Conditional Independence in a Clustered Item Test

Richard C. Bell and Philippa E. Pattison
University of Melbourne

Graeme P. Withers
Australian Council for Educational Research

Although the assumption of local independence underlies all latent trait theories in mental testing, it has rarely been empirically examined. In this study of a clustered item test (the Australian Scholastic Aptitude Test), a loglinear modeling approach was used to examine the conditional independence of items both within and between clusters. In general, although relationships between items were usually positive (as required for theories involving monotone item trace lines), conditional independence was not found. Departures from independence were more marked in items within clusters rather than between clusters, and also among items based on mathematical rather than verbal material. Another finding was the tendency for departures from independence to increase with ability (as measured by the score on other items).

The assumption of conditional independence is basic to most mental test theories (Lazarsfeld, 1950), especially so for latent trait theories (e.g., Birnbaum, 1968; Lord, 1953; Rasch, 1960). The assumption is expressed in relation to a latent variable U which a test is designed to measure, and takes the form of the conditional independence of item responses for given values of U. For example, the Rasch model proposes a latent "ability" parameter θ and proposes that an individual's response X_j to item j of a k-item test is governed by

$$P(X_j = x_j|\theta) = \frac{\exp[x_j(\theta - b_j)]}{1 + \exp(\theta - b_j)} \quad , \qquad (1)$$

where $x_j = 0,1$ are the possible responses to item j, and b_j is an item parameter describing the difficulty of item j. Implicit in the model is the assumption of "local" or conditional independence:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k|\theta) = \prod_j P(X_j = x_j|\theta) \quad . \qquad (2)$$

Despite the importance of the conditional independence assumption to item response models, the issue of its empirical status has been somewhat neglected (Goldstein, 1980). In 1953, Lord proposed a simple means of evaluating the assumption when he suggested a chi-square test for the independence of items for all persons at a given score level. From then until the end of the 1970s, however, the assumption of conditional independence was only used in the derivation of models, and its plausibility was not discussed. In their influential monograph, for example, Lord and Novick (1968) did not even refer to the simple early test devised by Lord. Other writers, such as Gustafsson (1980), simply used the approach devised by Lord and Novick to underpin theoretical speculation.

In the same journal issue as the paper by Gustafsson, however, Goldstein (1980) raised some problems of clarification in the traditional Lord and Novick approach to conditional independence and speculated that "the assumption of local inde-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 1, March 1988, pp. 15–26
© Copyright 1988 Applied Psychological Measurement Inc.
0146-6216/88/010015-12$1.95

pendence is such a strong assumption that it would be surprising if it were true other than in a few specially contrived circumstances'' (p. 239). Goldstein also showed that conditional independence was not equivalent to unidimensionality, a point recently reiterated by Hattie (1985).

Recently, renewed interest in the conditional independence assumption has emerged. Van den Wollenberg (1982) and Molenaar (1983) have extended Lord's (1953) proposal to provide a test of the fit of the Rasch model to a given set of data. Kelderman (1984) has developed the tests further by proposing a series of tests sensitive to different aspects of the Rasch model. His approach is based on the formulation of the Rasch model as a quasi-loglinear model. The model pertains to the $k(2^k)$ contingency table recording the frequency with which persons obtaining a given score on a k-item test possess each possible response profile. Included among the proposed tests is a test for the assumption of conditional independence, seen as a special case of the assumption of unidimensionality.

Holland (1981) has adopted a different and somewhat more general perspective to the question of conditional independence by establishing a condition necessary, but not sufficient, for conditional independence over a range of item response models. The condition, known as local non-negative independence, has been generalized further by Rosenbaum (1984) to provide a diagnostic test for the conditional independence assumption that does not depend on the assumption of a particular parametric form for the item response model. Rosenbaum suggested a test for negative association of an item pair which may be used to reject the assumption of conditional independence for all possible forms of monotone item characteristic curves.

