to this problem. Such a hope was quickly seen to be unfounded. In particular, an attempt to substitute the polynomial model in (A13), even just the linear approximation, yields seemingly intractable expressions. One result from this exploration may be of value and is therefore perhaps worth reporting.

If we accept the evidence given earlier that the linear model given by the first two terms of the polynomial series (24) is in general a good approximation to the normal ogive model, we first consider substituting this in the condition (A13) for a WLS, i.e. ML solution, to yield

$$\sum_{i=1}^{n} [b_{i1}/(b_{i0} + b_{i1}x)(1 - b_{i0} - b_{i1}x)](y_i - b_{i0} - b_{i1}x) = 0. \qquad \text{(A22)}$$

It seems fairly obvious that the linear approximation does not yield a solution in closed form. In desperation, with some, but not much, theoretical justification, we consider instead applying directly the ULS expression (A4), which yields

$$\hat{x} = \frac{\Sigma b_{j1}(y_j - b_{j0})}{\Sigma b_{j1}^2}, \tag{A23}$$

that is, by (25) and (27)

$$\hat{x} = \frac{1}{k} \sum_{j=1}^{n} \frac{1}{\alpha_j} n\left(\frac{-\mu_j}{\alpha_j}\right)\left[y_j - N\left(\frac{-\mu_j}{\alpha_j}\right)\right], \tag{A24}$$

where

$$k = \sum_{j=1}^{n} \left[\frac{1}{\alpha_j} n\left(\frac{-\mu_j}{\alpha_j}\right)\right]^2. \tag{A25}$$

This closed-form linear least squares estimate of x is not without a certain plausibility of expression. By theory given in Lord and Novick (1968, pp. 377-8), the quantity $1/\alpha_j$ is the correlation between x and y_j, and is a measure of the discriminating power of the item, while the quantity $N\{-\mu_j/\alpha_j\}$ is the proportion of examinees passing the item. In (A24), the contribution of an item to the estimate is weighted in three reasonable ways. First the item is weighted proportionally to its discriminating power, $1/\alpha_j$. Second, the item is weighted by $n\{-\mu_j/\alpha_j\}$, so that greater weight is given to items near the mean of the distribution of x than to items further away in either direction. Third, if the item is difficult, it gives a larger absolute value to the (positive) contribution from passing the item than to the (negative) contribution from failing it, and conversely for an easy item. It is conjectured from the form of (A24) that x will prove an acceptable closed-form estimate of the examinee's latent trait, given a reasonable number of items. The approximation involved should be best at the middle of the distribution. At the extremes, corresponding to a total test score of zero or n, where the maximum likelihood estimate is infinite, x cannot be correct. But this is not necessarily a disadvantage.

The Improvement of Measurement in Education and Psychology
Edited by Donald Spearritt
Copyright © ACER 1982

10

A Perspective on the Seminar

Donald Spearritt

It was recognized in the planning of the invitational seminar that the authors of the invited papers would examine the contributions and potentialities of latent trait measurement procedures at a number of levels. Some papers would emphasize the place of latent trait procedures within the general stream of the theory and practice of measurement in education and psychology. Others would emphasize theoretical issues in latent trait measurement which have arisen in the course of finding solutions to practical problems. Some would be more directly concerned with demonstrating practical applications of latent trait models in large-scale educational testing and in particular areas of long-standing interest in the development of tests in education and psychology. With the expected diversity in the papers, it was anticipated that it would be a useful exercise to draw together the main themes of the seminar and to assess its contribution to the improvement of measurement in education and psychology. This task fell to the chairman of the seminar. The approach taken has been to consider first the trends raised in Thorndike's opening address, and then to reflect on the main themes.

In his introductory address, Thorndike provided the seminar participants with an excellent overview of the origins and broad trends of psychometric theory and practice over the past 75 years. His references to Binet and Simon and Spearman are a reminder that the notion of latent traits has been in use for a long time, though the conception of latent traits or underlying abilities in the Spearman model is rather different from that coming under the rubric of latent trait theory today. He indicates why and how two of the main streams of psychometric theory were developed—the theory of measurement error with its notions of true score and reliability, and the theory of the organization of human abilities, both deriving in large measure from Spearman's early work. Lest we hasten to demolish the old temples too quickly, he gives us a timely reminder that the classical measurement model was responsible for producing a body of useful knowledge about tests. He also notes the emergence of an alternative notion of true score in more recent times, in which true score represents the universe of behaviour that would be approached if the number of relevant test tasks were increased without limit. But neither this notion nor the components-of-variance model to which it leads received much attention in the seminar.

