1996

Accounting for and quantifying dependencies in dichotomous test data

Barry Sheridan
Edith Cowan University

Edith Cowan University

Copyright Warning

EDITH COWAN UNIVERSITY

PERTH WESTERN AUSTRALIA

ACCOUNTING FOR AND QUANTIFYING DEPENDENCIES IN DICHOTOMOUS TEST DATA

Barry Sheridan

RESEARCH REPORT No: 16

April 1996

MEASUREMENT, ASSESSMENT and EVALUATION LABORATORY

ACCOUNTING FOR AND QUANTIFYING DEPENDENCIES IN DICHOTOMOUS TEST DATA

Barry Sheridan

RESEARCH REPORT No: 16

April 1996

Measurement, Assessment and Evaluation Laboratory
Edith Cowan University

This Report is based on a Paper presented at the
American Educational Research Association Meeting
New York, April 1996

ISBN: 0-7298-0295-7

ACKNOWLEDGEMENTS

The Secondary Education Authority, Perth, Western Australia
for permission to use data from the ASAT

and to

Professor David Andrich, Murdoch University
during the conduct of the research featured for,
and for a critical review of the draft revision of,
this Report

Accounting for and Quantifying Dependencies in Dichotomous Test Data

Abstract

This paper reports on a strategy employing the Extended Logistic Model of Rasch to quantify dependencies in dichotomous test data by providing standard errors of measurement which are commensurate with the loss in information arising from dependencies among subsets of items. Data collected for the Australian Scholastic Aptitude Test were analysed and shown to exhibit strong dependencies among the items involved. A technique is presented for estimating the number of equivalent independent items (found to be 74 in this study) to the original test number (100).

Key words: local dependence, dichotomous, Rasch, item analysis, subtest

Accounting for and Quantifying Dependencies in Dichotomous Test Data

1.0 Introduction

Despite the increased attention to performance assessment procedures in recent years, the multiple-choice format is still employed in many testing situations. They are used in assessing levels of achievement and as instruments to assist with job selection and placement of students in remedial courses in higher education institutions. However, one limitation of multiple-choice tests, with their emphasis on recall rather than on generation of answers (Wainer & Thissen, 1993), has no doubt hastened the present growth in development of performance assessment procedures. As Yen (1993) has observed, these procedures "...require qualitatively different performance of students than do multiple-choice tests" (p 187).

Another concern associated with multiple-choice tests is the need to account for the presence of local item dependence. While "items in traditional multiple-choice tests are usually carefully designed to be independent of one another" (Yen, 1993, p 187), the construction of specific forms of multiple-choice tests are such that dependence among items is an inevitable consequence of the test design. Because of the high profile placed on selection and achievement outcomes, it is important that the precision of test measures not be compromised by the presence of such dependencies among items.

In addressing these issues, attention is focussed in the present paper on dependencies among items. Interest in local, or conditional, dependence was renewed in the early 1980's, and a link between this issue and that of the attenuation paradox was raised by Andrich (1983, 1984). Most of the techniques for addressing this issue and reported at the time (Kelderman, 1984; Molenaar, 1983; Rosenbaum, 1984; van den Wollenberg, 1982) involved either the Simple Logistic Model (SLM) of Rasch or traditional-based item response models. A quite different conceptualisation for the resolution of the problem was provided by Andrich (1985b) who introduced the idea of combining sets of dichotomous items comprising a test into a smaller number of item groups, called subtests, and analysing the transformed data set with the Extended Logistic Model (ELM) of Rasch. This idea of

combining test items appeared later but in a different guise, where Wainer and Kiely (1987) introduced a new label, testlets, but the concept they espoused had the same logical basis as subtests within the Andrich formulation. However, while the conceptualisation of the problem area was similar in both cases, the approach adopted by Wainer and Kiely for the testlets notation towards the measurement situation was fundamentally different from that offered by Andrich, as discussed later.

Subsequent investigations in this area have concentrated on the testlet notion (Thissen, Steinberg & Mooney, 1989; Thissen, Wainer & Wang, 1994; Wainer & Lewis, 1990) but the measurement models employed are elaborations of the dichotomous situation only and do not address the fundamental structure of the multiple category format which is central to the whole strategy. On the other hand, the ELM addresses this very issue through the adoption of a sequential reparameterisation formulation. Here, a more parsimonious approach to a variety of testing situations presents itself, including the issue identified for the present study, that of accounting for conditional dependence among test items. The ELM also identifies as problematic those items that over discriminate, a feature associated directly with that of dependence between items (Andrich, 1985b). As highly discriminating items introduce a bias favouring one group against another (Andrich, 1985c, Masters, 1988) it is important that this problem also be addressed in the present context.

Another advantage of the technique developed by Andrich is that it also provides a direct means of investigating the theoretical framework guiding test construction. This aspect of the technique was explored by Sheridan and Puhl (in press) who examined the measurement properties of a 188 item multiple-choice test in common use in Australia, the English Skills Assessment (ESA) test, which was in turn adapted for Australian conditions from two prominent American tests: the Sequential Tests of Education Progress Series I, for grades 10 to 12, and the Descriptive Tests of Language Skills for College Freshman (ACER, 1982b). As the conceptual framework presented in the Test Manual for the ESA specified a design which increases significantly the likelihood of dependence between the individual multiple-choice items, Sheridan and Puhl demonstrated how the Andrich technique could account for dependencies and at the same time provide a means of assessing the theoretical basis of the test design itself.

Following this introduction, Section 2 presents a brief overview of the ELM where the emphasis is placed on the special features of the model relating to the dependency problem. In

Section 3, an investigation of the dependencies between the items of a widely accepted aptitude multiple-choice test is reported by examining the relationship of the dependencies to the specific design characteristics of the test. Section 4 provides an examination of the relationship between the effect of the dependences accounted for and the number of equivalent test items these data actually represent. The paper concludes in Section 5 with a discussion on the implications for measurement and test construction that these findings have for test analysts.

2.0 Theoretical Framework and Measurement Model

The essential difference between the Wainer and Kiely approach to the measurement situation relating to item independence and that provided by Andrich is a difference between traditional test theory (in the testlet situation) and that of providing parameter separation using appropriate sufficient statistics (as with Andrich). In the former situation, discrimination and/or guessing parameters are included in the models adapted from those developed by Birnbaum (1968), Bock (1972), Lord (1980) and Samejima (1969). On the other hand, Andrich (1985a) employs the ELM which uses the notion of thresholds between categories which are scored in accordance with the familiar Likert format.

