British Journal of Mathematical and Statistical Psychology (1989) 42, 139–167 *Printed in Great Britain*

Five decades of item response modelling

Harvey Goldstein†
Institute of Education, University of London, 20 Bedford Way, London WC1H 0AL

Robert Wood
Psychometric Research and Development Ltd, St Albans

An historical and theoretical review is provided of so called item response theory (IRT), more accurately described as item response modelling (IRM). This paper looks at 50 years of IRM and finds a disappointing lack of advance. It is shown how a linear model framework, involving different response transformations, unifies separate approaches to the study of test item responses.

1. Introduction

In reviewing 25 years of test theory, Lewis (1986) asserted that, compared to 1961, there were now probably closer to 50 test theory models than five. He saw this as a healthy sign of a growing discipline and one which would ultimately result in a deeper understanding of that elusive relationship between 'the ability of the individual and . . . (her) observed score on the test'; Thissen & Steinberg (1986), faced with what they saw as the 'growing (and bewildering) array of models . . . proposed for use in item response theory', thought the time had come to catalogue them.

It is possible to doubt whether the proliferation we have undoubtedly seen does signify health. True diversity would have been good, but we have seen embellishment and tinkering. There has been little or nothing in the way of invention. It is still the same scenario first projected in 1942: person with ability x encounters an item, which has difficulty and discrimination, and the outcome is a binary response (or, perhaps, graded). Whether that response is correct or incorrect or even partially correct is the subject of the theory. But what sort of theory is it? As the title of Lord & Novick's (1968) book made clear, the theory is statistical, not psychological. It is not about why an individual should get an item right or wrong, or what conditions should be present for a particular outcome to happen, but about the supposed probabilistic nature of item response conditioned on something called ability. When person responses are modelled by a response function, the model *is* the theory. With rare exceptions, such as White's (1979) efforts to introduce psychologically grounded person parameters like 'speed', 'accuracy' and 'persistence' into such (latent trait) models, all the embellishments have been manipulations of the basic model, as in

†Requests for reprints.

person measurement (Trabin & Weiss, 1983), or, generalizations on a theme, as in the types of graded scoring, or purely statistical, as in the competing estimation procedures.

What has been going on since Ferguson's (1942) paper is not item response theorizing, but item response modelling. As Thissen & Steinberg remark, the goal in item analysis (which is what IRT has really been about even if persons have intruded occasionally) is to describe the observed item response data; such a goal is served by a model (or, at times, more than one model) that can account for the observed data. Our purpose, in writing this paper, has been to confront IRT for what it is, item response modelling (IRM). In applying a standard linear modelling approach, we are able to bring out the essential unity among apparentiy disparate models and to show that no one model has a claim to special status.

2. Item response theory (1942–)

Item response theory (IRT) hinges crucially on the assumption that only a single latent trait underlies performance on an item. The same applies to so called 'classical' test theory although it is not so often remarked upon. Fred Lord, looking back on over 30 years of pioneering work in IRT, thought this still a reasonable assumption; 'most tests are constructed to measure a single trait, for example, verbal ability' (Lord & Stocking, 1985, p. 2745). If that seems a shade underconceptualized, Lord had been more forthcoming a few years earlier:

> It seems plausible that tests of spelling, vocabulary, reading comprehension, arithmetic reasoning, word analogies, number series, and various types of spatial tests should be approximately one-dimensional. We can easily imagine tests that are not. An achievement test in chemistry might in part require mathematical training or arithmetic skill and in part require knowledge of non-mathematical facts (Lord, 1980, p. 20).

For Bejar (1983, p. 18) this was inadequate. In practice, he said, dimensionality is situation specific. It is not a property of the items but rather of the responses to items under a specified set of conditions, much like reliability is not a property of a test but rather of responses to a test. If the population being tested included a substantial number of dyslexic children, an otherwise unidimensional spelling test may turn out to be multidimensional. Whether logically distinct constructs show up as such depends on the nature of inter- and intra-variation, which in turn depends on the homogeneity or otherwise of nurturant and educational experiences to date. Differential instruction or training can create multidimensionality where before there had been unidimensionality (Traub, 1983).

Bejar is right; it is always an empirical question whether a test is unidimensional or not and indeed whether IRT provides the appropriate model. Lawley, to whom we owe IRT, was content to write, 'we shall assume that all items composing a given test are measuring the same ability' (Lawley, 1943, p. 273) thus, in effect, attributing unidimensionality to the items. Lawley's uncomplicated approach is understandable. In his time, analysis of even the one-dimensional model appeared formidable, and it is unsurprising that multidimensional models were not entertained. Furthermore, Lawley, a statistician, was not concerned with unpacking what 'ability' might mean.

Yet those psychometricians who followed, have, until recently, been content to stay with single ability models.

The founders of IRT all worked in Edinburgh and were responsible for three seminal papers: Ferguson (1942), Lawley (1943) and Finney (1944). But there was a pre-Edinburgh period, as it were, when some eminent American psychologists, Thurstone, Richardson and Terman, but above all, Thurstone, nosed their way towards what the Edinburgh people formalized, the characterization of item performance in terms of a probabilistic item response curve and the parameters of location and slope.

3. Pre-Edinburgh

Interest in item analysis and the item characteristic curve concept was present at the beginning of psychological and educational measurement. Binet & Simon (1916) presented tables in which proportions of correct response were arranged as a function of age. Thurstone (1925) took this further and described how items might be scaled on an age metric in terms of the average age corresponding to 50 per cent success rates; he also drew ogival curves to show how performance improved with age.† Later, Terman applied Thurstone's methodology to the scaling of Stanford–Binet items (Terman & Merrill, 1937) and intuited that the steepness of the empirical item characteristic curves gives a graphic indication of the validity of the tests. What Baker (1965, pp. 168–169) attributes to Terman was essentially published by Thurstone in a series of papers (1925, 1927, 1928); also Thurstone & Ackerson (1929). Bock (1983, p. 206) observed that Thurstone's 'method of absolute scaling', while no longer in use, is important as a forerunner of modern IRT procedures.

The first attempt to fit a normal ogive to item response data seems to have been by Richardson (1936), who used the mean and standard deviation of the fitted curve to describe the instrument (Baker, 1965, p. 169). Richardson also showed, and in this he truly preceded Lawley, how many concepts of classical test theory could be expressed as functions of the item parameters. This work was extended by Tucker (1946) and tidied up by Lord & Novick (1968) and Bock & Lieberman (1970). In a memoir, Tucker (1987) mentions that he coined the term 'item characteristic curve' (in 1945) and also confirms the seminal nature of Thurstone's 1925 paper.

4. The Edinburgh period

Ferguson (1942) was the first to speculate that the probability p_{ij} of person j giving a correct response to item i is given by the normal ogive function

$$p_{ij} = (2\pi)^{-0.5} \int_{-\infty}^{a_i + b_i f_j} \exp(-0.5y^2)\,\mathrm{d}y,$$

†Psychometricians still do not have a convenient term for what 'insecticide people' (to quote Thomson) call the median effective dose (ED50), or what psychophysicists call the limen. Baker (1965) tried to introduce x_{50}, modelled on ED50, but it did not stick.

where f indexes the latent, or unobserved, ability of the person and a, b are parameters for that particular item (facility and discrimination as it turns out). Working in the psychophysical tradition, Ferguson attempted a least squares solution using the constant process developed much earlier by Fechner, Muller and Urban. The result was somewhat unwieldy. Lawley, a colleague of Ferguson's at Moray House, took his problem and gave it a mathematical treatment, specifically maximum likelihood (Lawley, 1943). He also extended Ferguson's results considerably; it is in every respect a benchmark paper. The psychophysical tradition persisted so that when Lawley came to characterize the ability value corresponding to a 50 per cent chance of getting an item correct, we find him using the term limen.

Meanwhile, Finney (1944), working in the field of bioassay, was developing probit analysis for estimating animal and plant tolerance to drugs, insecticides, etc. Finney saw a parallel with psychological testing; the dosage was the ability and the organism was the item. He did more than that; he also saw the connections with psychophysics and the constant process. By applying the probit method to Ferguson's data, the 1944 paper, to use Finney's own words (1971, p. 42), brought the two streams together. Godfrey Thomson tells us that Lawley, like Ferguson, did not know of the developments in probit analysis. It was only because Finney was a friend and correspondent of Lawley that connections were made (Thomson, 1947, p. 72). Lawley acknowledged as much in his 1944 paper. All Finney had to say was that he, Lawley, 'came near' to an independent derivation of the maximum likelihood solution (Finney, 1971). Once the bioassay–psychophysics connections had been made, the tendency to think of ability as something of which you had more or less, was reinforced, but then this was entirely in keeping with the prevailing 'strength' model of intelligence.