The pervasive influence of multiple-choice items on testing theory and practice is also noted, as is the development of indices for item analysis purposes, viz. item difficulties or facilities, and item discrimination indices. Despite their shortcomings, these indices have influenced educational and psychological measurement at the workface as well as in the research laboratory.

It is enlightening to have had Thorndike's perspective on what he sees as the two major competing models to interpret a test score, that is, the domain sampling model and the latent trait model. While the seminar was largely concerned with the latter, we are reminded that a very substantial amount of work has been done in recent years on defining the universe of behaviour to which we wish to generalize from our tests, on criterion referenced tests and on testing for mastery. Though the domain sampling model is not without its problems, it represents an important aspect of educational measurement which must continue to be explored.

Thorndike's concept of the latent trait model as a vertical dimension on which the individual person is to be located is a useful one. There are, as he notes, some real difficulties in conceptualizing some aspects of educational achievement in terms of the latent trait model, though setting up dimensions such as 'competence in history' may go some way towards meeting these difficulties.

Latent trait models, the aspect of educational and psychological measurement which forms the main subject of this conference, were considered in the final section of Thorndike's paper. He distinguishes two schools of thought – the 'one-parameter' school represented in the Rasch approach, and the 'three-parameter' school led largely by Lord. The pros and cons of these approaches were argued by Thorndike and debated in greater detail in subsequent papers in the seminar.

In their own way, the individual papers have each made a contribution to the theory and or practice of latent trait measurement. But what has been the contribution of the seminar to the field, considering the set of papers in toto? One convenient way of making this assessment is to formulate some fundamental questions and to see what light has been thrown upon these by the various papers.

WHICH OF THE LATENT TRAIT MODELS GIVES THE BEST FIT TO ITEM DATA?

This has been a controversial question for some time. The contest has been largely between the Rasch one-parameter model, involving item difficulty only, and the three-parameter model, involving item discrimination indices and guessing parameters also. As Lord (1970) demonstrates, good fits appear to be obtainable with the three-parameter model, which takes account of the differing slopes of item characteristic curves and the level of the lower asymptote, as well as item difficulties.

On this criterion, the three-parameter model, not unexpectedly, has the advantage. But it is relevant to ask, as Thorndike does, whether such a complex model is really required. Are its advantages outweighed by the need for substantial computing facilities and for large samples for item tryouts? With respect to these questions, McDonald's study provides some promising findings. His results suggest that precise estimates of ability can be obtained with sample sizes as low as 500, which is about a desirable minimum value of *N* for item analysis studies.

DOES THE RASCH ONE-PARAMETER MODEL GIVE A SUFFICIENTLY GOOD FIT TO ITEM DATA?

From a practical point of view, the Rasch model obviously has the advantage of being simple to operate and requiring smaller tryout samples of persons. But Thorndike indicates that the model provides only a rough fit to the data in some cases, depending on the differences among the values of item discrimination indices for a set of items and the extent to which guessing is involved in students' selection of answers. He sees the Rasch model as being rather more successful with constructed-response items than with multiple-choice items, though with the qualification that carefully selected multiple-choice items are likely to provide good fits. Choppin recognizes that the model is designed to give an approximate rather than an exact representation of data, but argues on the basis of his extensive experience with the model in studies carried out by the National Foundation for Educational Research in England and Wales that the model is robust with respect to violations of its underlying assumptions, and presents empirical evidence to support this argument. Working in the area of cognitive development, however, in which ability is likely to change over a period of time, Keats shows that a one-parameter model is inadequate. A two-parameter model of cognitive development which uses as individual difference parameters both the asymptotic value of a person's ability and the rate at which he approaches that level gives a rather better fit to the data than a one-parameter model involving some index of IQ.

