Latent Trait Item Analysis and Facet Theory—A Useful Combination

John R. Balla and Roderick P. McDonald
Macquarie University

Computer programs for fitting latent trait models to data provide indices of item misfit. An analysis of the consistency of item misfit determination is presented. Two content-equivalent forms of 71 items representing the behavioral domain of arithmetic skills were generated. Each item was defined in terms of its combination of facet elements, and the ith item on each form represented the same selection of facets. The dichotomously scored responses to the two forms were analyzed using the computer programs NOHARM and BICAL. Misfitting items were identified by use of the residual covariances in the case of NOHARM and Total-t and Between-t in the case of BICAL. The consistency of misfit was measured by the extent of agreement in selection of misfitting items across the parallel forms. It was found that the analysis of residual covariances provided a more consistent means of determining item misfit. It was concluded that the use of the Between-t and Total-t indices as a basis for editing items should be viewed cautiously. In addition, misfitting items were grouped according to common facet elements and reasons for misfit were postulated. Thus, the analysis of residual covariances of items defined in terms of their combination of facet elements seems to provide a very satisfactory method of item analysis.

Computer programs have been developed to fit a variety of models based on item response theory (IRT; latent trait theory) to test data. Typically, values of item parameters are estimated by maxi-

mizing, in some sense, the fit of the model to data. In addition to estimates of parameters representing item difficulty, discrimination, and the lower asymptotes, some programs yield measures of item misfit (e.g., BICAL 3; Wright, Mead, & Bell, 1979).

The development of an item misfit index has been justified by the need to measure an item's suitability for measurement of the latent trait in question. Commonly, a function of the mean of squared residuals for the item is recommended. There has been considerable concern expressed in the literature that the use of presently recommended measures of item misfit may lead to unwarranted exclusion of items. Wood (1978) asserted that by aiming for homogeneous groups of items the measurement goal may be sacrificed, whereas Rentz (1978), recognizing some problems with the misfit statistics, recommended not deleting items purely on the basis of one calibration run. Realizing the problems with these misfitting items, Gustaffson (1980) suggested a strategy involving the development of homogeneous subsets of items which independently fit the model. Quite correctly Gustaffson suggested that any proposed scale should be cross-validated on another sample of persons with further items. An interesting question then is: "How well do the measures of item misfit identify misfitting subsets of items?" That issue was explored in this investigation.

A number of researchers (e.g., Gustaffson & Lindblad, 1978; Hambleton, Swaminathan, Al-

gina, & Coulson, 1978) have advocated the use of content analysis to supplement the measures of item misfit used in item analysis. Such content analysis requires a knowledge of the subject matter being tested. It can be reasonably claimed that the process of content analysis will be best served if the items included in a test are well-defined in terms of the facets of the behavioral domain they represent. The technology of defining a behavioral domain in terms of a number of facets is well developed (Foa, 1965, 1968; Guttman, 1954). In addition, a number of workers have outlined procedures of item specification by item forms (Hively, 1974; Hively, Patterson, & Page, 1968; Osburn cited in Martuza, 1977). However, it can be claimed that their work is more neatly conceptualized within the facet theory developed by Guttman and Foa, where a behavioral domain is expressed in terms of a cartesian product of the facets taken to define the domain. Items can then be generated that represent all aspects of the domain, and indeed some types can be generated by computer (Balla, 1982). The analysis of the behavioral domain in terms of facets has been applied in the present study.

The normal ogive and logistic IRT models have been shown to have virtually indistinguishable item characteristic curves (see e.g., Lord & Novick, 1968). McDonald (1967) demonstrated, in effect, that the normal ogive model can be fitted by expressing the model, and hence the item covariances, in terms of Hermite-Tchebycheff polynomials. Christoffersson (1975) fitted the normal ogive model by expressing the item covariances in terms of tetrachoric functions. More recently, McDonald (1985) has shown that these two treatments are essentially mathematically equivalent. The computer programs for these treatments differ mainly in that FADIV (Andersson, Christoffersson, & Muthén, 1974) uses estimation by generalized least squares and a ten-term series approximation to the tetrachoric functions, whereas NOHARM (Fraser, 1984), written for McDonald's work, uses ordinary least squares and a three-term approximation to the tetrachoric functions. Consequently, NOHARM can manage much larger data sets than FADIV and thus is suitable for the present investigation, which requires analyses of up to 71 items. NOHARM prints a matrix of residual interitem covariances. Mc-

Donald (1982) has suggested that the model can be said to fit only if these residual covariances are small, since by the weak form of the principle of local independence the model must explain the item covariances.