Besides providing person free measurement in accordance with its properties as a Rasch model the ELM can account for, in a meaningful way, the threshold structure inherent with the scoring function for items employing an extended number of response categories. This model takes the general form where person n of ability β_n responds to item i of difficulty δ_i and where there are m ordered thresholds τ_{ki}, for k = 1, m, on the measurement continuum:

$$\Pr\left\{X = x;\beta_n,\delta_i,\tau_{ki}\right\} = \exp\left\{x(\beta_n - \delta_i) - \sum_{k=1}^{x}\tau_{ki}\right\} \Big/ \gamma_{ni} \qquad (1)$$

where the score $x \in \{0,1,\ldots,m\}$ and the normalising factor is

$$\gamma_{ni} = 1 + \sum_{k=1}^{m}\left\{\exp k(\beta_n - \delta_i) - \sum_{j}^{k}\tau_{ji}\right\}$$

The constraints $\sum_i \hat{\delta}_i = 0.0$ and $\sum_k \hat{\tau}_{ki} = 0.0$ are imposed, without loss of generality, for each item i in estimating these parameters.

Thresholds are conceptualised as a set of boundaries between the response categories of an item and specify the change in probability of a response occurring in one or the other of two categories separated by each threshold. These thresholds can also be reparameterised, through the category coefficient, to form a hierarchy of item parameters which are directly related to the Guttman (1954) principal components, where the number of parameters is governed by the number of categories in the scoring function of the item. For example, with four categories, three item parameters can be estimated. To date, four parameters have been identified and clarified, although it is possible to have more, provided the number of categories per item is greater than five. The reparameterised form of the general expression of the model presented in (1) is

$$\Pr\{x;\beta_n,\delta_i,\theta_i,\eta_i,\psi_i\} = \frac{1}{\gamma_{mi}} \exp\{-x\delta_i + x(m-x)\theta_i + x(m-x)(2x-m)\eta_i + x(m-x)(5x^2-5xm+m^2+1)\psi_i + x\beta_n\} \quad (2)$$

The item parameters are labelled, in hierarchical order, as *location* (δ_i), *scale* (θ_i) *skewness* (η_i) and *kurtosis* (ψ_i). In a real sense, the higher order parameters (from *scale* onwards) qualify the location of an item on the latent trait continuum, with the second parameter, *scale*, defining the unit of measurement for that item. If the threshold estimates τ_{ki} for a particular item do not appear in a sequential, ordered, manner then this is evidence of misfit to the construction of the model (Andrich, 1985a; Sheridan, 1993). Threshold disorder can often provide valuable insights into the nature of the variable under review.

The reparameterisation of the thresholds creates a model that is very versatile. Threshold order can be assessed usually by examining the threshold estimates directly and the alternative reparameterised estimates consulted for additional, more specific, information. With the familiar Likert format, the threshold estimates relate directly to the boundaries between categories whose meaning is clear within a sequence such as "Strongly Agree, ... , Strongly Disagree"; "Always, ... , Never", and so on. In other situations, however, attention must be focussed on the second-order, or *scale*, parameter where the summary information provided assists with an understanding of the technique of combining sets of dichotomous items into subtests within a test. In this case, the total score obtained

for a set of individual dichotomous items provides a multiple category scoring function for the subtest, where the scores range from 0 (no items correct) to a maximum value equal to the total number of items in the subtest. As *different* combinations of items can produce the *same* total score, the association between the scoring sequence across categories does not have a unique meaning. Therefore, threshold disorder is not directly interpretable as is the case with the Likert format, so attention must be directed to the reparameterised formulation of the category function (Andrich, 1985a).

For the purposes of the present paper, attention is now focussed on an understanding of the meaning and interpretation of the second order, or *scale*, parameter. Andrich (1985b) demonstrated that a meaningful relationship exists between the difficulty estimates of the individual dichotomous items within a subtest and the degree of dependence between these items, and that this association could be captured in terms of the average half-threshold distance ($\overline{\theta}$) for each score, m_i, of item i under ideal conditions. This special set of values for the scale parameter provides an upper-bound value for the parameter for each value of m_i and applies when all items in a subtest have equal difficulty estimates while at the same time exhibiting no dependence between them. This situation is represented schematically in Figure 1 (structure [a]). The behaviour of the scale parameter (θ_i) when the condition of equal item difficulties is relaxed in the presence of item dependence therefore constitutes the assessment of dependence between items. As a consequence, the presence of dependencies between items of a subset can be detected by examining the relationship between the size of the scale estimates relative to the upper bound value derived for the items of the subset.

Consider the sequence of steps involved in this process and as summarised schematically in Figure 1. If the constraint of equal difficulty of items in a subset is relaxed (when the items are independent and as presented in structure [c] in Figure 1), the response distribution becomes more peaked, resulting in an increase in the threshold distance beyond the upper bound value (see Andrich, 1978). Conversely, if dependence between items is present but the item difficulties are equal (as represented in structure [b] of Figure 1), the effect is for increased responses in the extreme categories to the exclusion of the middle categories, resulting in a smaller value of the threshold distance associated with a flatter response distribution. These two features of unequal item difficulties and