Choppin notes that one of the suggested applications of the Rasch model involves the identification of responses which are 'lucky guesses', and the editing out of the items which produce such responses. This type of editing would have the effect of improving the fit to the one-parameter model. It would seem highly desirable in such applications to seek some independent verification of the statistical aberrations. Some interviewing of students about their lucky guess responses would indicate whether the editing was justified.

The question at issue concerns the robustness of the Rasch model. How far can the data depart from the model before leading us to draw incorrect inferences? Arguments about the relative virtues of the Rasch model and the three-parameter logistic model are reminiscent of other controversies over the last three decades concerning the violation of assumptions underlying statistical tests, for example, the robustness of the F test in analysis of variance. In such controversies, it has often been found that the models allow considerable relaxation in their underlying assumptions before they begin to support false inferences. It would not be surprising to find that the Rasch model exhibited a similar degree of robustness.

There is a further question that may be asked in considering the goodness of fit of the Rasch model. If better fits to a set of data are possible, does this mean that less good fits must be discarded, even if they take a fraction of the time to obtain and provide a satisfactory approximation to the questions being asked of the data? While further empirical studies will be of assistance in answering this question, it would seem reasonable to draw the tentative conclusion that the Rasch model is a satisfactory model for estimating item and person ability parameters, unless its applicability to a set of items is obtained at the expense of discarding too many items.

ARE EXISTING TESTS OF FIT FOR LATENT TRAIT MODELS SATISFACTORY?

This question was taken up by Douglas and McDonald. Both authors regard the existing chi-square tests of fit as unsatisfactory. They are more likely to yield a significant non-fit with increase in sample size.

Douglas urges that the present approximate tests of fit be used with caution. One of the advantages of the generic Rasch model which he derives in his paper by means of conditional inference approaches is that it leads to a test of fit based on likelihood functions, though the applicability of the test is limited by numerical analysis problems. Douglas notes that an exact test of fit of data to a Rasch model is theoretically possible, through the use of conditional inference procedures free of all parameters in the model. Such a test has still to be developed, but he sees a promising line of development in the approaches taken by Agresti.

McDonald's paper is largely concerned with improving methods of fitting, and testing the fit of, latent trait models. He notes that difficulties such as lack of convergence have been experienced with programs for fitting latent trait models which involve the simultaneous estimation of item parameters and person parameters and that the difficulties have been magnified if guessing parameters have also been estimated. He doubts that the Rasch model is free of these difficulties. He questions the appropriateness of the chi-square approach, since no account is taken of the size of the residuals. He concludes that satisfactory criteria for testing the fit of latent trait models have been lacking, and notes that, even when a prior test of unidimensionality of the data has been made, it has relied on a linear factor analysis of a non-linear set of data.

McDonald's use of non-linear factor analysis is a promising approach to the fitting of latent trait models, since it avoids the false assumption of linear item characteristic curves. He shows that the normal ogive model, for example, can be expressed through orthogonal polynomials as a linear combination of nonlinear functions of the latent trait. Constraints on the model can be introduced into the program (COSAN) used for fitting the model, a particular set of constraints providing the equivalent of the Rasch model. He demonstrates that the program provides a satisfactory fit to the normal ogive model in a one-, two-, or three-parameter form for a range of data sets. Since this approach requires no prior test of the dimensionality of the data, and takes account of the size of the residuals in assessing the goodness of fit, there is some support for McDonald's claim that it is superior to the usual methods of testing fit. There remains the problem of determining whether his approach is robust with respect to violation of the assumption of normality in the distribution of the latent trait.

There is obviously scope for improvement in testing the fit of latent trait models, and the rather different approaches of Douglas and McDonald to the problem provide useful directions for further investigation.

HOW EFFECTIVE ARE CONDITIONAL AND UNCONDITIONAL PROCEDURES FOR THE ESTIMATION OF ITEM PARAMETERS?