Thus, the three objectives of the investigation reported here were:
1. To attempt to identify misfitting items (a) by an analysis of the residual covariance matrix after fitting the Rasch model using NOHARM, and (b) by the use of BICAL 3 output relating to item misfit, namely Total-t and Between-t as defined by Wright et al. (1979).
2. To determine whether items so identified possess any common properties in terms of the facets of the behavioral domain they represent, and
3. To observe whether the selection of items misfitting can be cross-validated in the analysis of a parallel form given to matched subjects.

Method

Two content-equivalent forms of 71 items drawn from the behavioral domain of arithmetic skills were generated. The forms were content-equivalent in the sense that the ith item on each form represented the same selection of facets. The two forms, divided into subgroupings according to their combinations of facet elements, are presented in Table 1.

The two forms were constructed to be appropriate for late Grade 3 and early Grade 4 (8- to 9-year-olds), and were thus administered to third and fourth graders in six Sydney metropolitan primary schools: Form A to 279 students, Form B to 309 students. The student responses were coded as 1 or 0 (correct or incorrect, respectively), and the results from the two forms were then separately analyzed by NOHARM and BICAL.

As described previously, NOHARM expresses misfit in terms of a residual covariance matrix from which an overall mean square index of fit of the Rasch model to data can be obtained. Further, the availability of the entire residual covariance matrix provides an opportunity to investigate patterns of contributions of misfit by each pairing of items. The standard error for a residual can be crudely

Table 1
Items in Forms A and B of the Arithmetic Skills Test

Form A			Form B		
Addition and Subtraction (24 items)			**Addition and Subtraction (24 items)**		
(1) $3+4=\_$	(7) $5+\_=14$	(11) $\_+7=88$	(1) $6+3=\_$	(7) $6+\_=13$	(11) $\_+3=49$
(2) $2+\_=7$	(8) $\_+7=16$	(12) $56+5=\_$	(2) $4+\_=8$	(8) $\_+9=17$	(12) $39+8=\_$
(3) $\_+1=9$	(9) $23+5=\_$	(13) $84+\_=93$	(3) $\_+2=9$	(9) $12+7=\_$	(13) $74+\_=81$
(6) $3+9=\_$	(10) $43+\_=49$	(14) $\_+4=42$	(6) $4+8=\_$	(10) $36+\_=38$	(14) $\_+5=62$
(19) $7-3=\_$	(26) $14-\_=8$	(30) $\_-8=21$	(19) $9-2=\_$	(26) $15-\_=6$	(30) $\_-6=42$
(20) $6-\_=2$	(27) $\_-4=8$	(31) $43-8=\_$	(20) $8-\_=5$	(27) $\_-8=4$	(31) $54-7=\_$
(21) $\_-5=4$	(28) $63-2=\_$	(32) $64-\_=55$	(21) $\_-4=2$	(28) $57-4=\_$	(32) $73-\_=66$
(25) $15-7=\_$	(29) $89-\_=84$	(33) $\_-4=27$	(25) $17-8=\_$	(29) $76-\_=73$	(33) $\_-3=38$
Multiplication and Division (15 items)			**Multiplication and Division (15 items)**		
(45) $4\times5=\_$	(48) $7\times4=\_$	(51) $7\times6=\_$	(45) $5\times3=\_$	(48) $8\times4=\_$	(51) $8\times9=\_$
(46) $3\times\_=9$	(49) $3\times\_=30$	(52) $8\times\_=64$	(46) $5\times\_=20$	(49) $4\times\_=36$	(52) $6\times\_=48$
(47) $\_\times2=10$	(50) $\_\times3=21$	(53) $\_\times9=54$	(47) $\_\times3=12$	(50) $\_\times2=18$	(53) $\_\times7=49$
(61) $4\div2=\_$	(64) $32\div4=\_$		(61) $8\div4=\_$	(64) $27\div3=\_$	
(62) $12\div\_=4$	(65) $40\div\_=5$		(62) $16\div\_=4$	(65) $36\div\_=9$	
(63) $\_\div2=4$	(66) $\_\div2=8$		(63) $\_\div5=3$	(66) $\_\div7=2$	
Properties of Identities (13 items)			**Properties of Identities (13 items)**		
(4) $6+0=\_$	(24) $8-8=\_$	(43) $0\times4=\_$	(4) $0+5=\_$	(24) $3-3=\_$	(43) $0\times7=\_$
(5) $0+\_=3$	(40) $1\times5=\_$	(44) $5\times0=\_$	(5) $\_+0=4$	(40) $5\times1=\_$	(44) $6\times0=\_$
(22) $6-0=\_$	(41) $1\times\_=8$	(58) $6\div1=\_$	(22) $4-0=\_$	(41) $\_\times1=8$	(58) $5\div1=\_$
(23) $\_-0=5$	(42) $1\times1=\_$	(59) $\_\div1=5$	(23) $\_-0=6$	(42) $1\times1=\_$	(59) $\_\div1=8$
		(60) $0\div7=\_$			(60) $0\div4=\_$