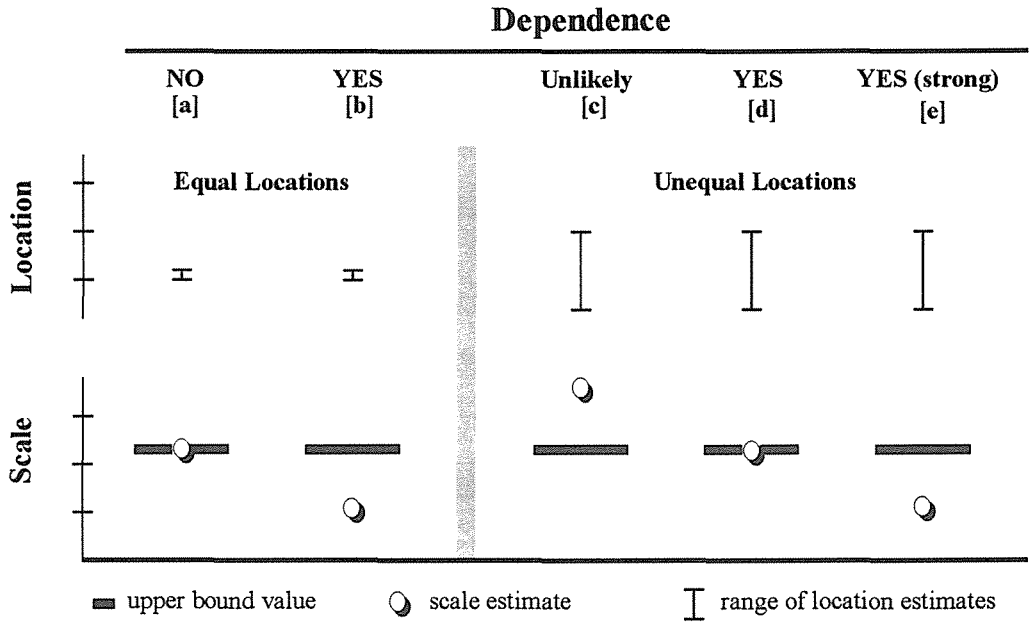

Figure 1 Schematic representation for detecting dependencies in terms of the Distribution of Location Estimates for the Component Items of a Subtest and the size of the Scale Estimate of the Subtest relative to the Upper-bound value.

dependencies between items work against each other and thus provide the basis for detecting the presence of dependencies, as the threshold distance incorporates any differences in difficulties and the averaging effect of dependencies among items of a subset. The usual cases present in real testing situations are represented by structures [d] and [e] in Figure 1. The way the threshold distances are interpreted in an analysis of a test likely to involve dependencies, and how the ELM can account for these dependencies in providing improved parameter estimates, is now considered using data collected for a multiple-choice type aptitude test in common use in Australia.

3.0 Methods and Techniques

This investigation involves the Australian Scholastic Aptitude Test (ASAT) which is employed for scaling examinations used for entrance to universities. The test is composed of 50 multiple-choice items designed to measure humanities and social science concepts (HUM/SocSc) and 50 multiple-choice items related to mathematics and science concepts (MATH/Sci). For the HUM/SocSc section of the test, a series of text passages are followed by a set of multiple choice items, where each set relates to the content of the passage immediately preceding it. A similar logic is followed for the MATH/Sci segment,

where sets of items are related to an initial statement describing a system of interest or a process specific to some scientific or mathematical principle.

The structure of the ASAT is presented in Table 1. For the purpose of this paper, the term "subtest" will be used to designate a group of items within the ASAT while the term "item" will refer to each component item within the original construction of the ASAT.

Table 1
Subset Structure of the ASAT

Discipline Group	Label Code	Max Score	Original Item Numbers per Subset							
HUM/SocSC	VA01	4	1	2	3	4				
HUM/SocSC	VB02	5	5	6	7	8	9			
HUM/SocSC	VC03	3	10	11	12					
MATH/Sci	SA04	8	13	14	15	16	17	18	19	20
HUM/SocSC	VD05	8	21	22	23	24	25	26	27	28
MATH/Sci	SB06	5	29	30	31	32	33			
HUM/SocSC	VE07	5	34	35	36	37	38			
MATH/Sci	SC08	7	39	40	41	42	43	44	45	
MATH/Sci	SD09	5	46	47	48	49	50			
MATH/Sci	SE10	6	51	52	53	54	55	56		
HUM/SocSC	VF11	6	57	58	59	60	61	62		
HUM/SocSC	VG12	5	63	64	65	66	67			
MATH/Sci	SF13	8	68	69	70	71	72	73	74	75
MATH/Sci	SG14	5	76	77	78	79	80			
HUM/SocSC	VH15	6	81	82	83	84	85	86		
HUM/SocSC	VI16	8	87	88	89	90	91	92	93	94
MATH/Sci	SH17	6	95	96	97	98	99	100		

The number of items per subtest varies, ranging from a minimum of three items in one subtest (VC03) to a maximum of eight per subtest. Of the 17 subtests present, nine subtests derive from the 50 HUM/SocSc items and eight subtests from the 50 MATH/Sci items. The column headed "Label Code" in Table 1 lists the identification tag to be used for each subtest in the analysis described in Section 4.3.1 Accounting for dependence.

Because a subtest structure occurs as a deliberate consequence of the design of the ASAT, it is likely that dependencies exist between the items within each subset.

Therefore, items from the same subtest are likely to violate local independence across the whole set of items — items within a subtest would be more dependent than items from subtests. Within each subtest, dependence is accounted for by variation in the scale parameter — the smaller the value, the greater the dependence. The importance of accounting for this dependence is that it provides standard errors of measurement which are commensurate with the loss of information arising from dependencies among subtests of items. In contrast, if one *assumes* without qualification that local independence holds equally well among all items, then the standard errors of measurement will be smaller than they should be.

3.2 Quantifying dependence.

Using the inflation in information when all items are assumed equally locally independent, and comparing with the more accurate information when dependencies are accounted for, it is possible to quantify the relative dependence among items. In particular, it is possible to estimate the information in terms of an equivalent number of locally independent items. Details on this technique appear in Section 4.

3.3 The Sample

The data for the present analysis was obtained from the version of the ASAT administered to Western Australian secondary school students sitting for the tertiary entrance examinations in November 1989. From a total population in excess of 15,000, the responses of 500 male and 500 female students were selected at random to provide the calibration sample for the analysis. These data were collected at the time the students sat for their tertiary entrance examinations as the culmination of five years of secondary schooling.

The computer program used to analyse these data was RUMM (Andrich, Lyne & Sheridan, 1995), a program for analysing test data using Rasch Unidimensional Measurement Models including the Extended Logistic Model (ELM). All techniques described in the next Section are available in the RUMM program.

4.0 Results

Responses of the calibration sample were analysed in two stages. The first stage involved an analysis of the original 100-item dichotomous-scored test followed by an analysis of the 17 subtests as displayed in Table 1. An examination of the test-of-fit statistics for the individual dicotomous items within each subtest points to the presence, or otherwise, of dependencies between these item groupings while the scale parameter estimates for the subtests provide additional information in this regard. The second stage of the investigation relates to the reliability indices for the two sets of analyses produced in Stage 1 and examines the number of independent component items that will produce measures equivalent to the subtest parameter estimates.