The 1970s have been marked by a considerable amount of interest in the development of unconditional and conditional maximum likelihood procedures for the estimation of item parameters. Unconditional procedures which involve the simultaneous estimation of both item and person ability parameters have the disadvantage of yielding inconsistent estimates of item parameters. Conditional procedures yield estimates of item parameters conditional on person ability parameters and are commonly accepted as possessing theoretical advantages, especially with respect to the testing of goodness of fit. The seemingly intractable numerical problems of estimation with conditional procedures led investigators to make some progress with the refinement of unconditional estimation procedures.

As mentioned in the previous section, McDonald questions the acceptability of procedures which involve the simultaneous estimation of item and person ability parameters, on grounds which include the uncertainty of convergence. The difficulties are exacerbated by the inclusion of guessing parameters, unless a wide range of abilities is involved.

In presenting his generic Rasch model, Douglas provides a more generalized framework within which estimation procedures for item parameters can be considered. By determining how many data sets.could have produced the observed marginal totals for all parameters, he can estimate the conditional likelihood of the observed data set given the observed marginal totals. This enables him to focus on any designated set of parameters, say, item difficulty, and to estimate the conditional likelihood for the set of item marginals given the set of raw scores, thus allowing him to arrive at item parameter estimates which are not dependent on subject ability parameters. A number of numerical analysis problems have to be solved, however, before these conditional maximum likelihood estimation procedures can be put into operation. He recognizes that Gustafsson (1980) has recently developed conditional estimation procedures which can be successfully applied in the Rasch model for up to 100 dichotomous test items, but has some reservations about the effects of extreme item parameters and rounding errors on the estimates yielded by these procedures. Pending further work on conditional estimation procedures, he recommends the use of unconditional estimates of parameters, which can be subsequently 'corrected' to the corresponding conditional estimates, though the extent of applicability of such corrections has also to be explored.

Douglas's paper is an important one with respect to the estimation of both item parameters and person parameters, and opens up new avenues for investigation in this technically complex aspect of latent trait measurement.

CAN LATENT TRAIT MODELS COPE WITH THE FACT THAT ABILITY PARAMETERS CHANGE AS A RESULT OF INSTRUCTION AND OVER TIME?

Keats points out that ability is a trait which is likely to change with time, and that the ability parameter being estimated through person parameters in the Rasch model and some other latent trait models is time-bound. Taking cognitive development as an area in which ability can be expected to change, and using in turn both ratio IQs and deviation IQs as the individual differences parameter, he finds that the one-parameter model fails to account satisfactorily for cognitive development. Rather better results were obtained with a two-parameter model of cognitive development, which incorporates as individual differences parameters both the asymptotic value of a person's ability and the rate at which he approaches that level. While this approach provides a procedure for accounting for a change in ability, it was seen to involve some difficulties by some of the seminar participants. Reliable information about both the asymptotic value of ability and the rate of development may be difficult to obtain for a significant proportion of persons prior to adulthood. More fundamental was the question of whether latent trait models should be expected to cope with changing ability. Should the ability parameter be an estimate of ability at the time of measurement only, or an estimate which also took into account the likely ultimate level of development of that ability?

The use of latent trait models in the measurement of change was considered also by Spada. In their paper, Spada and May set out the rationale of, and some practical applications of the Linear Logistic Test Model (LLTM), which was developed during the 1970s by a number of European latent trait theorists to overcome the problem involved in measuring change. In effect, the problem is handled by breaking down the usual item parameters into linear combinations of the operations involved in finding the solution to the item; these include not only cognitive operations but components such as the effect of different types of instruction. Whereas change in item difficulty with time or instruction is difficult to represent in the Rasch model, it can be adequately represented in a model such as the LLTM which analyses the difficulty of each component operation. Operation difficulty parameters can be used to arrive at estimates of item difficulty parameters.

Spada and May argue significantly that the structure of a task or item is not likely to be the same for all persons in a sample, as is assumed by both the LLTM and the Rasch models. Intellectual development does not necessarily take the form of increasing mastery of the same solution algorithm, but may be characterized by the appearance of different solution algorithms. By allowing change to be examined at a basal level, the LLTM provides more possibilities for coping with structural change than does the Rasch model. It has distinct advantages for the evaluation of factors contributing to change in item difficulties, and considerable potential as an approach to the study of change. As McDonald noted during the seminar, the significance of this procedure lies more in its new approach to the modelling of change than to the measurement of change.