Thus, for latent trait models such as the Rasch model, conditional independence should be evidenced by zero associations between items in, for example, the loglinear tests; nonzero associations would mean a failure to meet this requirement in such models. Positive associations, however, would indicate that some (as yet unspecified) models with monotonic trace lines may exist for which the assumption is satisfied, while negative associations would rule out the existence of any such models.

Despite these recent developments in techniques for investigating the conditional independence assumption, the only empirical research to date appears to be a study by Yen (1984). She compared a number of indices of departure from conditional independence in both simulated and real data. In the simulated data, departures were caused by the generation of data under a multidimensional model; in the real data, they were argued to follow from similarity of item content in the achievement tests used. Yen found that the type of index proposed by Lord and van den Wollenberg was able to detect violations of conditional independence for the three-parameter logistic model.

The purpose of this paper is to report a further empirical inquiry into the conditional independence assumption. This study adds to the results of Yen by investigating conditional independence in a clustered item or unit-structured test, and by using some procedures for detecting failure of local independence in the context of the Rasch model (Kelderman, 1984). Some results of Rosenbaum's (1984) diagnostic approach to examining the failure of conditional independence irrespective of item response model are also reported.

The present method for investigating conditional independence is based on a loglinear approach. Such an approach has become increasingly popular in the item-response model domain. Apart from Kelderman's (1984) formulation of the Rasch model in loglinear terms, Mellenbergh and Vijn (1981) have constructed a loglinear version of the Rasch model, while Bart and Palvia (1984) specifically examined interactions among items using a loglinear approach. It should be noted that although tests of fit for latent trait models, such as the Wright and Panchapakesan (1969) test or the Andersen (1973) likelihood ratio test for the Rasch model, are sensitive to departures from conditional independence, they are also sensitive to other failures of assumptions in the models and cannot be used to specifically identify failures of local independence.

The test for which these evaluations of conditional independence were conducted is the Australian Scholastic Aptitude Test (ASAT). In ASAT, the items are clustered into units about different stimulus materials. Such clustering is not taken into account in scoring the test, although the units determine how an item is to be classified in selecting the proportions of four broad domains in the test: Mathematics, Science, Humanities, and Social Science. A number of tests possess a unit structure of this kind, including many reading comprehension tests, sections of ETS admissions tests such as SAT, GMAT, and LSAT, and other tests such as the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980).

The assumption of conditional independence is probably less plausible for clustered-item tests than for tests whose items are not clustered according to stimulus material. The assessment of the assumption is therefore a particularly important concern where scaling models incorporating this assumption are to be used. In this initial study, an exploratory approach to the question was adopted. In particular, examination of conditional independence was restricted to items within the Mathematics/Science and Humanities/Social Science domains. Other issues investigated were independence of items within particular clusters (or units) of these domains, and relationships occurring across clusters which fall in the same domain.

Method

The Test

The ASAT is a 100-item multiple-choice test with four options per item. Items are grouped into units about stimulus material drawn from mathematics, science, social science, or humanities. Although the material is drawn from these content areas, no content-specific skills are required to answer the questions. Examinees are instructed to complete all items. There is no penalty for guessing.

Analysis

A constraint on any study of this kind is the amount of data required to test the assumptions. For a small unit of, say, 5 Mathematics/Science items, the number of cells in the contingency table required to represent all possible response profiles is $2^5 = 32$. If all possible score levels on remaining Mathematics/Science items, say $50 - 5 = 45$ score levels, are then considered, the approach to testing aspects of the Rasch model outlined by Kelderman (1984) would require $32 \times 45 = 1{,}440$ cells. To have reasonable numbers in each cell would require a very large sample. Accordingly, a number of constraints were imposed on the study:

1. Units in the Mathematics/Science domain were the main focus, although some illustrative analyses used Humanities/Social Science material.
2. The number of items in a table was restricted to less than 8. For larger units, items were split into two sub-units of sequential items (due to the belief that local dependence would be more likely to be "local," i.e., found in contiguous items). This was necessary in only one Mathematics/Science unit, but was necessary in all but one Humanities/Social Science unit.
3. Rather than considering separate score levels, the sample was split into deciles. Rosenbaum (1984, p. 428) hypothetically considered the use of deciles in a test of a complex case of his model. However, as a check on this grouping, data from four separate neighboring score levels were tabulated. If, say, the Rasch model were to hold for these data, then these score levels would almost directly represent four separate ability levels and any problems of heterogeneity within deciles would not be present.
4. As noted earlier, conditional independence was examined both within and across units, but the latter investigation was restricted to units in the same domain.