4.1 Test-of-fit and the Scale Parameter

The special interest in this first stage of the item analysis is the relationship between the tests-of-fit for individual items within a subtest and the size of the scale estimate for the subtest. Because Rasch models identify as problematic those items with unusually high or low discriminations, any pattern revealed within the subtest groupings would require further investigation, especially as the concept of discrimination is fundamental to item response theory. When the individual item-person interaction test-of-fit statistics were examined, a pattern did emerge across the subtest groupings. This pattern revealed a hierarchical ordering according to the nature and size of the discrimination evident in these fit statistics and that this pattern was repeated in the distribution of the scale estimates. As the item-person interaction fit statistics approximate a t distribution when items fit the model, values less than -2.00, or greater than +2.00, exhibit a departure from the model at the 5 percent level of significance. Further, a negative value for this statistic indicates that the item is fitting the model too well; the more negative the value the higher the item discriminates. Consider now those subtests with the lowest scale estimates. Table 2 displays details for subtests SD09, SH17, and SC08, those with the lowest scale estimates (θ_i = 0.104, 0.124 and 0.177 respectively), together with subtest VC03 which produced the highest scale estimate, θ_i = 0.529. The listing for each subtest contains the fit statistics for the component items of the subtest, that is, when these items are considered individually as the 100-item dichotomous-scored test. In Table 3, the location estimates

Table 2
Test-of-fit Statisitics for Subset Component Items by Scale (θ_i) Estimates
for Four Subtests of the ASAT

Subtest : SD09		Subtest : SH17		Subtest : SC08		Subtest : VC03	
θ_i = 0.104		θ_i = 0.124		θ_i = 0.177		θ_i = 0.529	
Items	Fit	Items	Fit	Items	Fit	Items	Fit
sd46	- 6.20	sh95	- 2.18	sc39	- 0.96	vc10	2.04
sd47	- 1.56	sh96	- 2.64	sc40	- 1.24	vc11	0.83
sd48	- 6.13	sh97	1.27	sc41	- 1.87	vc12	3.47
sd49	- 3.24	sh98	- 2.91	sc42	- 0.39		
sd50	- 1.38	sh99	- 1.66	sc43	0.26		
		sh100	- 4.38	sc44	0.69		
				sc45	1.07		

Table 3
Location Estimates for Subset Component Items by Location (δ) Estimates
for Four Subtests of the ASAT

Subtest : SD09		Subtest : SH17		Subtest : SC08		Subtest : VC03	
δ = 0.07		δ = − 0.14		δ = − 0.47		δ = − 0.30	
Items	Locn	Items	Locn	Items	Locn	Items	Locn
sd46	0.27	sh95	- 0.86	sc39	- 1.30	vc10	- 0.22
sd47	0.09	sh96	- 0.85	sc40	- 2.00	vc11	- 0.63
sd48	0.51	sh97	- 0.58	sc41	- 0.95	vc12	0.09
sd49	- 0.54	sh98	- 0.86	sc42	- 0.89		
sd50	0.40	sh99	0.03	sc43	0.38		
		sh100	0.60	sc44	- 0.61		
				sc45	- 0.68		
MEAN:	0.15		- 0.41		- 0.87		- 0.35
Av SE	0.07		0.08		0.08		0.07

for each subtest are displayed followed by the location estimates for those items comprising the subtest. For ease of identification, a subtest code has upper case letters (such as 'SD') while the component items of that subtest appear as with lower case letter (such as 'sd'). The numerals following each alpha code specify the sequence, or serial, order of the subtest, or item, within the ASAT.

The fit statistics for all five component dichotomous items of subtest SD09 are negative. This means that all five items comprising subtest SD09 overdiscriminate, with three of the five items (sd46, sd48 and sd49) highly discriminating. Subtest SH17 reveals a similar pattern but the degree of over discrimination is reduced compared to that for SD09. By exporing the pattern amongst the remaining subtests, as displayed in Figure 2, a trend

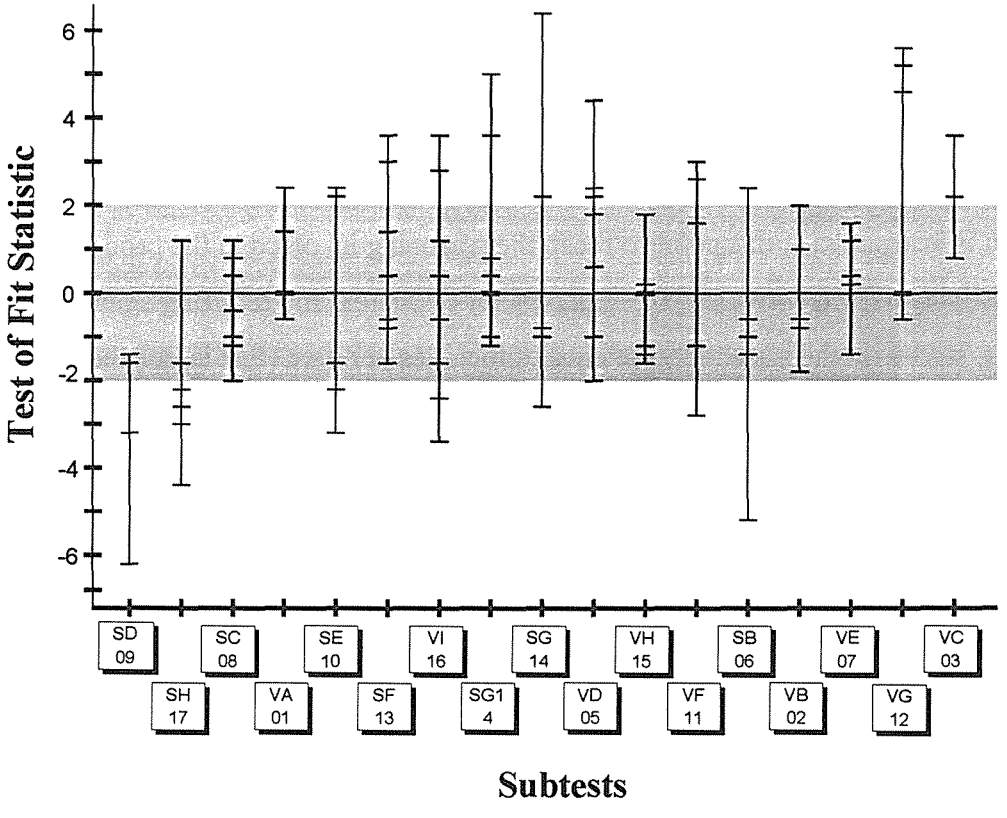

Figure 2 Distribution of test-of-fit statistics for component items within a subtest for all subtests in the ASAT.

becomes evident in which the sign of these statistics, collectively, slowly change from all negative (SD09) to all positive (VC03). This situation for subtest VC03 is the reverse of that observed in subtest SD09 whereby all component items are now underdiscriminating, two of them (vc10 and vc12) significantly.