HOW CAN LATENT TRAIT THEORY IMPROVE TEST DEVELOPMENT PROCEDURES?

One of the main purposes in singling out latent trait theory as the major focus of this seminar was the gap between the theoretical advances in test theory and the procedures being applied at the practical level in test development, which by and large have taken little or no cognizance of these advances. Hence it was highly appropriate that some attention be given in the seminar to practical applications of latent trait concepts in test development.

A commonly claimed advantage of latent trait models is that they facilitate the equating of scores across tests, because of the availability of sample-free item parameters and item-free person parameters. This particular application of the Rasch Simple Logistic model was examined by Morgan in relation to the equating of different trial and final forms of the Australian Scholastic Aptitude Test (ASAT), through the use of link items at both the whole-test level and sub-test levels. The procedures generally followed the Rasch common item method of equating tests as set out in Wright and Stone (1979).

A number of findings from this study are likely to be of general interest to test constructors. It was found that items which did not conform to the Rasch model were largely those with very high or very low item discrimination indices. The percentage of ASAT items conforming to the Rasch model was greater when items were calibrated within their respective sub-tests rather than across the whole test, presumably because of a greater degree of unidimensionality within the separate areas. In this type of equating exercise, test constructors should make sufficient allowance for the loss of link items which do not fit the Rasch models. The effect of the positioning of link items within a test also seems worthy of further study.

Morgan's study demonstrates that it is possible to use Rasch models to equate factorially complex scholastic aptitude tests at both the whole-test and sub-test levels. It would be useful to ascertain what was happening to the item pool in the process. Is the factor composition of the finally selected items less complex than that of the original pool? Morgan's suggestion of classifying a set of test items into homogeneous sub-tests before applying the Rasch model seems sensible, and akin to the old question of whether to use total score or verbal and quantitative subscores as the criterion against which to analyse individual test items.

Izard and White's paper is an attempt to make latent trait analysis procedures accessible to classroom teachers, a development which must occur if latent trait models are to have a significant impact on educational testing as distinct from educational and psychological measurement in research. In the classical measurement tradition, booklets such as Educational Testing Service's *Making the classroom test* or *Multiple choice questions: A close look* have had a major influence in spreading ideas about item analysis procedures and notions of reliability and validity among teachers. These ideas have become more readily available to Australian teachers in recent years through the publication by state education departments or examining bodies of series of booklets under such titles as 'School Assessment Procedures'.

Izard and White describe the use of Rasch procedures to develop a pool of calibrated items for use by teachers. They distinguish between progress tests consisting of small numbers of items to indicate the degree of mastery of specific skills, and review tests, a large collection of items designed to give a broader coverage of a student's performance in a content area. Their example, using a uniform test with items evenly spaced across the difficulty range, suggests that tests with small numbers of items can be satisfactorily prepared for progress tests to allow a student to be regarded as having mastered a skill if he scores 5 or 4 on a five-item test, or having not mastered it if he scores 0 or 1. While their procedure for developing progress tests depends on applying the characteristics of a narrow test, they use the size of the standard error of measurement as a criterion to determine the appropriate length of review tests.

Once banks of calibrated items are prepared, their use by teachers in constructing classroom tests will depend very much on whether simplified and easily understood item calibration procedures are available. The simplified PROX procedure from *Best Test Design* (Wright and Stone, 1979) and the method of calibrating teacher-made items on to the item bank scale by using link items selected from the latter scale are illustrated by Izard and White, and would seem to have a reasonable chance of implementation by teachers who are prepared to put this extra effort into their assessment practices.

The use of worksheets of the type suggested by Izard and White will be essential if the Rasch procedures are to be applied by teachers in their own tests. The task of making clear to teachers the assumptions and concepts of Rasch measurement is likely to be more difficult, but should be aided by manuals such as *Best test design* (Wright and Stone, 1979) which provides exceptionally clear and simple presentations of basic concepts of measurement.

The work done by Izard and White in the development of short progress tests exemplifies Thorndike's argument that it is in the areas of individualized and adaptive testing that latent trait models, which are well suited to the estimation of a subject's precise location on a trait dimension, will have a considerable impact on test development procedures.