Data

The data used are from 30,465 examinees who were administered the 1984 version of ASAT-M in Queensland and Western Australia. In the decile approach, therefore, there were approximately 3,047 persons in each group; the task was then to divide them among the tables ranging from 8 (three-item) to 128 (seven-item) cells.

In the separate score level approach, a simplifying step was taken to select persons at four fixed score levels (21, 26, 31, 36) irrespective of the unit being considered, although these scores would not represent the same ability for different units. Thus the numbers of persons at each of these score levels varied with the unit being considered. This was only done for the Mathematics/Science section (because the only concern here was to establish the validity of the decile approach); numbers in these score groups ranged from 938 (score level 21 for Unit 11) to 1,428 (score level 31 for Unit 9).

Models

The conditional independence assumption was investigated by attempting to fit an independence model to each of the tables using the loglinear modeling available in the computer program GLIM (Baker & Nelder, 1978). If such a model could not be fitted, then the nature of any failure of fit was elaborated by the fitting of models with (1) all two-factor (two-item) interactions for consecutive items, (2) all two-factor interactions, and (3) if necessary, all three-factor (three-item) interactions.

Model parameters computed by GLIM differ from the "usual constraints" parameters of normal loglinear modeling (Bishop, Fienberg, & Holland, 1975) in that the GLIM parameter for a particular interaction between items is calculated from a different computational base. In particular, the GLIM parameter is derived from data of persons obtaining a score of 0 on all items in the subset not involved in the interaction, rather than from all data in the table. In the absence of higher-order interactions, the two sets of parameters should not differ greatly. Accordingly, several sets of item parameters were calculated under the "usual constraints" as well as under the default GLIM constraints in order to assess their similarity.

Results

Preliminary Results

Deciles versus separate score groups. Parameter estimates for pairs of items were obtained for four separate score levels in four Mathematics/Science units and were also obtained with the deciles containing these four score levels. There were 84 item pairs thus considered. When score level estimates were regressed on the decile estimates, a slope of .95, an intercept of 0.0, and a standard error of estimate of .15 were obtained. The correlation between the two sets of estimates was .97. Hence it was concluded that the decile grouping did not introduce substantial heterogeneity of ability into the procedure, and accordingly this approach was used in the remainder of the study.

GLIM parameters versus usual-constraints parameters. A detailed comparison of GLIM and usual-constraints parameters was conducted for items 20 and 21 from Unit 3 (a mathematics-based unit). It was found that the parameters were identical (to three significant digits) despite their different computational bases. As a result of this comparison, it was decided to use the GLIM parameterization because it avoided some problems associated with usual-constraints estimation in large tables by GLIM.

Conditional Independence of Items Within Units

Models were fitted to the data for all Mathematics/Science units and for two Humanities/Social Science units. One Mathematics/Science unit was split (Unit 6), as was one Humanities/Social Science unit (Unit 15). A summary of model fit is shown in Table 1.

In no case did the independence model provide an adequate fit to the data, although it can be seen from Table 1 that independence was more strongly violated in some units than in others. The data for most units could be adequately modeled when all two-item interactions were included in the model specification; in many cases, not all two-item interactions were necessary for adequate fit. Only one unit showed strong evidence of a three-item interaction. This was in a mathematics-based unit (Unit 7), where the interaction was largely among the last three items and took the form of a lower than expected number of persons giving a correct response to all three items. For the most part, models

Table 1
Number of Score Levels for Which the Model Fits at the .05, .01, and .001 Levels for the Independence Model and for the All 2-Item Interaction Model