As highlighted earlier in Section 2, the familiar Likert scoring function is employed for the 17 subtests. For example, as the first subtest (labelled VA01 in Table 1) contains four dichotomous scored items, the possible scores available for this subtest are 0, 1, 2, 3, and 4, making five response categories in all. With this structure, it is now possible to estimate additional item parameters per subtest beyond the single (location) parameter only that is

available for each of the original 100 dichotomous items. It should be emphasised that the number of parameters referred to in this discussion relate to the items only. There is, of course, a person parameter present as well.

Apart from subtest VC03 (for which only three parameters can be estimated from the four categories present) the remaining 16 subtests have sufficient categories to allow all four item parameters to be estimated for each subtest. However, and as also indicated earlier, the present analysis concentrates on the behaviour of the first two item (that is, subtest) parameters only. The first-order, or *location*, parameter specifies the average difficulty estimates for each subtest while the second-order, or *scale*, parameter provides evidence of dependence among component items of the subtest.

It is instructive at this point to recap the reasoning involved in the process for identifying dependence using these parameter estimates, and referring again to Figure 1. First, start with the assumption of equal difficulties per item (that is, for the components of a subtest) and no dependencies present between these items. Under these conditions, the scale estimate for the subtest is equal to the upper-bound value corresponding to the number of categories per subtest and as displayed in Columns 3 and 4 of Table 4. Next, observe that if a scale parameter estimate is less than the upper-bound value, then evidence for the presence of dependencies is revealed. Then, note the distribution of the location estimates for the component items within the subtest. If these estimates are not equal, then stronger evidence now exists for the presence of dependence because the scale estimate would need to be *higher* than the upper-bound value to counteract the opposing effect due to the unequal item difficulties.

An examination of Table 3 and Figure 2 for the individual location parameter estimates for the 100-item dichotomously-scored ASAT reveals that the set of estimates for the items within each subtest are not equal. A comparison between the scale parameter estimates, θ_i and the upper bound value for each subtest, $\overline{\theta}$, in conjunction with the knowledge that the difficulties of the component items are not equal, indicates that dependencies are present in all subtests. In Columns 3 and 4 of Table 4 the two sets of scale estimates for each subtest are listed, where it is observed that the upper-bound value varies with the number of items (that is, the number of categories) per subtest.

Table 4

Scale Estimates (θ_i) and Least Upper Bound Values ($\overline{\theta}$) for these estimates by Number of Categories (m+1) for all Subtests of the ASAT (N = 1000)

Subtest	No. categories m	Upper Bound $\overline{\theta}$	Scale Est θ_i	Dependence Present
VA01	4	0.41	0.22	Yes
VB02	5	0.35	0.31	Yes
VC03	3	0.55	0.53	Yes
SA04	8	0.22	0.23	highly likely**
VD05	8	0.22	0.24	highly likely**
SB06	5	0.35	0.30	Yes
VE07	5	0.35	0.35	highly likely**
SC08	7	0.25	0.18	Yes
SD09	5	0.35	0.10	Yes
SE10	6	0.29	0.22	Yes
VF11	6	0.29	0.27	Yes
VG12	5	0.35	0.39	marginal
SF13	8	0.22	0.22	highly likely**
SG14	5	0.35	0.23	Yes
VH15	6	0.29	0.25	Yes
VI16	8	0.22	0.22	highly likely**
SH17	6	0.29	0.12	Yes

** component items (for this subtest) are not of equal difficulty (refer to Table 3). In the absence of dependencies, the value of θ_i would be well above $\overline{\theta}$. Thus, dependencies must be present to supress θ_i to the level of $\overline{\theta}$ as noted in Table 4.

As Table 4 and Figure 3 reveal, dependencies are definitely present in more than half of the subtests (VA01, VB02, VC03, SB06, SC08, SD09, SE10, VF11, SG14, VH15 and SH17), while dependencies are highly likely in a further five subtests (SA04, VD05, VE07, SF13 and VI16). The prognosis for the remaining subtest (VG12) could best be described as marginal. In Figure 3, the subtests have been grouped by order of the upper-bound value whose value is represented by a black bar. The scale estimate for each subtest is represented by a [❍], while the range of location estimates for the component items of each subtest appear across the top of the diagram.