Item banking is an area in which latent trait theory could make a major contribution to measurement, again because of its sample-free item parameters. The value of item banks is enhanced by the availability of invariant item difficulty parameters. The importance of item banking is underlined in Choppin's paper, both with respect to national monitoring programs and its general application for use in schools. Some reservations were expressed at the seminar as to whether the Rasch model was sufficiently robust as a base for monitoring programs, and particularly about the effects of the proposed multi-chaining procedures on item parameters, which might be expected to become considerably less stable with moves away from centralized curricula to school-based curriculum development. There is little doubt that latent-trait-based item banks will be of assistance to schools in their assessment procedures. Their use in monitoring programs is somewhat more controversial, although it is difficult to determine the mix of technical measurement problems and the educational and political overtones in such controversies. Theoretical questions about the viability of latent trait models for such purposes will need to be considered in the light of the extensive practical experience that the National Foundation for Educational Research in England and Wales has acquired in the use of such models.

Choppin also describes some novel applications of latent trait models to particular practical problems in educational testing, such as the determination of between-marker agreement and the handling of score matrices with incomplete observations.

WHAT WERE THE MAJOR CONTRIBUTIONS OF THE SEMINAR TO MEASUREMENT THEORY?

The reader may have gained the impression from the previous pages that widespread consensus of views was the order of the day. Such was not the case. A number of participants, and especially McDonald, felt that the Rasch and other latent trait modellists may be in danger of cutting themselves off from other areas of psychometric theory, and warned against a complete rejection of older models in the search for new models. Despite the differences in orientation on this issue, it is apparent that some important contributions towards the unification of test theory were made in the seminar, particularly by Andrich, Douglas, and Keats.

Andrich has made an undoubted contribution to bridging the gap between older and more recent models of measurement by showing that the Rasch latent trait model synthesizes the Thurstone and Likert approaches to attitude measurement. He perceptively observed that Thurstone was searching for an attitude measure which was invariant across different groups of persons, and a person measure which was invariant across different sets of statements, whereas Likert was not. He expresses Thurstone-type scale values in the form of the simple logistic model for dichotomous-response rating scales, and derives a 'rating response model' for the ordered polychotomous response case by taking the threshold and statement values as additive components in the model, and bringing together the results obtained from considering each threshold independently. In effect, the Thurstone and Likert approaches are considered as special cases of the dichotomous and polychotomous form respectively of the ordered response model. This development of a Rasch rating model represents both a theoretical and a practical contribution to the improvement of measurement of attitudes.

Douglas in turn has made a substantial theoretical contribution in generalizing the theory behind the Rasch model to incorporate variants of the model, including Andrich's model for polychotomous attitude scale items and the Rasch/Andrich essay grading model. Douglas's generic model reduces to the standard Rasch approach in the case of dichotomously scored items.

Keats was the only author to make a direct examination of the relationship between classical test theory and latent trait theory. He arrived at the important generalization that the true scores in classical test theory show an explicit relationship with latent ability values only when all items have identical item characteristic curves. This condition might be regarded as unnecessarily restrictive by test developers, although there would be practical advantages in having tests of equivalent items available at a number of different age levels.

Given that these three authors were successful in achieving some further integration of measurement theory, there was nevertheless a feeling on the part of some of the seminar participants that latent trait theorists were failing to take sufficient account of the mainstream of measurement theory. This point of view was resisted by some of the latent trait theorists, who thought that premature attempts to integrate differing theories might obscure the special features of new theories. McDonald's plea for a more concerted effort on the part of latent trait theorists to consider their models in the context of other aspects of measurement theory is worth heeding. It would seem incongruous, for instance, not to expect some correspondence between the latent trait measures yielded by factor analysis of item data and those estimated through latent trait models.

WHAT CHANGES IN MEASUREMENT PRACTICE ARE LIKELY TO RESULT FROM THE INCREASING USE OF LATENT TRAIT MODELS?