	Independence Model			All 2-Item Interaction Model	
Unit	.05	.01	.001	.05	.01
Math/Science Unit					
1 Science	0	0	0	10	10
3 Math	0	0	0	9	9
4 Science	0	0	0	10	10
6i Science	0	0	0	8	9
6ii Science	0	1	1	10	10
7 Math	0	0	0	1	4
9 Math	0	0	0	7	8
10 Science	0	0	0	10	10
11 Math	0	0	0	8	8
13 Math	0	0	0	10	10
14 Science	0	0	0	8	9
Humanities/Social Science Unit					
12 Soc Science	0	0	0	9	10
15i Humanities	0	0	0	8	9
15ii Humanities	0	0	0	10	10

characterized by two-item interactions for consecutive items did not provide adequate fit to the data, although they did yield much better fit than the independence model.

In order to describe the nature of the dependence between the items, the estimates of interaction parameters were calculated for Units 3 and 6 (the latter split into two sub-units) based on Mathematics/Science material, and for Units 12 and 15 (the latter again split) based on Humanities/Social Science material. Table 2 shows the pairwise estimates for Unit 3.

Although not shown here (the tables are available from the authors), most (81%) of the parameters for the four units were positive, indicating an increased likelihood of a correct response to one item if another is correctly answered. The main exception to this was the last item of Unit 6 (item 46), which had a small negative relationship with the other items in the second half of this unit. In another study (Withers & Bell, 1985), this item was shown to perform poorly in that good students could only eliminate one incorrect alternative and could not discriminate among the other three. Among the Humanities items, negative interactions were

Table 2
Parameter Estimates for Item Pairs in Unit 3 for the Model With All Two-Item Interactions

Score Group	Item Pair		
	20-21	20-22	21-22
1	.98	.13	.25
2	1.25	.54	.30
3	1.52	.50	.45
4	1.69	.91	.31
5	1.55	.67	.38
6	2.06	.63	.58
7	2.12	.75	.68
8	2.08	.54	.53
9	2.02	.44	.86
10	2.86	.20	.98

obtained between items 79 and 80 in Unit 12, and between item 95 and items 91 through 94. In this latter case it is interesting to note that item 95 pertains to the second passage in the unit whereas items 91 through 94 pertain to the first.

The finding of generally positive relationships between items was confirmed using a test suggested by Rosenbaum (1984). For a given pair of items, a test of the hypothesis of an odds ratio of less than 1 may be conducted. The test relies on the Mantel-Haenszel procedure; the statistic whose values are given in Table 3 is referred to the lower tail of the standard normal distribution. The statistic was calculated for Units 3, 6, 12, and 15; as Table 3 indicates, all units demonstrate non-negative association for almost every pair of items tested.

Table 3
Tests for Negative Item Association
in Some Units of ASAT

Unit and Item Pair	z	Unit and Item Pair	z
Unit 3: Math		Unit 15i: Hum	
20-21	50.54	91-92	4.02
20-22	21.58	91-93	33.92
21-22	21.75	91-94	7.90
Unit 6i: Sci		91-95	-5.01
39-40	11.34	92-93	-.30
39-41	18.28	92-94	2.95
39-42	17.00	92-95	3.75
40-41	11.13	93-94	11.27
40-42	7.10	93-95	-1.40
41-42	64.12	94-95	1.82
Unit 6ii: Sci		Unit 15ii: Hum	
43-44	5.83	96-97	10.30
43-45	13.44	96-98	15.89
43-46	-1.79	96-99	8.66
44-45	5.59	96-100	3.32
44-46	2.34	97-98	12.28
45-46	-1.49	97-99	11.13
Unit 12: Soc Sci		97-100	6.32
77-78	15.88	98-99	16.76
77-79	11.38	98-100	12.43
77-80	.89	99-100	11.50
77-81	10.72		
77-82	2.83		
78-79	4.73		
78-80	3.37		
78-81	10.31		
78-82	4.97		
79-80	-9.06		
79-81	5.84		
79-82	1.02		
80-81	5.52		
80-82	.70		
81-82	14.49		

The second general finding was that, for a number of item pairs, there was a tendency for the size of the positive association between items to increase steadily across decile groups. This can be seen in Table 4, where average estimates (across item pairs) are shown by unit and decile. Simple regression slopes of average estimates on decile level are shown to highlight these trends of increase. Thus, as the total score for items not included in the subset increased, there was an increased likelihood of a correct response to both items in a pair of the subset. The effect was shown by item pairs in both Mathematics/Science and Humanities/Social Science subsets, but it was more pronounced for some consecutive mathematics-based items (e.g., the first and second items of Unit 3), as shown in Table 2.