Vertical Format (14 items)   **Vertical Format (14 items)**

Form A				Form B			
(15) $62+$ $\underline{25}$	(16) $45+$ $\underline{38}$	(17) $74+$ $\underline{63}$	(18) $56+$ $\underline{88}$	(15) $27+$ $\underline{32}$	(16) $46+$ $\underline{27}$	(17) $82+$ $\underline{36}$	(18) $45+$ $\underline{67}$
(34) $47-$ $\underline{23}$	(35) $58-$ $\underline{29}$	(36) $74-$ $\underline{66}$	(37) $147-$ $\underline{79}$	(34) $68-$ $\underline{25}$	(35) $67-$ $\underline{29}$	(36) $83-$ $\underline{78}$	(37) $187-$ $\underline{99}$
(38) $160-$ $\underline{77}$	(39) $106-$ $\underline{68}$			(38) $180-$ $\underline{56}$	(39) $104-$ $\underline{75}$		
(54) $23\times$ $\underline{3}$	(55) $42\times$ $\underline{4}$	(56) $18\times$ $\underline{5}$	(57) $63\times$ $\underline{8}$	(54) $21\times$ $\underline{4}$	(55) $53\times$ $\underline{3}$	(56) $24\times$ $\underline{4}$	(57) $54\times$ $\underline{6}$

Fractions/Brackets (5 items)   **Fractions/Brackets (5 items)**

Form A			Form B		
(67) $\frac{1}{2}\times14=\_$	(68) $\frac{1}{3}\times\_=10$	(69) $\frac{3}{4}\times40=\_$	(67) $\frac{1}{2}$ of $14=\_$	(68) $\frac{1}{3}$ of $\_=10$	(69) $\frac{3}{4}$ of $40=\_$
(70) $4\times(3+5)=\_$	(71) $(10-8)\div2=\_$		(70) $3\times(4+3)=\_$	(71) $(8-2)\div3=\_$	

approximated by $1/n^{1/2}$, where n is the sample size. Sample sizes in the present study resulted in approximate standard errors of .059 for Form A and .057 for Form B. This suggests, as a not unreasonable decision rule, that two or more items may be considered a misfitting group if their pairwise residual covariances are all greater than .05 in the samples (see Fraser, 1984).