It will be appropriate to complete this overview of the seminar with some personal predictions of the changes which are likely to occur in measurement practice in Australia as the result of an increasing use of latent trait models. These will coincide to some extent with the predictions made in

Thorndike's paper, since some of the new approaches made possible by latent trait models will be likely to be adopted irrespective of any national differences in the philosophy or practice of measurement.

Except in the case of short 'progress' tests designed to estimate fairly precisely a person's position on a dimension, there seems likely to be little change in the types and characteristics of items selected for achievement and ability tests because of the use of latent trait models. Most of the items which are acceptable in terms of traditional item facility and discrimination indices are likely to be acceptable in a latent trait model. If experience indicates that some 'good' items under traditional indices are rejected by latent trait models because of lack of fit, practitioners may well show some inclination to question the model as well as the items. In effect, tests of any reasonable length can be expected to have similar distributions of item facility and discrimination indices, and similar levels of reliability as they have at the present time.

Major changes can be expected in the provision of short progress tests with uniform distribution properties of the kind described in the Izard and White paper. Testing is likely to be used more for instructional purposes and somewhat less for survey purposes, especially if computer facilities became more readily available at the local level. This will stimulate demand for individualized testing, for adaptive or tailored tests, and the greater availability of such tests will in turn promote a greater use of tests for instructional purposes. The fact that latent trait models can be used to provide fairly precise item-free estimates of person ability will probably lead to their widespread adoption in the development of such tests. There may well be some development of fine-grained tests in accordance with the linear logistic test model to assess whether the individual operations required to answer an item have been mastered.

Latent trait models are also likely to improve the quality of item banks and to generate more interest in their use. Teachers are likely to become more accepting of item banks as an additional resource in their teaching and testing, and more so if the items are accompanied by adequate sample-free information about their parameters.

Some slackening of demand for norm-referenced tests can be anticipated, though it is not likely to be very pronounced. Teachers are still likely to be interested in comparing the performance of their students with other appropriate reference groups, even if they have item-free indices of their students' achievement in different subjects. Their reliance on norm-referenced tests will be greater if it proves to be difficult to apply latent trait models to achievement tests in some of the traditional content areas. Norm-referenced tests are likely to retain their appeal also for educational administrators and for psychologists involved in the assessment of ability and aptitude.

A major obstacle in gaining widespread acceptance of latent trait models will be the inherent complexity of the measurement notions on which they rest. It takes years to wean the public and even teachers away from the idea that marks can only be interpreted on a percentage scale. There has obviously been some increase since the 1940s in the percentage of the community with some understanding of the ideas of standard deviation and percentiles, but these notions are still not understood widely. Given the recent public controversies in Australia about the conversion of raw scores to scaled scores, one must view with some trepidation the public's likely degree of understanding of a student's score on an underlying ability or achievement scale which does not range from 0 to 100. A great deal of effort will have to be expended in communicating the meaning of the new score scales to teachers and the public. If these ideas remain inexplicable to people at the level at which they are to be implemented, their implementation is unlikely to be successful.

Although the prospects of an early widespread acceptance of latent trait measures in the public educational domain seem dim, they are probably much brighter in the areas of educational and psychological research. Research can only benefit from the use of sample-free item scores and item-free person scores. If the seminar was successful in raising the level of understanding of latent trait models among researchers and measurement specialists in Australia, it will have proved to be a significant event in the improvement of measurement in education and psychology in Australia, and a fitting event to mark the ACER's achievements in educational and psychological measurement on the occasion of its golden jubilee year.

REFERENCES

Gustafsson, J. A solution for the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, 1980, **40**, 377-85.

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.

Wright, B. D. and Stone, M. H. *Best test design: Rasch measurement*. Chicago: MESA Press, 1979.