Thirdly, as shown in the column means in Table 4, the size of the positive associations between items in the Mathematics/Science units tended to be larger than that found between items in Humanities/Social Science units. Mathematics-based units again showed large positive associations between items.

These general characteristics of the model parameters can be illustrated by examining the odds ratios for responses to items 20 and 21 (from Unit 3), items 45 and 46 (from Unit 6), and items 91 and 95 and items 98 and 99 (from Unit 15). The odds ratios are presented in Table 5.

Items 20 and 21 and items 98 and 99 illustrate the case of positive interaction parameters, increasing across score groups. In both cases, the tables of odds ratios show that, as the total score on other items increases, the proportion of persons correctly answering the second item given a correct answer to the first increases more rapidly than does the proportion correctly answering the second item given an incorrect answer to the first. The effect is more pronounced for items 20 and 21, as also shown by the simple regression slopes.

Table 5 also shows that for items 91 and 95 and items 45 and 46, there is a negative interaction for most score groups. The proportion of students correctly responding to the second item is larger for the group giving an incorrect answer to the first item than it was for the group giving a correct answer to the first item.

Table 4
Average Estimates Across Item Pairs By Unit and Decile

Decile	Unit 3 Math 3 items	Unit 6i Sci 4 items	Unit 6ii Sci 4 items	Unit 12 Soc Sci 6 items	Unit 15i Hum 5 items	Unit 15ii Hum 5 items	Decile Average
1	.45	.39	.17	.13	.10	.06	.16
2	.70	.45	.05	.10	.09	.16	.18
3	.82	.45	.03	.08	.16	.14	.19
4	.97	.46	.05	.13	.16	.21	.23
5	.87	.47	.05	.17	.19	.25	.25
6	1.09	.50	.10	.22	.22	.25	.26
7	1.18	.53	.09	.12	.20	.27	.28
8	1.05	.56	.14	.13	.22	.37	.30
9	1.11	.53	.18	.12	.19	.39	.30
10	1.34	.67	.14	.16	.21	.39	.35
Average	.96	.50	.10	.13	.17	.25	.25
Slope*	.078	.023	.008	.004	.013	.036	.020
S.E.**	.012	.004	.006	.004	.003	.003	.001

*Simple regression of average estimates on decile level.
**Standard error of slope.

Table 5
Odds Ratios for Selected Item Pairs

Score	Item Pair			
Level	20-21	45-46	91-95	98-99
1	2.69	1.08	.99	1.51
2	3.61	.99	.91	1.51
3	4.83	.91	.85	1.43
4	5.78	.85	1.00	1.62
5	5.01	.81	.74	1.70
6	8.52	.89	.89	1.77
7	9.28	.90	.75	1.52
8	8.43	.93	.78	2.05
9	7.93	1.12	.64	1.72
10	17.82	1.11	.80	2.01
Simple*	1.203	.009	-.028	.054
S.E.**	.267	.012	.009	.016

*Simple regression of odds ratios on decile level.
**Standard error of slope.