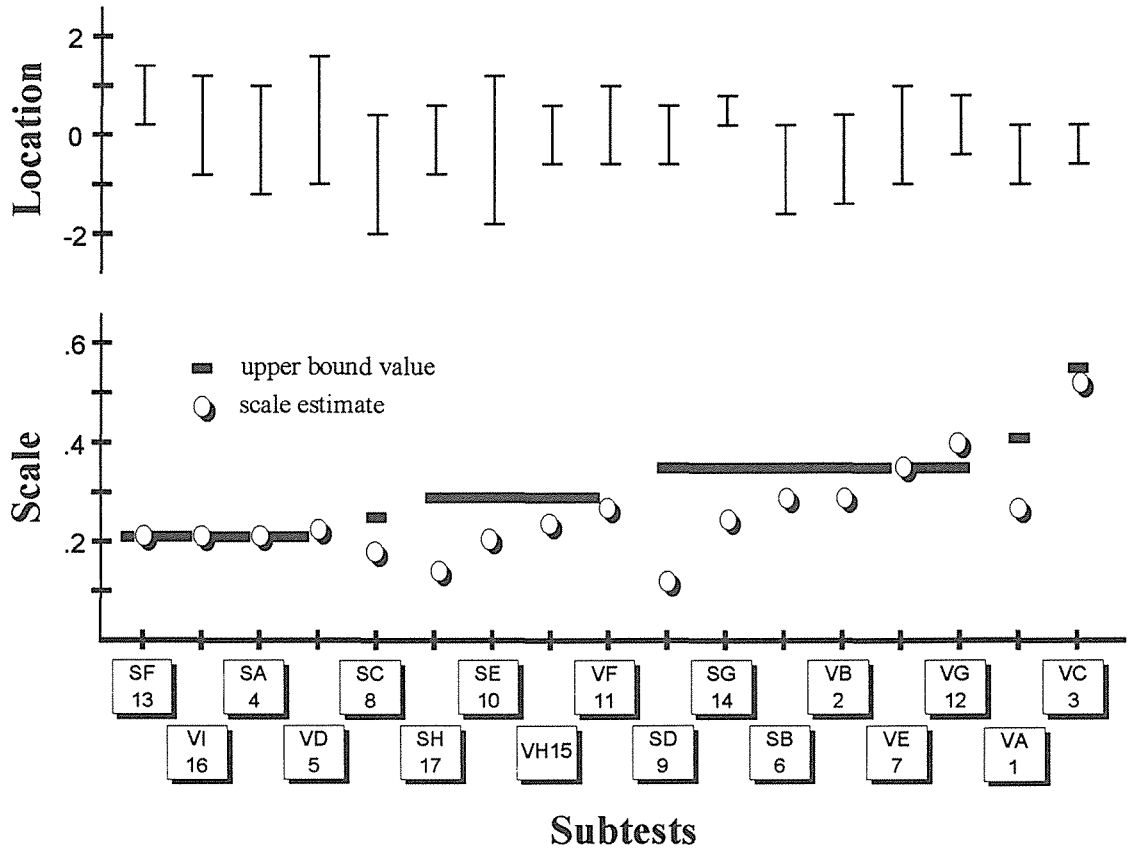

Figure 3 Scale estimate relative to the upper-bound value of each subtest of the ASAT in association with the distribution of location estimates for the component items of each subtest

4.2 Threshold estimates and the Scale Estimates

An examination of the relationship between threshold estimates for each subtest and the corresponding scale parameter estimate provides additional insights into the analysis of dependencies between test items. Unless there is a hierarchical ordering of categories for an item (as required for the Likert format), then that item exhibits an extreme pattern and is deemed to not fit the model. The set of threshold estimates for each of the 17 subtests is displayed in Table 5 and, apart from subtest SD09, the thresholds are ordered in the correct sequence across the categories. Appropriately, this same subtest with the reversed threshold estimates also has the lowest scale estimate ($\theta_i = 0.104$) which accords with the interpretation placed on threshold parameters. This is, high values of θ_i indicate large distances between thresholds — interpreted as an increase in the number of responses in the middle categories relative to the extreme categories — while low values imply increased responses located in the extreme categories. Low, even negative, values of θ_i

Table 5

Threshold Estimates for the 17 Subtests of the ASAT
(N = 1000)

Subtest	Threshold Estimates							
	1	2	3	4	5	6	7	8
VA01	-.0.64	- 0.12	- 0.01	0.078				
VB02	- 1.31	- 0.47	0.01	0.48	1.30			
VC03	- 1.10	0.08	1.02					
SA04	- 1.69	- 1.10	- 0.60	- 0.16	0.26	0.66	1.08	1.55
VD05	- 1.73	- 1.04	- 0.58	- 0.23	0.09	0.49	1.07	1.94
SB06	- 1.64	- 0.14	0.32	0.46	1.01			
VE07	- 1.49	- 0.54	- 0.01	0.52	1.52			
SC08	- 1.16	- 0.49	- 0.21	- 0.10	0.06	0.49	1.41	
SD09**	- 0.85	0.13	0.35	0.22	0.15			
SE10	- 1.30	- 0.67	- 0.08	0.41	0.75	0.89		
VF11	- 1.43	- 0.74	- 0.22	0.24	0.74	1.41		
VG12	- 1.53	- 0.79	- 0.05	0.74	1.63			
SF13	- 1.75	- 0.93	- 0.43	- 0.12	0.12	0.43	0.93	1.75
SG14	- 1.11	- 0.40	0.25	0.65	0.62			
VH15	- 1.52	- 0.45	- 0.02	0.17	0.49	1.34		
VI16	- 1.71	- 0.97	- 0.47	- 0.14	0.15	0.48	0.96	1.70
SH17	- 0.75	- 0.12	- 0.02	- 0.07	0.10	0.87		

** subtest exhibiting threshold reversal

can result when thresholds are reversed. Discussion on the nature and significance of thresholds in item analyses is found in Andrich (1978b; 1982; 1988); Andrich, de Jong, and Sheridan (1994); Andrich and van Schoubroeck (1989), and Sheridan (1993).

If the threshold estimates for the subtests with the lowest and highest scale estimates respectively are compared with the equivalent values for the binomial situation — where items exhibit independence and are of equal difficulty — the nature of dependencies is demonstrated further. As Table 6 indicates, the threshold estimates for a subtest with a low value of θ_i (equal to 0.18 with SC08, for example) are closer together than is the case for the binomial situation with the same number of categories (where $\overline{\theta} = 0.25$). Conversely, for a high value of $\theta_i = 0.53$ (for subtest VC03), the threshold estimates are in accord with the respective binomial values, where $\overline{\theta} = 0.55$. The corresponding values

Table 6

Actual Threshold Estimates and the Binomial Equivalents by Scale Estimates (θ_i) for Three Subtests of the ASAT (N = 1000)

Subtest	Threshold Estimates							Scale
	1	2	3	4	5	6	7	θ_i
SD09**								
Actual:	- 0.85	0.13	0.35	0.22	0.15			0.10
Binomial:	- 1.61	- 0.69	0.00	0.69	1.61			0.35
SC08								
Actual:	- 1.16	- 0.49	- 0.21	- 0.10	0.06	0.49	1.41	0.18
Binomial:	- 1.95	- 1.10	- 0.51	0.00	0.51	1.10	1.96	0.25
VC03								
Actual:	- 1.10	0.08	1.02					0.53
Binomial:	- 1.10	0.00	1.10					0.55

** subtest exhibiting threshold reversal

for subtest SD09 are also included in Table 6 where the presence of threshold reversal is evidence of even more extreme dependence.

Use of the least upper bound criterion ($\overline{\theta}$) for all 17 subtests of the ASAT reveals that dependence is present between component items in virtually all subtests. Even with the marginal case (VG12), as the observed scale estimate ($\theta_i = 0.39$) is only slightly higher than the upper bound value ($\overline{\theta} = 0.35$), it is reasonable to conclude that a competing dependence effect must be present among the items of the subtest to counter balance the elevating effect due to the unequal difficulties that exist between these items.