Lawley took Ferguson's formulation and developed it. With the psychophysical analogue so inviting, it is easy to understand Ferguson's choice of model. But now we may ask whether it was such an apt choice. Certainly, anyone looking at the tangles of item characteristic curves produced by Raven (1986, pp. 69–70) would be left wondering. In particular, why an ogive when you are not especially interested in the tails? Why not, as we argue later, use a linear function to start with? The form of the model seems immaterial since it is robust against mis-specification in the tails (Reese, 1986, p. 205). In fact, when estimating individual 'abilities' we are seldom interested in the tails.

Lawley's 1943 paper showed that many concepts of classical test theory could be expressed as functions of the item parameters. Simultaneously with this work, he was working on maximum likelihood solutions for factor analysis (Lawley, 1944). Just how close Lawley got to synthesizing factor analysis and latent trait theory, accomplished later by McDonald (1967, 1982), is recounted by Wood (1987).

Lawley's 1943 paper was the foundation on which Lord built, as he was ready to acknowledge (Lord, 1968). Baker praised Lawley for emphasizing the importance of working at the item level yet both Lawley himself, in his 1944 paper, and Lord later, concentrated on the properties of tests, as the titles of Lord's (1952) monograph and Lord & Novick indicate. By 1965 Baker was complaining that theory began and ended with test scores, ignoring the composition of the test. In fact, it was not until Lord commenced his work on tailored testing (circa 1968), when the whole notion of fixed tests had to be sidelined, that the item was recognized as the appropriate unit of analysis, and the appellation latent trait theory was dropped in favour of item characteristic curve theory (Tucker's term) and then item response theory.

Guilford, reviewing Lord's monograph, pointed out that the assumption of unidimensionality and the exclusion of multiple-choice tests, place severe limits on the usefulness of the theory. 'Since most tests depart from this highly specialized situation, the generality of applications from his theory is limited' (Guilford, 1954, p. 363). It is perhaps little known that the Lord/Lawley/Ferguson model was intended for free-response items only. Perhaps it would be desirable to revert to the free-response only condition; certainly the introduction of a guessing parameter into the IRT models (or the decision not to include one, as with the Rasch model) has been a constant cause of friction, if only because it has been so obviously a property of items rather than people; White's model being an exception.

By 1960 Lord is well advanced in developing the original model, and then comes a surprise. Out of nowhere (or so it seems) appears Georg Rasch with a collection of item response models, one of which turns out to be the simplest IRT model, a one-parameter (ability) logistic ogive function (Rasch, 1960). This comes to be known as the Rasch model while other more interesting models (psychologically), like the Poisson model for misreading, are ignored. Whereas Lord and the others have felt it necessary to persist with Ferguson's original item discrimination parameter in the IRT model, Rasch decides it is not necessary nor is he interested in a guessing parameter, which Lord and others believe must be included. Data should be collected in such a way that the need for such parameters is obviated.

Ordinarily, this would be a routine check on whether Occam's Razor has shaved too much off. But proponents of Rasch are present in numbers and only now is it possible to adjudicate on whether it matters in utility terms, whether you use Rasch or a more elaborate model. There has been strong resistance to Rasch when used in educational applications. Baker (1977) wondered even whether latent trait models were appropriate at all. As Wilcox (1980, p. 443) observed, latent trait theory is the culmination of the work on the measurement of ability begun by Binet that was the major focus of psychometrics in the 1920s, 1930s and 1940s but the educational problems of an earlier era are not the problems of the 1980s and 1990s. The major trend in educational measurement today is that of instructionally related testing. How, given deliberately targeted remedial instruction, can the assumption that the relative difficulties of pairs of items, remain the same? How, if education is about change, which is bound to be differential, can a stationary model like Rasch be entertained? Thorndike (1982) thought that Rasch would be most suitable for aptitude tests, where what develops is not so susceptible to instruction and is liable to change at uniform rates. Hoover (1987) hazarded that Rasch would do best with musical aptitude tests which would be as unidimensional as anything might be, but again unidimensionality is an empirical issue.

There is no particular evidence to support the view that Rasch will do better with aptitude tests. Johnson (1986), writing of recruitment via aptitude tests in the British Civil Service, concluded that Rasch would not work well in that situation precisely because the heterogeneity of the candidature and the relative absence of homogenizing experiences resulted in wide variation in discrimination parameter estimates.

IRT, including Rasch, is generally weak on utility grounds. Lord conceded that applications of item response theory are generally more expensive than similar applications of classical test theory (Lord & Stocking, 1985). It is commonly supposed that the use of IRT (the logistic model) is imperative for individualized testing because classical test theory is inadequate.† As we show later, this is not so. Any properly formulated item response model will do.

5. The basic model

We replace the term 'Item Response Theory' by the more accurate descriptor 'Item Response Modelling' (IRM).

Unidimensional IRM's are latent variable or factor analytic models where the response variables are the correct $(=1)$ or incorrect $(=0)$ responses to the items in a test. Because the response is binary, the most common statistical assumption in IRMs is that it has a binomial distribution. We note in passing that this assumption will nearly always be questionable. Thus, for example, samples are often taken from 'clustered' populations such as schoolchildren, so that over and above any individual factors we would expect school specific effects (see Goldstein, 1987, ch. 6). This would then lead to biases in estimates for standard errors etc.

Let the $(0, 1)$ response be denoted by p_{ij} for the ith item for the jth subject. Denote the probability of a correct response by

$$\pi_{ij}=\text{prob}\,(p_{ij}=1\mid f_j)$$

$$=a_i+b_i f_j.$$

The term f_j represents the underlying assumed latent trait and we can write the observed response as

$$p_{ij}=a_i+b_i f_j+e_{ij}, \tag{1}$$

where e_{ij} is a random variable.

This is an example of a congeneric test score model (Jõreskog, 1971).

The term b_i is known as the discrimination of item i, measuring the average rate of change of π_{ij} with f_j. Furthermore, the average of π_{ij} over the distribution of f_j is the average probability of a correct response, or the item facility. More generally, if we suppose that f_j is actually measured, we can then plot p_{ij} against f_j to give the 'item characteristic curve' (ICC), and (1) assumes that this is a straight line. As we shall show this is not generally entirely suitable as an item response model, but supposing for now that we accept it as reasonable, the next task is to find an appropriate procedure to estimate the parameters, including the f_j also, since these are unknown and hence considered as parameters to be estimated along with the a_i, b_i. (We see later that we can avoid the estimation of every one of the f_j directly by making some distributional assumptions about them).

Assuming independent residuals, the parameters in (1) can be estimated using an

†The flexilevel procedure (Lord, 1971) did, however, manage without IRT.

iterative procedure with weights which depend on the predicted values, and subject to boundary constraints. We note that (1) is a regression model, where for each subject we have

$$p_i = a_i + b_i f + e_i. \tag{2}$$

Let us compute the ordinary estimate of the regression coefficient, namely

$$\hat{f} = \Sigma_i (p_i - a_i) b_i / \Sigma_i b_i^2.$$

This is a linear function of

$$\Sigma_i p_i b_i$$

which is simply a weighted sum of the item responses, the weights being the item discriminations. If these discriminations are equal then this coefficient is a simple function of the raw score, namely,

$$(x - \Sigma_i a_i)/(\Sigma_i b_i),$$

where x is the raw score. Moreover, even where the discriminations are unequal the expected value of the raw score is

$$\Sigma_i a_i + (\Sigma_i b_i) f.$$

So the raw score is an unbiased estimate of a fixed linear function of ability. It is only efficient, however, when the discriminations are equal, and we use the appropriate weights. Nevertheless, in many practical applications the discriminations do not differ markedly, and so the raw score will be a reasonably efficient estimate. A similar result will hold when we consider 'transformed' IRMs. Thus we see that the procedure in classical 'item analysis' for choosing the raw score as an estimate of ability can be justified in terms of a simple linear model. It does not of course follow that the choice of raw score to characterize an individual set of item responses therefore implies a model such as (1), since there may well be other grounds for such a choice.