Conditional Independence of Items Across Units in the Same Domain

Results of investigations of conditional independence across units are presented for the first items and last items in Mathematics, Science, and Humanities/Social Science. Each of these three areas was covered by 5 units in ASAT and, as in the preceding analysis, examinees were divided into 10 score levels according to their scores on remaining items in either Mathematics/Science or Humanities/Social Science. The goodness of fit of the independence model for each of the six 10 × 2^5 tables was evaluated and is summarized in Table 6. It can be seen from the table that the independence model fails to fit in every case, and that lack

Table 6
Summary of Model Fit for Cross-Unit Item Sets

	No. of Score Levels for Which Model Fits							
	Overall Goodness of Fit		Independence Model Alpha			All 2-Item Interaction Model Alpha		
Domain and Item Set	G	df	.05	.01	.001	.05	.01	.001
Math								
First	1473.0	260	1	1	2	10	10	10
Last	660.7	260	0	1	2	9	10	10
Sci								
First	348.9	260	7	8	9	10	10	10
Last	450.7	260	3	5	8	9	9	10
Hum/Soc Sci								
First	369.5	260	6	6	7	9	10	10
Last	306.6*	260	8	9	10	9	10	10

*p < .05, other fits p < .001.

Table 7
Test for Negative Item Association
Across Units of ASAT
(Results for First Items in a Pair of Units Shown
Above the Diagonal; Last Items Shown Below)

Unit and Item	Item				
Math	3	7	9	11	13
3	--	7.28	14.51	13.89	2.61
7	9.49	--	18.92	13.13	2.61
9	10.73	5.53	--	16.24	6.02
11	9.66	5.09	3.62	--	4.01
13	1.91	3.30	3.42	.01	--
Sci	1	4	6	10	14
1	--	2.05	4.74	3.90	2.65
4	10.10	--	3.41	4.49	.99
6	-.56	-.68	--	3.79	.21
10	3.38	3.92	-.02	--	4.01
14	4.64	2.50	.73	1.24	--
Hum/Soc	2	5	8	12	15
2	--	4.61	-.13	4.90	3.24
5	3.89	--	-.59	1.35	5.79
8	4.09	2.24	--	-.56	5.48
12	.51	.99	.74	--	2.38
15	1.72	-.55	2.22	1.33	--

of fit is greater for Mathematics units than for Science or Humanities/Social Science units.

The direction of item dependence was investigated further using the test discussed by Rosenbaum (1984). The results, presented in Table 7, show that no pair of items in the set exhibits an odds ratio of less than 1 across score levels. Illustrative odds ratios for a pair of items of each type are given in Table 8. These results demonstrate that the tendency for odds ratios to increase across score levels is confined to pairs of items belonging to the same unit, rather than pairs of items similarly based on material from a given domain, in that the slope coefficients of Table 8 for between-unit item pairs were not significant in contrast to those in the earlier tables for items within units.

Discussion

The consistent lack of independence of items within each score group suggests that the assumption of conditional independence may be questioned. In particular, for any item response model for which total score provides a good estimate of ability (e.g., the Rasch model), the results indicate that the conditional independence assumption is violated by these data. In part, the result may be attributed to the unit structure of the test, because lack of independence is most evident for items in the same unit and particularly for items in the same Mathematics unit.

The results also suggest, however, that item dependence occurs across unit boundaries and is to

Table 8
Odds Ratios for the First Items in Units 3 and 7 (Math), Units 1 and 4 (Sci), and Units 2 and 5 (Hum/Soc Sci)

	Item Pair		
Score Level	20-47 (Math)	1-23 (Sci)	6-29 (Hum/ Soc Sci)
1	1.15	1.03	1.17
2	1.14	1.04	1.27
3	1.36	1.02	1.20
4	1.45	1.06	1.11
5	1.28	1.29	1.19
6	1.26	1.11	1.03
7	1.43	.98	1.21
8	1.42	1.06	1.07
9	.97	.96	1.09
10	.80	1.25	1.11
Slope*	-.025	.007	-.014
S.E. of Slope	.023	.012	.007

*Slope of simple regression of odds ratio on decile level.

an extent dependent upon the content domain. Both within and across units, the dependence is more substantial for items based on mathematical material. In general, the associations between items were positive, indicating an increased probability of correctly answering one item given a correct response to the other. Some validity for the approach could be adduced, however, from the negative associations between item 46 and earlier items (where there appeared to be substantial guessing and hence where no systematic relationships might be anticipated). Similarly, the negative relationship between items 91 and 95 follows from a greater probability of correctly answering item 95 given failure on item 91, indicating perhaps a mental set from the earlier material in this unit working against the demands of the second question relating to subsequent material.