4.3 Reliability

The second stage of the analysis reported in this paper involves the nature of the reliability indices and in particular, the behaviour of the index of person separation, r_β. This index is similar in interpretation to the traditional reliability coefficients (Andrich, 1988), in particular, Cronbach's alpha, α (Cronbach, 1951) and "is constructed as the ratio of the estimated true variance among the persons and the estimated observed variance

Table 7

Ability Estimates and Reliability Index Estimates for the Dichotomous and Subtest Designs of the ASAT (N = 1000)

Dichotomous Analysis (100 items)		Subtest Analysis (17 items)	
Ability Range	Separation Index	Ability	Separation Index
- 4.89 to 4.85	0.923	- 4.22 to 4.23	0.896
Mean ability:	0.42	Mean ability:	0.33
SD ability:	0.84	SD ability:	0.62
Error Variance:	0.054	Error Variance:	0.040
Variance explained:	92.3%	Variance explained:	89.6%

among the persons using the estimates of their locations and the standard error of these locations" (Andrich & van Schoubroeck, 1989, p.483). Because the standard errors used for the estimation of r_β are assessed on an individual basis, this index is able to provide, routinely, a proper indication of the relative quality of the separation of the persons on the measurement continuum. This feature is not available to Cronbach's α, making the index restricted in meaning due to the necessity of knowing independently the dimensionality of the scale.

Table 7 displays the details of r_β together with related information on the mean of the person estimates, the standard deviation of these estimates, and the percent of variance accounted for by r_β. It is clear from the variance accounted for in the two analyses that the presence of dependencies within the subsets inflates the variance by three percent when the ASAT is considered as a 100 item dichotomous test.

To explore the meaning of the data in Table 7, the following elaboration is provided. By examining the association between r_β for the two forms of the ASAT, it is possible to assess the number of independent dichotomous items equivalent to the subtest design which accounts for dependence. Because the dichotomous case is the prime interest here, consider first the relationship between the error variance $V[E]$ (estimated at 0.0543 for the dichotomous analysis and as displayed in Table 7) and the model probabilities in terms of the existing number of items, N = 100:

$$V[E] = \frac{1}{\sum_i p_i(1-p_i)} = 0.0543 \qquad (3)$$

Then, on average, for the 100 dichotomous items:

$$V[E] = \frac{1}{100(p)(1-p)} = 0.0543$$

$$\text{or} \quad \frac{1}{(p)(1-p)} = 5.43 \qquad (4)$$

This value is then used to estimate the equivalent number of independent dichotomous items.

There are two ways to proceed from here. On the one hand, the equivalent error variance can be determined in terms of the observed person separation for the *dichotomous* case under the constraint of the subtest reliability estimate. Starting with (5):

$$r_\beta = \frac{V\left[\hat{\beta}\right] - V[E]}{V\left[\hat{\beta}\right]} \qquad (5)$$

and substituting the values $r_\beta = 0.896$ and $v\left[\hat{\beta}\right] = (0.84)^2 = 0.7056$ (refer to Table 7), the value for the error variance, V[E], is then estimated as:

$$0.896 = \frac{0.7056 - V[E]}{0.7056}$$

where
$$V[E] = 0.7056 - 0.6322$$
$$= 0.0734 \qquad (6)$$

By substituting this new value into (3):

$$V[E] = \frac{1}{\sum_i p_i(1-p_i)} = 0.0734 \qquad (3)$$

and solving for the number of items, N, in terms of the basis value for the model probabilities obtained in (4):

$$\frac{1}{N}(5.43) = 0.0734$$

provides the number of items, N, as:

$$N = 73.86$$
$$N \approx 74 \qquad (7)$$

The alternate approach involves deriving the equivalent error variance in terms of the observed person variance for the *subtest* situation under the constraint of the subtest reliability estimate. Starting again with (5) and substituting the values $r_{\beta} = 0.896$ and $v\left[\hat{\beta}\right] = (0.62)^2 = 0.3844$ (refer to Table 7), the value for the error variance, V[E], is then estimated as:

$$0.896 = \frac{0.3844 - V[E]}{0.3844}$$

where $V[E] = 0.3844 - 0.3444$

$$= 0.0400 \qquad (8)$$

By substituting this new value into (3):

$$V[E] = \frac{1}{\sum_i p_i(1-p_i)} = 0.0400$$

and solving for the number of items, N, in terms of the basis value for the model probabilities obtained in (4):

$$\frac{1}{N}(5.43) = 0.0400$$

provides the number of items, N, as:

$$N = 136 \qquad (9)$$

These results can be interpreted in one of two ways. If the observed person separation reliability (traditional) obtained for the dichotomous case is to prevail, then an increase in item number from 100 to 136 (a ratio of 136 to 100, or 1.36) would be required. Alternatively, the amended observed person separation obtained in the subtest case from the 100 dependent items, is equivalent to 74 independent items. It should be noted that these results are equivalent relative to the original 100 items: 74:100 = 1/(136:100).

5.0 Discussion and Conclusions

The focus for this paper has been the issue of dependencies between items of a test and whether conditional independence can be assumed routinely when using item response models. If dependencies are present, it is important that account be taken of such dependencies so that the precision of the test measures is not inflated. While one approach to this problem would be to restructure the presentation of the test to minimise the

likelihood of dependencies, it is not always practical, or even feasible, to adopt this strategy. As item structure is an integral part of test design, then for tests such as the ASAT and many reading comprehension and achievement/content orientated tests, the mechanism for analysing these test data must also be capable of accounting for the highly likely event that dependencies exist between items. That is, the dependencies must be quantified in terms of the original test design which, in turn, reflects the theoretical or conceptual framework and hence the fundamental basis of the measurement process itself.

The ELM discussed in this paper is a measurement model capable of addressing this issue. As presented, this model accounts for dependencies by estimating parameters for item subgroups, or subtests, such that these subtests are considered in the same scoring manner as the Likert format. Thus, if tests are structured in such a way that dependencies are highly likely, then the ELM would be suitable for employment in subsequent analyses to provide estimates of item parameters, in this case, of subtests as specified by the test design. Because of the way this model incorporates the scoring function for multiple categories per item, a more parsimonious solution is available to this problem of dependencies than those proposed in recent papers on this topic.

The method proposed by Bell et al. (1989), for example, requires a special procedure for the selection of the calibration sample from a large number of respondents. Under these conditions, a compromise is required to arrive at a manageable sample size. On the other hand, the ELM requires a modest calibration sample and two analysis runs only of the RUMM program — which incorporates the ELM — for assessing the extent of any dependencies that may be present. For the first run, the items are treated as individual dichotomously scored entities, while the second run employs the familiar Likert format by considering subtests as the test items, such that sets of the original dichotomous items provide the scoring categories for the respective subtest items. If dependencies are present, then the item location and scale estimates provided by the second run would be the values adopted for the calibration of the test. The estimates obtained from this second run provide the more accurate and precise calibration values for the item parameters and should be the ones used for the person measures derived from the test.