A difficulty with (1) is that it places no restrictions on the probabilities, so that these could lie outside the admissible range (0, 1). A simple procedure for avoiding this is to constrain the solution so that upper and lower boundaries are defined for the probability of a correct item response. These can be interpreted as 'guessing' and 'carelessness' parameters respectively, and they will need to be estimated. The model also constrains individual abilities to lie between the corresponding points on the ability scale.

This '4 parameter' model would seem to provide a reasonable description of an ICC for a unidimensional ability with the linear form of IRM given by (2), and was suggested by McDonald (1967). We could improve the description where necessary for particular items by introducing higher order terms into (2), for example quadratic or cubic terms to allow for curvature. This will introduce further item parameters but does not introduce serious new difficulties into any estimation procedure. We note that although there are now more individual terms, the model is still unidimensional. We shall return to non-linear models shortly where the response probability is transformed, using a non-linear transformation, and the transformed probability then related to the item and individual parameters. Before we do this, however, we shall further explore (2) and see how it leads us to adopt the traditional range of item analysis procedures.

6. Reliability

We now explain how IRM's have been used to estimate reliability.

If we start with equation (1) and form a raw score by summing over the n items in a test we obtain

$$\begin{aligned} y_j &= \Sigma_i p_{ij} = a + bf_j + e_j \qquad (3) \\ &= x_j + e_j. \end{aligned}$$

This is in the common form 'Observed score = True score + Error'.

Reliability is defined as

$$R_y = \frac{\operatorname{var}(x)}{\operatorname{var}(y)}. \qquad (4)$$

In classical test models it is supposed that the e_j vary randomly from occasion to occasion. The e_j in (3), however, may contain a component which does not vary across occasions. We recall that the assumption made about the residuals e_{ij} was that, given f_j, they were independent across items and across individuals. If we introduce, say, a second independent administration of the test, we could then write

$$e_{ijt} = e_{ij} + d_{jt},$$

where t $(=1;2)$ refers to the measurement occasion and d_{jt} represents between-occasion random variation for the jth individual. Thus the item mean response is given by

$$\pi_{ijt} = \pi_{ij} + d_{jt}.$$

We note that d_{jt} will typically include such circumstances as the general environment, day-to-day variation, etc. Some of these variables may be common to subsets of individuals and this will create estimation problems which are discussed below. Such a 'hierarchical' partitioning of the residual is analogous to certain kinds of longitudinal data models (Goldstein, 1987, ch. 4), but the practical estimation of the separate components does not seem feasible in general and in practice we have to work with the actual residual e_{ijt}. In effect this constitutes a particular definition of measurement error which includes 'within' and 'between' occasion error. We note in particular that even where $e_{ij}=0$, we may still have a between-occasion component of error variance.

It will be noticed that the definition of reliability is with respect to a particular population. In a different population the true score and/or residual variance may change as may any factor structure underlying the true score. In particular, in moving from a heterogeneous to a homogeneous population we might expect this to occur. A common example is where a test is given to students covering a wide age span. This will tend to strengthen the factor structure spuriously due to the propensity of students to acquire knowledge and understanding at different rates, but in a similar order (Levy, 1973, p. 6).

More importantly perhaps, the conditions of testing will generally affect each response probability in a similar direction. If these conditions vary, as is typically the case, then this will strengthen the common factor structure, and will act as if there are further factors along which individuals differ. Such conditions may include presentation of material, time of day, environment, health, etc. As pointed out above, d_{jt} will incorporate such between-individual structures. Such variation should really be counted as a component of the error variation and not the common factor structure. In other words, circumstances to do with test administration which have a common effect on test items will lead to spuriously high estimates of reliability based on coefficient alpha.

In fact there are many possible definitions of reliability. Thus, if we think of a test as one possible selection of items from a universe of items, then we can define a test–retest reliability which measures variation in a well-defined population of possible tests based upon a universe of test items. This notion is formalized in so-called generalizability analysis (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Another possibility is the 'alternate forms' reliability, concerned with variation between tests specifically designed to have items with characteristics as similar as possible. This can be modelled similarly to the between-occasion model above. The form of reliability to be chosen must depend on its use and its context.

7. The general linear item response model (GLIRM)

We can extend (1) in an obvious way to include further explanatory variables measured on individual items so that a_i is replaced by

$$\Sigma_t a_t z_{it} + d_i$$

$$p_{ij} = \Sigma_t a_t z_{it} + b_i f_j + e_{ij} + d_i \qquad (5)$$

$$t = 1, \ldots, k; \quad z_{i0} = 1.$$

Thus we might have $t=1$ where z_{i1} is a measure or classification of the presentation format of the item, say. The estimation of the parameters of (5) is a straightforward extension of the procedures already outlined. We can also model b_i as a function of further variables, giving

$$p_{ij}=\Sigma_t a_t z_{it}+(\Sigma_m b_{im} w_{im}) f_j+e_{ij}+d_i.$$

We can also generalize (1) to the case where individual subjects are classified by further observed variables

$$f_j=\Sigma_m c_m v_{jm}+u_j$$

giving,

$$p_{ij}=a_i+b_i\Sigma_m c_m v_{jm}+b_i u_j+e_{ij}. \tag{6}$$

Some simple examples of such models are given by Fischer (1983) and Reiser (1983) for the logistic IRM described below. It is also possible to have models where both items and individuals are classified by further variables. Finally, some of the parameters in (6) may themselves be random variables, varying randomly across schools for example, so giving rise to random coefficient models (Goldstein, 1987).

8. Generalizability models

Consider the special case of the GLIRM

$$p_{ij}=d_i+u_j+e_{ij}, \tag{7}$$

where the item facilities have a random distribution as do the person abilities. Thus, rather than seek to estimate a parameter for each item and subject, we would like to estimate the mean and variance of the d_i, u_j. Equation (7) is the basic model of generalizability 'theory' (Cronbach *et al.*, 1972). The model typically is extended by introducing further classifications and levels of nesting for items and individuals. Thus, for example, students might be grouped by tester or by school so that we could write

$$p_{ikj}=d_i+w_{kj}+e_{ikj}, \tag{8}$$

where

$$w_{kj}=v_k+u_{kj}$$

and v_k varies randomly over schools and u_{kj} varies randomly within schools across students. It is also clear that we can introduce further explanatory variables as in the previous section and so obtain a multilevel general linear model, which includes cross classifications of units (Goldstein, 1987, ch. 7). A review of generalizability models is given by Shavelson & Webb (1981). In practice, generalizability analyses tend to use scores derived from sets of items rather than individual item responses. They are an example of 'random effect' unidimensional models described below. Their main distinguishing feature is that they explicitly consider random samples of items, although it may be difficult to see how a group of items which are chosen for a test can be regarded as a *random* sample from a universe of items, unless very strict item definition and selection rules are chosen.

9. Response transformations

One method of avoiding predicted probabilities in the standard linear model outside the range (0, 1) is to transform the scale of the response probability. The most common transformation is the logistic, namely

$$\log\{\pi_{ij}/(1-\pi_{ij})\}=a_i+b_if_j. \tag{9}$$

Another possibility is the log-log function

$$\log(-\log\pi_{ij})=a_i+b_if_j. \tag{10}$$

Another common function is the cumulative distribution of the normal or gaussian curve, but the shape of this curve is very similar to the logit curve and need not be considered in detail. It needs to be stressed that the principal purpose in introducing such transformations is largely mathematical. They eliminate in a simple way inadmissible probability values. Whether this is substantively and practically important and whether such curves actually give better fits to observed data is an empirical matter upon which there seems to be little evidence. In fact there is an infinity of transformations which would satisfy the probability constraints, some of which are symmetrical like the logit and others which are asymmetrical like the log-log. Goldstein (1980), in a simple example, compares the logit and log-log transformations and finds little to choose between them in terms of overall fit to the data, but shows that they lead to quite different estimates of individual abilities, and indeed to different rank orderings of ability.

Thus, in a situation where the choice of transformation lacks empirical justification and where different transformations give different ability estimates, we do need to be rather cautious in our interpretations of the parameter values.