List of Participants

Chairman: Professor D. Spearritt, Department of Education, University of Sydney

Professor J. Anderson, School of Education, Flinders University of South Australia

Dr D. Andrich, Department of Education, University of Western Australia

Dr M. Bailey, School of Education, Macquarie University, New South Wales

Professor S. Ball, Department of Education, University of Sydney

Dr W. Bartlett, Department of Psychology, University of Melbourne

Dr N. Baumgart, School of Education, Macquarie University, New South Wales

Mr R. C. Bell, Research Unit in University Education, University of Western Australia

Dr K. D. Bird, School of Psychology, University of New South Wales

Mr S. F. Bourke, Australian Council for Educational Research

Mr G. Bradshaw, Faculty of Education, University of Melbourne

Dr T. M. Caelli, Department of Psychology, University of Newcastle

Dr B. Choppin, National Foundation for Educational Research in England and Wales (*now* with Centre for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles)

Professor K. F. Collis, Faculty of Education, University of Tasmania

Dr G. Cooney, School of Behavioural Sciences, Macquarie University, New South Wales

Professor M. Cooper, School of Education, University of New South Wales

Mr G. Cornish, Victorian Institute of Secondary Education

Dr J. Davidson, Department of Psychology, University of Tasmania

Dr G. Douglas, Department of Education, University of Western Australia

Mr S. S. Dunn, Education Research and Development Committee, Australian Capital Territory

Dr T. Dunn, Centre for Higher Education, University of Melbourne

Dr J. Elkins, Schonell Education Research Centre, University of Queensland

Professor G. Evans, Department of Education, University of Queensland

Mr S. Farish, Australian Council for Educational Research

Mr I. Firth, New South Wales Department of Education

Professor D. Fitzgerald, Centre for Behavioural Studies in Education, University of New England, New South Wales

Professor R. A. M. Gregson, Department of Psychology, University of New England, New South Wales

Professor S. B. Hammond, Department of Psychology, University of Melbourne

Mr J. Hattie, Centre for Behavioural Studies in Education, University of New England

Profesor Hsu Lien-tsang, Institute of Psychology, Chinese Academy of Sciences, Peking

Dr J. F. Izard, Australian Council for Educational Research

Emeritus Professor P. H. Karmel, Commonwealth Tertiary Education Commission, Australian Capital Territory

Professor G. E. Kearney, Department of Behavioural Sciences, James Cook University of North Queensland

Professor J. A. Keats, Department of Psychology, University of Newcastle, New South Wales

Dr J. P. Keeves, Australian Council for Educational Research

Ms R. Koponen, University of Jyväskyla, Finland

Mr R. G. Lamb, Mt Lawley College, Western Australia

Dr H. G. Law, Department of Psychology, University of Queensland

Dr J. Lokan, Australian Council for Educational Research

Dr J. Lumsden, Department of Psychology, University of Western Australia

Professor R. P. McDonald, Ontario Institute for Studies in Education (*now* with School of Education, Macquarie University, New South Wales)

Professor B. McGaw, School of Education, Murdoch University, Western Australia

Dr L. Mackay, Victorian Institute of Secondary Education

Dr G. Maxwell, Department of Education, University of Queensland

Mr J. Miles, Newcastle College of Advanced Education, New South Wales

Dr W. F. Moore, New South Wales Department of Education (*now* with Catholic Teachers College, Sydney)

Mr G. Morgan, Australian Council for Educational Research

Mr D. G. Palmer, Tasmanian Education Department

Mr C. Poole, Faculty of Education, University of Melbourne

Mr N. Reid, New Zealand Council for Educational Research

Dr H. Rowe, Australian Council for Educational Research

Dr G. Rowley, School of Education, La Trobe University, Victoria

Dr E. Rump, Department of Psychology, University of Adelaide

Mr B. E. Sheridan, Claremont Teachers College, Western Australia

Dr A. G. Smith, Department of Education, University of Newcastle, New South Wales

Dr. G. A. Smith, Department of Psychology, University of Melbourne

Professor Dr Hans Spada, Psychologisches Institut, Albert-Ludwigs Universitat Freiburg

Professor G. Stanley, Department of Psychology, University of Melbourne

Professor J. P. Sutcliffe, Department of Psychology, University of Sydney

Professor R. L. Thorndike, Department of Educational Psychology, Teachers College, Columbia University, New York

Professor S. C. Tseng, Department of Psychology, Shanghai Teachers University

Mr P. Varley, Queensland Department of Education

Mr J. White, Australian Council for Educational Research (*now* with Victorian Education Department)

Mr M. Wilson, Australian Council for Educational Research

Name Index

Subject Index