While the ASAT was found to exhibit dependencies between the items comprising the test, the extent of the dependencies was variable across the different subtest groups. This variation was detected directly from an examination of the scale value estimates associated

with the subgroups. In addition, it was demonstrated that the extent of the dependence within a specific subtest was directly related to the level of discrimination observed for the component items comprising the subtest.

In the case of the ASAT, it was also shown that the 100 items (which contain the dependencies) are equivalent to approximately 75 conditionally independent items. An examination of the degree to which the index of person-separation, r_{β}, obtained for the original dichotomous structure is reduced when the composite subtest format is adopted, provides a means of estimating the effective number of original items when the effect of the dependencies has been quantified. While the variance explained by the ASAT drops only three per cent when the subtest structure is employed, this reduction translates into a 25 per cent reduction in the number of items producing independent contributions to the actual test variance. The key to this issue is in determining the correct value for the error variance as the presence of dependencies reduces this value, thus resulting in an increased reliability index. It is this situation that leads to the so called attenuation paradox whereby increased reliability produces a decrease in the validity of the test (Andrich, 1983, 1984). Once the correct error variance has been quantified by accounting for the presence of highly discriminating items in the *original* test structure — through the employment of the subtest strategy devised by Andrich — the attenuation paradox no longer prevails.

To appreciate the merit of the ELM as a measurement model, it is important to understand the role that the threshold parameters play in assessing order among categories for an item. Threshold order is an informative indicator in this regard, especially as this ordering in not a requirement of the solution algorithm for the model (Andrich, 1985a). The continual references to hierarchical ordering within testlets (Wainer & Kiely, 1987; Wainer, H., & Lewis, 1990) is, if fact, leading the discussion away from the issue of dependencies within tests as originally presented by Andrich and as addressed in this paper.

The issue of discrimination is fundamental to an understanding of the process of dependencies between items. Masters (1988) has argued that respondents to a test who possess low ability, for example, are disadvantaged more than respondents possessing high ability "because of the greater penalty imposed for failing (an item with high discrimination) than for failing a less discriminating item" (p. 19). As argued in this paper, it is the ability of Rasch models to identify unusually high discriminating items as

problematical that leads to a clearer understanding of the process of dependencies between test items.

The use of highly discriminating items in tests is therefore not recommended as this leads to bias of one section of a population against that of another section. As many of the models recommended in earlier papers for addressing the issue of dependencies between items contain discrimination parameters, this situation must be cause for concern. In addition, these same models are each elaborations of the dichotomous situation only and do not address the fundamental structure of the multiple category format. On the other hand, the ELM obviates this problem. Further, and because of the sequential reparameterisation formulations employed, this model is capable of addressing the variety and range of testing situations presently confronting test analysts in a more parsimonious way than is possible with the models discussed in recent years in relation to the conditional independence problem.

The study reported in this paper shows that the measurement model employed accounts for and quantifies dependencies found in the test data analysed and that the multiple-choice format involved can best be accommodated by the creation of subtests as described. This technique can also be extended, as Sheridan and Puhl (in press) have demonstrated, to address the theoretical construct guiding the structure of extended multiple-choice tests leading to a more meaningful interpretation of the variable constructed to explain the measures derived.

References

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581-594.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105-113.

Andrich, D. (1983). The Attenuation Paradox in Latent Trait Theory. Paper presented at the Annual Conference of the National Council of Measurement in Education, Montral, Canada.

Andrich, D. (1984). The Attenuation Paradox of Traditional Test Theory as a breakdown of local independence in Person-item Response Theory. Paper presented at the Annual Conference of the National Council of Measurement in Education, New Orleans, U.S.A.

Andrich, D. (1985a). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.). *Sociological Methodology*, (pp. 33-80). San Fransisco: Jossey-Bass.

Andrich, D. (1985b). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*, (pp. 245-275). New York: Academic.

Andrich, D. (1985c). The social construction of education and psychological variables: A half century of unwitting bias in testing. Paper presented the Annual Conference of the Australian Association for Research in Education, Hobart, Australia.

Andrich, D. (1988). A general form of Rasch's Extended Logistic Model for partial credit scoring. *Applied Measurement in Education, 1*, 363-378.

Andrich, D., de Jong, J., & Sheridan, B. (1994). Diagnostic opportunities with the Rasch model for ordered response categories. Paper presented at the IPN Symposium on *Applications of latent trait and latent class models in the social sciences*, Kiel, Germany.

Andrich, D., Lyne, A., & Sheridan, B. (1995). RUMM (Version 2.0) [Computer software: A Windows program for performing item analyses according to Rasch Unidimensional Measurement Models]. Perth: Edith Cowan University, Measurement, Assessment & Evaluation Laboratory.

Andrich, D., & van Schoubroeck, L. (1989). The General Health Questionnaire: a psychometric analysis using latent trait theory. *Psychological Medicine, 19*, 469-485.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novivk, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Guttman, L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), M*athematical thinking in the social sciences*. New York: Free Press.

Kelderman, H. (1984). Loglinear Rasch model test. *Psychometrika, 49*, 223-245.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*, 15-29.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch Model. *Psychometrika, 48*, 49-73.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425-435.

Samejima, F. (1969). Estimation of latent ability using a reponse pattern of graded scores. *Psychometrika, 34*, Suppl. 17.

Sheridan, B.E. (1993). *Threshold order and Likert-style questionnaires.*. Paper presented at the Seventh International Objective Measurement Workshop, American Educational Research Association, Atlanta.

Sheridan, B., & Puhl, L. (1996). Evaluating an indirect measure of student competencies in higher education using Rasch measurement. In G. Engelhard, Jr. and M. Wilson (Eds), *Objective Measurement: Theory into Practice (Volume 3)*. (pp. 19-44). New Jersey: Ablex Publishing.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of Multiple-Categorical-Response Models. *Journal of Educational Measurement, 26*, 247-260.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123-140.

Wainer, H., & Kiely, G. L. (1987). Item clusters and Computerised Adative Testing: A case for Testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Lewis, C. (1990). Towards a psychometics for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Towards a Marxist Theory of Test Construction. *Applied Measurement in Education, 6*, 103-118.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.