These remarks apply to multidimensional as well as unidimensional models. In the former case there is the additional issue that if the 'wrong' transformation is chosen then we may still be able to fit a set of data adequately by including extra factors. Thus, for example, if a log-log model is a good fit with 2 factors, we may in fact require 3 factors if a logit model were to be used. A similar point arises with single factor models, since it may be possible to have a unidimensional model under one data transformation which requires more than one factor under another data transformation. Hence when we talk about the unidimensionality of a test this always must be with respect to the particular transformation used.

For both the logit and log-log transformations we see that both transformations imply that, at the extremes of the scale, small changes in the probability of a correct answer are associated with very large changes in ability. It is difficult to envisage a substantive definition of ability, say in an educational context, which accords with such a model. It implies that there are, in principle, individuals who have a virtually infinite ability! In their own way, therefore, although they have eliminated inadmissible probabilities, these transformations have introduced further difficulties which our original model (2) did not possess. Moreover, we know that we can overcome the inadmissible probability problem by introducing upper and lower bounds into the model, interpreted as guessing and carelessness parameters. As we shall see in the next section, the introduction of a lower bound is common practice when using logit transformed models and an upper bound likewise can be used (though it is rarely attempted). Nevertheless, this does not remove the possibility of infinite abilities. In the absence of empirical or substantive theoretical reasons for choosing a particular transformation it seems that the modified version of (2), which we can refer to as the constrained identity transformation, may have theoretical advantages as a unidimensional IRM since it neither allows inadmissible probabilities nor infinite ability values. Furthermore, in the 2-parameter case, even though the identity transformation leads to inadmissible probabilities, it does provide perfectly sensible ability estimates, since these are just weighted sums of item responses. We return to this point in the next section. It may also have practical advantages, and provides a rather more straightforward interpretation.

In spite of the advantages of the 2-, 3- or 4-parameter identity model it has found little favour in the IRT literature, despite the implicit reliance of traditional item analysis upon a simple version of it. The reason for this appears to be partly historical, and partly due to certain mathematical properties of the logit model which, in the 2-parameter case, provide relatively straightforward estimation procedures, although when we move to 3- and 4-parameter models, estimation problems become more apparent.

Having said all this, the logit model still remains the standard one in use and we shall therefore devote the next section to a study of it.

10. The logit item response model

Before we consider the basic model given by (9), a few remarks are necessary concerning the simpler model where the discriminations b_i are all equal. (Note that we call the parameters in this model by the same names as in the identity model since they have the same general interpretation.) This is the Rasch model. It can be written as

$$\log\{\pi_{ij}/(1-\pi_{ij})\} = a_i + f_j. \tag{11}$$

Efficient (maximum likelihood) estimates of the ability parameters are in fact monotone transformations of each subject's raw score. Our previous discussion of the use of raw scores, when item discriminations really do differ, applies also to (11). Moreover, as pointed out in the previous section, the logit model suffers from the difficulty of interpreting infinite abilities. Related to this are estimation problems, since subjects who answer all items correct or all items incorrect cannot be assigned finite ability estimates. It can be shown (Goldstein, 1980) that the presence of such subjects leads to biases in parameter estimates.

In our view the 1-parameter logistic model has little to recommend it over its main competitor, the 1-parameter identity model, and if we wish to choose a 1-parameter model, then the identity model, which is simpler to use, seems to have the advantage. Moreover, as we have already pointed out, the estimates from the logistic model are simply monotone functions of the raw scores (and correspondingly for the facilities) so that the same rank orderings appear whichever model is used. We also note that while (11) gives ability estimates which are monotonic functions of the raw scores, other transformations such as the log-log (10), for example, do not.

In the logit 2-parameter model the ability estimates are obtained by forming a weighted sum of the item responses using the discriminations as weights and applying a non-linear monotone transformation to this sum. This is analogous to the procedure for the identity model and similar arguments apply to the comparison of the two models as for the 1-parameter models.

We have already discussed the advantages of the identity transformation in the case of the 4-parameter model. There appear to be few empirical analyses of 4-parameter logit models although Barton & Lord (1981), found no advantage over the 3-parameter model. Nevertheless, there have been several applications of the logit 3-parameter model. These suggest (Traub & Lam, 1985) that the lower bound or asymptote typically is not estimated very accurately, and imprecision in its estimate can affect the values of the other parameters. We would expect the 3-parameter identity model to behave similarly.

So far, it should be noted, we have been considering fixed effects models. In a subsequent section we consider random effects models and show how these overcome certain difficulties we have encountered.

To summarize the position: despite the considerable effort which has gone into developing unidimensional models with 1, 2 and 3 parameters, fixed effect logistic IRMs appear to possess no persuasive advantage over the corresponding simpler identity models, and some clear disadvantages. No compelling empirical evidence indicates the clear superiority of any particular transformation.

11. Goodness of fit

Goodness-of-fit tests are informative only when they are sensitive to a specific alternative model. Thus, we may wish a test of the fit of a 2-parameter model to be sensitive to the alternative of a 3-parameter model. In this case it would be appropriate to fit a 3-parameter model and study the values of the third parameter. If we wished to test whether a 2-factor rather than a 1-factor model should be used then a 2-factor model should be fitted and the loadings on the 'minor' second factor studied. In some cases a specific alternative model may be suspected. For example, in the mathematics data analysed by Goldstein (1980), separate 'algebra' and 'geometry' factors could be identified, giving (at least) a 2-dimensional model. Yet when a 1-dimensional model was fitted, a general 'goodness of fit' test indicated a satisfactory fit. This illustrates the way in which such tests can be highly misleading.

A number of alternative procedures for judging unidimensionality have been proposed, but these tend to be statistically unsound, lack sensitivity or adopt an unsatisfactory definition of unidimensionality. The factor analytic approach of McDonald (1982) provides one of the most useful formulations of this problem.

12. Conditional independence

We have assumed, as is common, that the residuals e_{ij} in the GLIRM are independent. This is often referred to as the 'local' or 'conditional' independence assumption. It implies that, for a group of individuals with the same set of (multidimensional) abilities, the observed item responses are independent.

Now, for a group of such equivalent individuals, the responses to a p-item test can be arranged as a 2^p table, and the conditional independence assumption implies that there is complete independence is this table. It is nevertheless possible to have a situation where there are dependencies in the table, for example as a result of a common autoregressive sequence (see for example, Ackerman & Spray, 1987). In the usual case, however, each of a group of non-equivalent individuals responds just once to each item. In this case, a p-dimensional model can explain all the *observed* dependencies, yet the true model might be p-dimensional together with further, say autoregressive, dependencies. Thus the confounding of the dimensionality of the between-individual variability and residual dependencies is of practical but not theoretical importance. Goldstein (1980) and Bell, Pattison & Withers (1988) point out that this confounding appears to have misled Lord & Novick (1968) into *defining* conditional independence in terms of dimensionality.

13. Random ability models

One of the major problems with the fixed effects models is that we have to estimate an ability parameter for every subject in the estimation sample. Thus, as the sample size increases the number of parameters increases and we cannot therefore guarantee consistent estimates (Cox & Hinckley, 1984). An alternative is to follow McDonald (1967) and revert to the traditional factor analysis model and suppose that the ability has a distribution among the population of individuals and then seek only to estimate the parameters of this distribution plus the item parameters.

While such random ability models have advantages, and solve certain problems, all of our previous discussions concerning choice of transformation, admissible probabilities and scaling assumptions apply equally to these models. As before, we can fit either identity models or logit or guassian models. There have been attempts, especially in the literature on generalizability modelling, to fit multidimensional identity transformation models treating the observed (0, 1) data in the same way as continuous measurements, but as might be expected, these run into difficulties. To our knowledge there have been no attempts to fit constrained identity models with random abilities, all the theoretical and empirical work being with logit or gaussian transformed models. We therefore restrict ourselves to these transformations.

We can write the 2-parameter logistic model as

$$\operatorname{logit}(p_{ij}) = a_i + b_i f_j + e_{ij}. \tag{12}$$

We now are interested primarily in estimates for the parameters a_i, b_i. The f_j are assumed to have a gaussian distribution with zero mean and unit standard deviation. The estimation procedure is iterative, where working estimates of the f_j lead to new estimates of the other parameters and a new set of values f_j are then obtained based upon the gaussian distribution assumption. Details of one version of the procedure can be found in Sananathan & Blumenthal (1978). The procedure avoids the problem associated with individuals who have all zero scores or all scores equal to 1, since the assumption of a distribution for the f_j allows finite values to be predicted for such individuals. We will have more to say about such models when we discuss multidimensional models.