These results suggest that models such as the Rasch model are inappropriate for unit-structured tests like ASAT. At the level of investigating local independence for a more general class of item response models, however, the results are less discouraging. The dependence between items is generally positive, a condition which Rosenbaum (1984) has shown to be necessary but not sufficient for local independence (see also Holland, 1981).

An unexpected finding was the generally steady increase in association between some items as a function of score group level. This would seem to pose a problem for possible item response models for the data, and this trend would not be apparent in approaches such as those of Lord (1953) or Rosenbaum (1984), where statistics are aggregated across score levels to provide an overall test of independence.

It may be hypothesized that the clustered item structure of the test (rather than, say, "blind" guessing) is responsible for the finding, because its presence was confined to the analyses conducted within item clusters. This trend was characteristic of items relating to most kinds of stimulus material, but was most clearly shown for some mathematics-based units such as Unit 3. The stimulus material for mathematics-based units tends to constrain the possible questions which might be asked in a rigorous fashion; hence the probability of "complete understanding" increases across score groups, and consequently the probability of correctly responding to a number of items in the unit increases as well.

Another consideration might be that the possession of specific information derived from the study of mathematics might be more associated with the higher score levels than the lower ones. This possibility could not be investigated here, but it is worth noting that Yen (1984), in examining local independence in mathematics achievement tests, also found evidence of some general failures of local independence. Yen attributed this to specific knowledge as well as to isolated dependencies which may have reflected specific item relationships.

Conclusions

The results indicate that the items within units of ASAT were not independent in a way that is desirable from a test theory point of view. In particular, the assumption of independence between items at a given level of ability is violated for any model in which total score is a sufficient measure of ability. This is not surprising given the view of Goldstein (1980) about the strength of such an assumption and the previous results of Yen (1984), and in fact may be overcome by recourse to models which require weaker assumptions, as indicated by Rosenbaum (1984). However, the departures from local independence were not consistent throughout the test (i.e., they were more evident for mathematics-based units); similar problems may affect models based on homogeneity of items in ASAT.

The other conclusion drawn from this study was that this lack of independence often varied systematically as a function of the score level considered. No previous studies have considered this possibility (nor have they been in a position to observe it, due to the aggregating approaches used); its implications for model fitting for clustered-item tests such as ASAT are clearly in need of further attention.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Baker, R. J., & Nelder, J. A. (1978). *The GLIM system: Generalized linear interactive modeling.* Oxford: The Numerical Algorithms Group.

Bart, W. M., & Palvia, R. (1984). Relationships among test factor structure, test hierarchical structure, and test inter-item dependency structure. *Applied Psychological Measurement, 8,* 199–205.

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

Bishop, Y., Fienberg, S., & Holland, P. (1975). *Discrete multivariate analysis.* Cambridge MA: MIT Press.

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

Gustafsson, J.-E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33,* 205–233.

Hattie, J. A. (1985). Methodology review: Assessing the unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223–245.

Lazarsfeld, P. F. (1950). The interpretation and computation of some latent structures. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction (Studies in social psychology in World War II, Vol. 4).* Princeton NJ: Princeton University Press.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–549.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Mellenbergh, G. J., & Vijn, P. (1981). The Rasch model as a loglinear model. *Applied Psychological Measurement, 5,* 369–376.

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

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

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

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

Watson, G., & Glaser, E. M. (1980). *Critical Thinking Appraisal*. New York: Psychological Corporation.

Withers, G., & Bell, R. (1985). *Information in wrong answers: An optimal scaling of* ASAT. Paper presented at the ASAT Research Conference, Canberra, Australia.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23–48.

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.

Acknowledgments

The authors thank the Queensland Board of Secondary School Studies and the Western Australian Tertiary Institutes Service Centre for making these data available.

Author's Address

Send requests for reprints or further information to Richard C. Bell, Department of Psychology, University of Melbourne, Parkville, Victoria 3052, Australia.