14. Non-linear models

Until now we have assumed that the relationship between the (transformed) item response and the factor is linear. It is perfectly possible, however, to envisage a non-linear model, say of the form

$$\operatorname{logit}(p_{ij}) = a_i + b_i f_j + c_i f_j^2 + e_{ij}. \tag{13}$$

This, clearly, is unidimensional in the sense that only one individual attribute or variable determines the item response. We shall not discuss estimation procedures, but a general discussion of non-linear factor analysis can be found in Etezadi & McDonald (1983). Sometimes, a model which is non-linear can be made into a linear one by a suitable transformation of the response variable. Furthermore, if the model (14) actually holds and we attempt to fit the model,

$$\operatorname{logit}(p_{ij}) = a_i + b_i f_j + e_{ij}, \tag{14}$$

then (13) will not provide an adquate fit although the following model will

$$\operatorname{logit}(p_{ij}) = a_i + b_i f_j + c_i g_j + e_{ij}. \tag{15}$$

A test for two dimensions will indicate the need for a second dimension (given sufficient data), when in fact the second factor is just the square of the first. The consequence is that the number of variables in (15) is really less than implied by the model. If we take the view that the aim should be to describe a data set using as few parameters as possible (consistent with fitting the data adequately) then, ideally, when fitting an item response model, the aim should be to search for a scale or transformation which reduces the number of separate dimensions. Non-linear models under various different transformations seem to offer considerable scope in this respect.

Another extension of the basic unidimensional model is to include 'interaction' terms between item parameters and ability parameters, thus

$$\operatorname{logit}(\pi_{ij}) = a_i + b_i(f_j + c_{ij}). \tag{16}$$

Separate values for each c_{ij} cannot be estimated, but, for example, if c_{ij} is assumed, for each j, to have a gaussian distribution across items or groups of items, then this introduces an extra variance parameter to be estimated (Levine & Rubin, 1979).

A further consequence of a non-linear model such as (13) is that p_{ij} is no longer necessarily a monotone function of ability. Some writers (for example Holland & Rosenbaum, 1985) essentially *define* items to be unidimensional only if such a monotone relationship exists. Such a definition, however, not only rules out general non-linear models, it also requires a linear model to have ICCs either which are all increasing or all decreasing, that is with discrimination parameters which all have the same sign. Of course, with simple unidimensional linear factor models we can always ensure this simply by changing the sign of appropriate response variables or by redefining what is regarded as a correct or incorrect response. Whether such redefinitions are substantively justifiable is another matter. The point is that it is the form of the model not the sign of its coefficients which is relevant to the definition of unidimensionality. The values of the discrimination parameters are of considerable interest, especially in the special case where the ICCs are all increasing or decreasing, but in a unidimensional model these values are not involved in the basic definition of unidimensionality.

15. The reference population

Suppose that there are two populations, and in each one there is a single factor, but the item parameters differ, and are not a simple transformation of each other. Thus for population 1 we could have

$$p_{ij}=a_i+b_if_j+e_{1ij} \tag{17}$$

and for population 2

$$p_{ij}=c_i+d_if_j+e_{2ij}. \tag{18}$$

Consider now population 3 formed by combining the two populations, where we suppose we can write a unidimensional model

$$p_{ij}=h_i+g_if_j+e_{3ij}. \tag{19}$$

We can see that the parameters of this new model will be weighted averages of the parameters from the original models. If we now fit (19) to individuals from the first population we see that the following relationship must exist between the residual terms

$$e_{3ij}=e_{1ij}+(a_i-g_i)+(b_i-h_i)f_j.$$

Hence, if (17) satisfies the definition for a unidimensional model, in particular that the residuals are mutually independent with zero mean values, then the residuals for the proposed model for population 3 do not satisfy this assumption, being related to the factor values. Thus we need at least a 2-dimensional model for population 3 (see McDonald, 1982, for a further discussion).

To give another example, we may suppose a two factor model where the first factor is discrete, that is takes a finite number, *p*, of values. We then divide the population into *p* subpopulations on the basis of this factor. It is clear that for each subpopulation there is just one factor exhibited since every individual now has the same value for the discrete factor. Such a situation may often arise in practice where, for example, there are substantial curricula differences between schools or cultural differences between groups in terms of the response variables. In the general population this variation will be exhibited in further factors.

It follows that the parameter values and the dimensionality of any factor model can be specified only with respect to a specific population of individuals. Any structure found for one population cannot be assumed to apply to any other population or subpopulation without an empirical demonstration. This has important implications not just for the interpretation of item response model parameters but also for individual factor or trait estimates based upon them. Thus, for example, we can have a minority subpopulation with its own factor structure which has little influence on the overall parameter values for the whole population. Application of the latter to determine minority group individual factor scores can then lead to biases in these. It is therefore important in any large scale implementation of IRMs to try to identify distinct subpopulations and to study their characteristics. We note that such efforts are rare and that most item response modelling is carried out on undifferentiated large general population samples. While large population samples are important, their indiscriminate use is not to be encouraged.

Most current procedures for constructing and analysing tests are based, in effect, upon unidimensional models. We now discuss some of the main procedures and relate them to our previous exposition.

16. Item bias and item anomalies

Much effort has gone into detecting so called 'biased' or 'atypical' items using IRMs (see e.g. Ironson, 1983). An issue in all these procedures can be summed up by quoting Shepard, Camilli & Averill (1981):

> an item is biased if two individuals *with equal ability* but from different groups do not have the same probability of success on the item (our italics).

The problem is how to decide whether the individuals do indeed have the same ability. Clearly, it would be possible to define ability in terms of an externally agreed criterion. Typically, however, such criteria do not exist in a simple form and so resort is often had to other test scores, ratings and so forth. We shall not enter into a debate of the merits of such procedures, save to note that such a debate would have to come to terms with different perceptions of ability and in many cases the result will be a negotiated compromise between different viewpoints.

In the context of IRMs, however, the notion of item bias is derived from an estimate of ability based upon the model itself. Typically, two groups of individuals with the same distinct ability estimates are compared in terms of the probability of a correct response, for each item in turn and over the complete range of ability estimates. It is clear that these procedures contain a circular logic, since the items being studied help to define ability and the unidimensional models used assume no differential group differences. If.we fit a model which does allow differential group differences, such as (8), then the notion of item bias is simply equivalent to the model assumption of group differences in item parameters. Other, more sophisticated versions of this procedure exist, but all suffer from the same problem. Thus, so called appropriateness measurement (Levine & Drasgow, 1982, 1988) attempts to identify individuals with 'unusual' item patterns. The approach, however, is based upon the assumption of an underlying item response model, typically a 3-parameter unidimensional logit model, so that any unusual patterns might well be indicating the inadequacy of the model. Of course, in one sense it is artificial to make a distinction between an aberrant individual and an inadequate model. In practice, however, choosing to regard individuals as aberrant rather then the model as inadequate to cater for such individuals, has important consequences. Thus, Levine & Drasgow (1982) maintain that:

> a few examinees may be so unlike the others that their multiple choice aptitude test scores have limited value as ability measures.

These authors go on to cite individuals who may have cheated on some of the questions, and also some 'exceptionally creative examinees'. There is in all this a temptation to tailor the complexity of actual test responses to conform to the structure of a simple model.

In like manner Tatsuoka & Tatsuoka (1982) use a unidimensional, so called perfect or Guttman scale, model which can be thought of as the special case where the discrimination parameters for each item become indefinitely large so that the ICC is a simple step function. They base their procedures on the notion of subtests within each of which items are parallel. Interestingly, these authors do recognize explicitly the difficulty which is posed by a heterogeneous reference group, implying departures from their model, but admit they have no way of coping with this. Because their discrepancy measures are also dependent on the assumption that the items within a subgroup are parallel, the existence of any discrepancies can also be viewed partly as a test of this assumption as well as the assumption of a unidimensional model.

In effect, the attempt to identify discrepant items or individuals on the basis of various models, must ultimately constitute tests for the adequacy of such models. Of course, some observed responses may actually be incorrect (e.g. through miscoding) and insofar as the study of residuals or discrepancies is directed towards that end the procedures have their uses. Nevertheless, the existence of 'extreme' individuals or unusual items will often be telling us that the model is inadequate. The term item bias, in our view, should be reserved for the cases where such things as item format or cultural stereotypes are thought to affect responses and thereby to reduce the accuracy of the measurement which we wish to make.

17. Test equating

In many applications of testing, it is required that different tests be given to different groups of individuals in different circumstances and perhaps at different times. It is also required that the test scores, or ability estimates, be reported on a common scale so that individuals or groups can be compared. The procedures developed for this are known as equating methods (see Angoff, 1971 for a comprehensive account). IRMs have recently become increasingly used for equating instead of earlier methods based upon raw test scores. We shall first deal with the latter since many of the problems of these carry over to IRM based methods.

The basic model for the equating of two tests is that where both tests are administered to a random sample of individuals from a specified population an equating relationship is derived whereby the score on one test can be transformed monotonically into a score on the second and vice versa. A variant is to administer the separate tests to distinct random samples from the same population. Of course, in many circumstances an adequate equating simply will not be possible. Rather than talking about whether tests can or cannot be equated, an index of equatability, ranging from 0 to 1 could be used. We shall not here go into the details of how the equatings are actually carried out and the attendant practical difficulties. A discussion can be found in Holland & Rubin (1982) and Goldstein (1986). We should point out, however, that there is an implicit unidimensionality assumption involved in requiring a single equating function to hold for each individual.

It is also important to note that the procedures are population dependent. There is no guarantee that an equating relationship found in one group will hold for another group. Moreover, equating typically takes place using random samples from general populations and this raises issues similar to those discussed earlier concerning subpopulations and groups. For various reasons to do with, say, curriculum or culture, equating relationships may vary over subpopulations, so that an overall relationship may not reflect at all accurately the relationships to be found within subgroups or subpopulations. This raises the potentially serious issue of bias and discrimination against certain subgroups. Unhappily there appears to be little formal recognition of this in the equating literature and a lack of serious empirical study of the issue (but see Angoff, 1986).

Where tests are constructed to be as nearly alike as possible, then it would seem reasonable to assume that subpopulation equatings will be similar, although again empirical verification of this is important. In many cases, however, tests are not so constructed and in these cases there must exist, *a priori*, doubt about the constancy of equating relationships. For example, it is often required to equate two tests, each of which is designed for a different age subgroup (vertical equating), with an area of overlap. In this case an equating relationship may be established within the overlapping age range and then used outside that range. Since the tests are necessarily dissimilar, however, there may be several factors which would cause the relationship to vary over the subgroups defined by age.

As far as IRMs are concerned, equating is based upon the assumption of a unidimensional model for which the items in both tests to be equated have a common set of parameter values. Having estimated these parameters via suitably overlapping samples or using equivalent random samples, then any individual score can be estimated. A variant of this procedure arises in the construction of so called scaled item banks or pools where a large number of items are 'calibrated', that is, parameters are estimated from suitable samples and then tests formed from sets of items selected from the pool. All our previous remarks about parameter variation between subpopulations and possible lack of unidimensionality, apply to IRM equating (see Section 19).

The literature which seeks to attempt to validate equating almost always does so without seriously questioning the basic model assumptions. Thus, for example, a popular technique is to equate a test to itself. This is done by having several tests and equating test 1 to test 2 then test 2 to test 3, etc., and finally equating the last test to test 1 to see whether, via the chain of equatings, the first test indeed is equated to itself. Such a procedure, however, does not address the crucial issue of subpopulation equating.

Although we have mentioned the dependence of equating procedures on a unidimensionality assumption, this is not strictly always necessary. If the tests to be equated are multidimensional but the loadings for the test items are such that the raw score or the unidimensional summary score (see below) represents the same function of the underlying factor for each test, then both raw score and IRM equating procedures are theoretically possible. In practice, this is unlikely to occur, but there may be situations where it is approximately true, and this is another topic for empirical investigation. (See Section 20).

18. Trends over time

A particular form of equating which arouses considerable interest is that whereby two different tests are administered to samples from the same population at different times. Generally, the same test cannot be used on more than one occasion because knowledge of the items may have become available so that items which are known in advance will acquire higher success rates and hence the parameters of an IRM fitted to them will change. Also, because curriculum objectives, etc., change over time, new items to cover such changes will need to be introduced and old and less appropriate ones eliminated.

When a test is multidimensional there is no guarantee that the dimensions at a second occasion will appear in the same form or with the same interpretation as at the first occasion. In fact, the updating of items in the test militates against this. More importantly, even if the same dimensions could be identified, we still have different items at each occasion so that identical solutions at each occasion are impossible. In particular, it is almost axiomatic that for those items which are discarded or newly acquired, their parameter values, for example their difficulties, implicitly change between occasions; that was why they were discarded or newly introduced. Nevertheless these items will contribute to the factor definitions, and conversely the factors are related to the items through the loadings. Instabilities or changes in the loadings imply a corresponding change in the interpretation of the factors.

There is a further fundamental problem with this form of equating which follows from the fact that the item parameter values, for example the difficulties, are defined with respect to the current population. Suppose, for a single item used at two occasions, that its difficulty increases from occasion one to occasion two. We have no way of knowing whether we should interpret this as the item becoming more difficult, or as the population becoming less 'able'. In constructing a common scale extending over time, we are free either to 'anchor' the parameter values of the items or of the subjects, or, indeed, some combination of them. Thus, if we carry out an equating or scaling so that items retain the same parameter values over time, this merely reflects an *assumption* that it is the population which has changed rather than the items. This fundamental ambiguity of interpretation effectively precludes statements concerned wtih absolute changes over time.

19. Item banks, two stage and tailored testing

We have already touched upon the idea of an item bank or pool. The fundamental element of this is a collection of items assumed to have a unidimensional structure, and there may be several such sets representing different domains. Note that we are not concerned here with the looser use of these terms to denote a collection of items available for test construction, rather with a structured set of items intended to conform to an IRM. Users of such a bank or pool are invited to make a selection of items to form a test, and since the item parameters are assumed known, on the basis of previous testing, any of the tests so formed will be capable of giving ability estimates for individuals. Thus, a user might wish to do this in order to replace a known test with a new one, and this can be viewed as an example of test equating and our previous remarks apply.

In addition to this use, however, item banks are sometimes advocated on the grounds that users will actually differ in their requirements, some wanting more items of one kind rather than another. Since the items are all intended to share a single common dimension these users can be satisfied in this respect and still have a common, unidimensional, scale along which to compare individuals. Unfortunately, such an argument is inconsistent. Suppose a set of items is truly unidimensional. Aside from considerations of estimation efficiency, it is a matter of indifference which kinds of items a user chooses since the same ability is being estimated by each one. Once it is admitted that different users require different kinds of items, then this becomes a tacit admission that the item set is not unidimensional with a common set of parameter values.

Suppose the true model for an item set incorporates group differences, which are not recognized in the operation of the item bank. Users are invited to select items for their own use, and to treat the parameters, e.g. the difficulties, as norms for their own populations. Thus, a user may well be able to select items especially 'favourable' to her own subgroup and hence be able to demonstrate an ability in excess of that obtained with, say, a random selection of items. Indeed, it is perfectly possible theoretically for every user to demonstrate a greater than average ability for their own subgroup! We should note also that this situation is generally true where users can choose to adopt any published test with national norms attached. By a judicious choice users may be able to present a particularly favourable picture for their own group. Phillips & Finn (1988) cite this as one possible reason for the so called 'Lake Wobegon Effect' where it was found that each of the 50 states of the United States had a mean score on certain standardized tests above the national mean.

An application of vertical equating procedures is to so called two-stage testing. In this, a first stage, usually fairly short, test is given to all subjects and then each subject is allocated one of a number of second stage tests. These second stage tests differ in their average difficulty and the allocation of them is made on the basis of the first stage score. Those subjects with the highest scores are given the most difficult test and so forth. The idea behind this is to give every subject a test which contains some easy (for them) and some difficult items—thus better to discriminate between subjects. A common scale for all the tests is formed by equating together all the second stage tests and the first stage test and hence is a case of vertical equating. Thus, the remarks made in Section 17 apply to this procedure, particularly so since several tests are being equated, not just two.

The unidimensionality assumption built into this procedure may produce some particularly undesirable effects when it is violated. Thus, for example, since the first stage test typically is short, many dimensions present in the second stage tests will be poorly represented if at all, and subjects who have high ability just on those dimensions will tend to perform poorly and be allocated to a less difficult second stage test. Because only one dimension is assumed to exist, each second stage test may contain different weightings for the several dimensions present, so that no fair single score summary is possible.

A further extension of two stage testing is so called 'tailored' or 'adaptive' testing. Here, the administration, typically by computer, is an item at a time with the choice of next item dependent on the response to the previous one. The items are all calibrated in an item pool, for example using a unidimensional 3-parameter logistic IRM (see Lord, 1980, and Hambleton, 1989, for details). Again, the procedure is heavily dependent on the unidimensionality assumption and similar remarks apply to it as to two-stage testing.

It would seem possible with two-stage and tailored testing schemes to adapt them to multidimensional item sets. Thus, for example, a first stage test of adequate length could be used to estimate scores on several dimensions and then several different second stage tests, each one strongly representing a particular dimension, applied to obtain more accurate estimates for each dimension. Naturally, the administration of such a procedure would be more costly, but it could avoid the more serious problems of unidimensional procedures.

20. Multidimensional item response models and unidimensional summaries

For simplicity we deal with the two dimensional model

$$\operatorname{logit}(\pi_{ij}) = a_i + b_i f_{1j} + c_i f_{2j}. \tag{20}$$

In order to fit a multidimensional model, as in ordinary factor analysis, we specify that the f_{1j}, f_{2j} are standardized random variables and in the usual formulation are also uncorrelated.

We shall not go into the various issues surrounding the interpretation of multidimensional IRMs. These parallel those of ordinary factor analysis (McDonald,

1985) with the additional complication imposed by the choice of transformation function. As with unidimensional models there are also the issues of models for different populations, use of covariates and so forth. No essentially new issues arise in these respects and the translation of the previous discussion to multidimensional models is straightforward. As yet these models have had little empirical exposure and hence their usefulness largely remains to be demonstrated. There is, however, one important area where there is something to be said; namely where a unidimensional model is fitted to multidimensional data.

In areas such as education, with diverse curricula, styles of teaching, etc., it is *a priori* reasonable to assume that tests of attainment or 'ability' will be multidimensional. At the same time it should never be overlooked that homogeneity of instruction or preparation may convert multidimensional responses into unidimensional ones. We have already explored the consequences of a failure to distinguish between separate subpopulations when considering unidimensional models. Now we look at what can happen when we fail to distinguish more than one dimension.

Several authors have studied the effect of fitting unidimensional models, typically the 2- and 3-parameter logistic fixed effects model, when at least two ability dimensions are present. Thus, for example, Drasgow & Parsons (1983) show that, for models with a general common factor and group factors (with non-zero loadings only for a particular group of items), the unidimensional fit tends to emphasize different item parameters depending on the relative importance of the different factors. Yen (1984) studied the effect of different discrimination parameters. In general, 'minority' dimensions will be poorly respresented in 1-dimensional fits and this will be the case whatever transformation is used.

Suppose we have a multidimensional IRM

$$y_{ij}=\Sigma_k c_{ik} f_{kj}+e_{ij}, \tag{21}$$

where y_{ij} may be any transformed item response. We now estimate a unidimensional model, say

$$y_{ij}=a_i+b_i f_j+u_{ij}. \tag{22}$$

The estimates of a_i, b_i are weighted functions of the y_{ij} and hence of the *c*s in (21). Hence by varying the latter, that is by varying the relationship between the y_{ij} and the factors, or by selecting items with particular value of the *c*s, we will vary the estimates of the a_i, b_i and also the average or expected values of these.

If we choose a set of items and have no knowledge of the actual factor structure (21), nevertheless that structure, among the items, will determine the parameters of any unidimensional model which is fitted. If we choose a set of items about which we do have knowledge concerning the loadings in (21), then for the particular unidimensional model we choose to fit we can fix the parameter values in (22) by suitable choices of items (or individuals). Reckase, Ackerman & Carlson (1988) discuss the special case where items are selected so that each has the same set of values of the *c*s. As pointed out in Section 17, this effectively produces a unidimensional test, which in turn means that information about the separate underlying factors cannot be recovered.

It follows, therefore, that where a multidimensional structure exists the choice of items to be included in a test will determine the parameters of a unidimensional model. These parameters are often complex functions of the parameters of the underlying multidimensional model and hence have in general no separate interpretation of their own. Where there is little knowledge about the underlying multidimensional structure and where attempts are made to obtain a unidimensional set of items by removing the 'misfits', any resulting unidimensional structure will inevitably reflect the original choice of items in relation to their underlying multidimensional structure.

21. Dynamic models

A feature of all the models we have described is their static nature. Only recently has there been any consideration of models which allow that responses to questions or items may depend on responses to previous questions or items (see Ackerman & Spray, 1987). In principle, however, there is no reason why we cannot model the probability of a correct response as a function of previous responses, together with measured individual and item or question characteristics.

Such dynamic models would lead to a more 'contextualized' view of individual responses, which recognized that individuals are actively interpreting the overall test situation, rather than providing a set of purportedly independent responses to items. Furthermore, such models have no need for particular dimensionality assumptions and to that degree may be able to provide more faithful descriptions of individual behaviour.

22. Practicalities

It should be clear from our exposition that in order to design a single summary test where items have a basic underlying multidimensional structure, it is likely that multidimensional factor analysis will have a role in assisting our understanding of the response patterns. In practice we would not in general wish to rely solely upon pure factor analytic models. There is often a great deal of knowledge about test items and information about individual characteristics which can be incorporated into models as covariates. Furthermore, test items are generally not selected haphazardly. The test constructor has in mind some idea of what factors or characteristics an item is supposed to reflect. All such information should be used when constructing a composite test.

The special case where there are only group factors is of some interest. Here, the weights can be approximated by choosing the number of items from each group in proportion to these weights and then forming an equally weighted raw score from the chosen items. This appears, in effect, to be what is done by index constructors. The index constructor acts as if each item in the index loads on a single factor and then selects items to correspond to predetermined weights as described. A similar procedure is often followed by constructors of educational tests, typically relying on prior knowledge or beliefs to classify items into distinct groups.

Thus, the procedure we have outlined can be viewed as a modification of index construction techniques, whereby the item responses are regarded as being influenced by several factors, and where empirical factor analyses combined with prior knowledge, beliefs or theoretical assumptions are used to create the composite score or index to reflect predetermined weights.

There are many situations where composite scores either are unnecessary or misleading or both. Often it will be advantageous to provide a multidimensional summary, perhaps in terms of factor scores, or perhaps in terms of separate subtest scores for groups of items selected as reflecting particular dimensions, constructs, etc. A useful intermediate position might be to provide a small number of summaries or indices, each based upon a different weighting of factors. This can be particularly useful for a multipurpose test.

Finally, all the above remarks may be applied to items or questions with a continuous or pseudo-continuous score. These include essay or constructed responses as well as multiple choice items with no single 'correct' answer but different values attached to each choice.

23. Conclusions

Item response *theory* is a misnomer. It is all very well to talk about IRT models having been developed in the context of a long history of psychological theory about the processes involved when people answer questions (Thissen & Steinberg, 1988), but the truth is that the theory as operationalized in those models is desperately thin. It is not about why people get questions right or wrong, but about the supposed probabilistic nature of responses conditioned on an assumed trait called ability. It is no accident that Lord & Novick (1968) titled their book '*Statistical* Theories of Mental Test Scores'. When responses are modelled by a response function the model *is* the theory. Substituting IRM for IRT is both a more accurate description of what is going on, and also an acknowledgement of the threadbare nature of the theory—as *substantive* theory. Practitioners have been too content to estimate item and ability parameters and leave it at that, a tendency encouraged by computerized adaptive testing. We believe that a switch of nomenclature will have the effect of emphasizing the essential task of the interpreter which is to glean insights from the results of modelling, perhaps to the point where the models have actual heuristic value. Very little of this style of work has been seen in the first five decades, although this issue is starting to receive attention (see, for example, Thissen & Steinberg, 1988).

The lack of progress can be seen as a direct result of the tendency to exploit a particular model to the exclusion of others. In this paper we have brought out the unity of item response models by siting them within an explicit linear modelling framework. The logistic models, for which some have claimed a special status, can be seen merely to be one class out of many possible classes of models. While they do, in certain respects, possess convenient statistical properties, they have no special claim to any desirable theoretical properties. Indeed, in our view, their relative mathematical complexity is often a disadvantage since it may help to obscure substantive reality. In practice, the simple identity models used over the effective response range, typically give near equivalent results.

We have shown how the statistical assumptions made in all item response models are also made in a number of other procedures, even where an explicit item response model is not used. Thus, the unidimensionality assumption is typically a key assumption in item bias studies, where the assumption of a single underlying dimension is needed to detect discrepant items. Equating studies rely upon the notion of a common unidimensional scale onto which each test score can be transformed monotonically. Scaled item banks also operate on the assumption of underlying unidimensionality. Because the bank is meant to enable user choice, the satisfying of such choices will undermine any attempt to establish a unidimensional set of items.

In our view, there has been too little concern with the consequences of fitting unidimensional models when reality is multidimensional. We have shown how an assumption of unidimensionality in the presence of multidimensionality will produce a composite dimension, the characteristics of which will reflect the choices of the test constructor. Thus, in this situation there can be no claim for the 'objectivity' of item response model estimates, nor for other procedures which rely upon a unidimensionality assumption. We would suggest that the best scientific practice is for the test constructor or analyst to seek vigorously for ways to reject an assumption of unidimensionality. Unfortunately, this seems to be the exception rather than the routine.

We are sure that there is an important place for statistical modelling of test item responses. Nevertheless, we believe that much of the mathematical development of such models during the last five decades has concentrated too exclusively on the particular case of the unidimensional logit linear model and at the same time shown an over-obsessive concern for mathematical elaboration at the expense of substantive empirical explorations.

Acknowledgements

We are very grateful to Professor Rod McDonald and two referees for their helpful comments.

References

Ackerman, T. & Spray, J. A. (1987). A general model for item dependency. *ACT Research Report Series, 87-89*. Iowa City: American College Testing Program.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*, 2nd ed. Washington, DC: American Council on Education.

Angoff, W. H. (1986). Some contributions of the College Board SAT to psychometric theory and practice. *Educational Measurement: Issues and Practice*, **5**, 7–11.

Baker, F. B. (1965). Origins of the item parameters x_{50}, β as a modern item analysis technique. *Journal of Educational Measurement*, **2**, 167–180.

Baker, F. B. (1977). Advances in item analysis. *Review of Educational Research*, **47**, 151–178.

Barton, M. A. & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Report 81-20*. Princeton, New Jersey: Educational Testing Service.

Bejar, I. I. (1983). Introduction to item response models and their assumptions. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Bell, R. C., Pattison, P. E. & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement*, **12**, 15–26.

Binet, A. & Simon, T. H. (1916). *The Development of Intelligence in Young Children*. Vineland, NJ: The Training School.

Bock, R. D. (1983). The mental growth curve re-examined. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, **35**, 179–197.

Cox, D. R. & Hinckley, D. (1984). *Theoretical Statistics*. London: Chapman & Hall.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioural Measurements: Theory of Generalisability for Scores and Profiles*. New York: Wiley.

Drasgow, F. & Parsons, C. K. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, **7**, 189–199.

Etezadi, J. & McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika*, **48**, 315–342.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika*, **7**, 19–29.

Finney, D. J. (1944). The application of probit analysis to the results of mental tests. *Psychometrika*, **8**, 31–39.

Finney, D. J. (1971). *Probit Analysis*, 3rd ed. London: Cambridge University Press.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, **48**, 3–26.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, **33**, 234–246.

Goldstein, H. (1986). Models for equating test scores and for studying the comparability of public examinations. In D. L. Nuttall (Ed.), *Assessing Educational Achievement*. Lewes, Sussex: Falmer Press.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. London: Griffin; New York: Oxford University Press.

Guilford, J. P. (1954). *Psychometric Methods*, 2nd ed. New York: McGraw-Hill.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement*, 3rd ed. Washington: American Council on Education, Macmillan.

Holland, P. W. & Rosenbaum, P. R. (1985). Conditional association and unidimensionality in monotone latent variable models. *Research Report 85-47*. Princeton, New Jersey: Educational Testing Service.

Holland, P. W. & Rubin, D. B. (Eds) (1982). *Test Equating*. New York: Academic Press.

Hoover, H. D. (1987). Personal communication.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver, BC: Educational Research Institute.

Johnson, C. E. (1986). Qualifying tests for appointments-in-administration and executive officer recruitment: A linkage feasibility study. *Recruitment Research Unit Report 28*. London: Civil Service Commission.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, **36**, 109–133.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61**, 273–287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburgh*, **62**, 74–82.

Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, **35**, 42–56.

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, **53**, 161–176.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, **4**, 269–290.

Levy, P. (1973). On the relationship between test theory and psychology. In P. Kline (Ed.), *New Approaches in Psychological Measurement*. London: Wiley.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika*, **51**, 11–22.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7. 17 (4, pt. 2).

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, **28**, 989–1020.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement*, **8**, 147–151.

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

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores* (with contributions by A. Birnbaum). Reading, MA: Addison-Wesley.

Lord, F. M. & Stocking, M. L. (1985). Item response theory. In T. Husen & T. N. Postlethwaite (Eds), *International Encyclopedia of Education: Research and Studies*, 2745–2748. Oxford: Pergamon Press.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monograph*, No. 15.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, **6**, 379–396.

McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerised Adaptive Testing Conference*. Minneapolis: University of Minnesota, Dept. of Psychology, Psychometrics Program.

Phillips, G. W. & Finn, C. E. (1988). The Lake Wobegon effect: A skeleton in the testing closet? *Educational Measurement; Issues and Practice*, **7**, 10–12.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Raven, J. (1986). *Manual for Raven's Progressive Matrices and Vocabulary Scales Research Supplement*, No. 3. London: H. K. Lewis.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, **25**, 193–203.

Reese, R. A. (1986). Data analysis: The need for models? *The Statistician*, **35**, 199–206.

Reiser, M. (1983). An item response model for the estimation of demographic effects. *Journal of Educational Statistics*, **8**, 165–186.

Richardson, M. W. (1936). The relation between difficulty and the differential validity of a test. *Psychometrika*, **1**, 33–49.

Sananathan, L. & Blumenthal, S. (1978). The logistic model and estimation of latent structures. *Journal of the American Statistical Association*, **73**, 794–799.

Shavelson, R. & Webb, N. (1981). Generalisability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, **34**, 234–246.

Shepard, L., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. *Journal of Educational Statistics*, **6**, 317–375.

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, **7**, 215–231.

Terman, L. & Merrill, M. A. (1937). *Measuring Intelligence*. Boston, MA: Houghton-Mifflin.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, **51**, 567–577.

Thissen, D. & Steinberg, L. (1988). Data analysis using Item Response Theory. *Psychological Bulletin*, **104**, 385–395.

Thomson, G. H. (1947). Review of 'Probit Analysis'. *British Journal of Psychology (Statistical Section)*, **1**, 71–72.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston, MA: Houghton-Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, **16**, 433–451.

Thurstone, L. L. (1927). The unit of measurement in educational scales. *Journal of Educational Psychology*, **18**, 505–524.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review*, **35**, 175–197.

Thurstone, L. L. & Ackerson, L. (1929). The mental growth curve for the Binet tests. *Journal of Educational Psychology*, **20**, 569–583.

Trabin, T. E. & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Traub, R. E. & Lam, R. Y. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology*, **36**, 19–48.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, **11**, 1–13.

Tucker, L. R. (1987). Developments in classical item analysis methods. *ETS Research Report 87-46*. Princeton, NJ: Educational Testing Service.

White, P. O. (1979). A latent trait model for individual differences in speed, accuracy and persistence (abstract). In R. Wood (Ed.), *Rehabilitating Psychometrics*. London: Social Science Research Council.

Wilcox, R. R. (1980). Determining the length of a criterion-referenced test. *Applied Psychological Measurement*, **4**, 425–446.

Wood, R. (1987). *Measurement and Assessment in Education and Psychology*. Philadelphia, PA: Falmer Press.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, **8**, 125–145.

Received 30 November 1988; revised version received 22 March 1989