Results

The misfitting items determined by analysis of the residual covariance matrices generated by NOHARM for Forms A and B can be found in Table 2. Altogether, there were 84 residual covariances larger than the chosen criterion. On Form A there were 74, while 73 were identified on Form B. Of these, 63 were identified on both forms. This constituted 75% agreement across the forms. In terms of items, there were 40 belonging to misfitting groups, 36 on Form A and 29 on Form B. Of these, 25 items were on both. This constituted 63% agreement.

A facet analysis of the misfitting groups commonly identified across the two forms was undertaken. This was an attempt to determine whether the items in these misfitting groups possessed any similarity in their facet composition, thereby suggesting reasons for their misfit. The results of this facet analysis are as follows.

Items 30 & 33: These were two of the four subtraction items expressed in horizontal format $(a - b = c)$ with the unknown in position a (see Table 1). Interestingly, the other two items with this combination of facets were also misfitting on Form B. Items 30 and 33 differed from 21 and 27 in that the replacement set for a was from 20 to 99, whereas it was from 2 to 10 for item 21 and from 11 to 20 for item 27.

Items 35 & 36: These were two of the six subtraction items expressed in vertical format (see Table 1). They were 2-digit minus 2-digit items requiring borrowing from the "tens" column to solve. Note that item 34, involving no borrowing, was not identified as misfitting on either form, but items 37 to 39, which were 3-digit minus 2-digit items that require borrowing, were identified as a misfitting group on Form A.

Items 43 & 44: These two items required knowledge of $0 \times b$ and $a \times 0$, respectively.

Items 52 & 53: These were two of the three multiplication items in horizontal format with both multipliers from 6 to 10 and unknown in position a or b.

Items 54 to 56: These were three of the four multiplication items in vertical format. Item 57, the fourth item, was included in the misfitting group on Form B. Item 57 involved a multiplier between 5 and 10 whereas the other three items involved multipliers between 2 and 5.

Table 2
Misfitting Items from Forms A and B Based on NOHARM Analysis

Test	Item Number	
	Form A	Form B
Addition	11, 14, 16, 18	
Subtraction	30, 33, 35, 36, 37–39	21, 27, 30, 33, 35–37
Multiplication	43, 44, 45–47, 49, 50 51, 52, 53, 54–56	43, 44, 48, 49, 52, 53, 54–57
Division, Fractions, & Brackets	58–67, 70, 71	58–67, 70, 71

Items 58 to 67, and 71: Items 58 to 66 comprised all the division items. Item 67 was the first fraction item, which could be considered as a division by two, and item 71 was the last brackets item, which included a division operation.

Items 70 & 71: These were the two brackets items.

Wright et al. (1979), in describing the fit statistics provided by BICAL 3, claimed that *Total-t* summarizes overall item fit from person to person and *Between-t* focuses on the variations between the ability groups formed by BICAL 3. Both of these statistics are hypothesized as being standardized values of weighted mean squares. However, based on their experience, Wright et al. suggested 1.5 as an upper cutoff and -2 as a lower cutoff for Total-t and 2 and -2 for Between-t in order to approximate the 95% confidence limits for a two-tailed t-distribution with the 200 to 300 degrees of freedom applicable in the present analysis.

Table 3 provides Total-t and Between-t item fit statistics for items identified as misfitting by BICAL 3 (Ucon option). Of the 26 items identified as misfitting when using the Total-t index, 12 were commonly identified on the two forms (46% agreement). Of the 24 items identified as misfitting by the Between-t index, 7 were commonly identified as misfitting (29% agreement).

A facet analysis of the misfitting items identified in both forms gave the following results:

Item 4 was $6 + 0 = $ ___ .

Items 35 & 36 were as discussed above in the NOHARM analysis.

Item 19 was the first subtraction item.

Item 25 was a question requiring subtraction from numbers greater than 9.

Item 32 was $64 - $ ___ $ = 55$, an item requiring borrowing.

Item 42 was $1 \times 1 = $ ___ .

Item 56 was a multiplication item in vertical format requiring carrying to the second column.

Items 59 to 63, 65 & 66: These were all of the division items, except for item 58, in which the unknown was in position c and division was by 1, and item 64, in which the unknown was in position c. These two items were misfitting on the Form B analysis.

Item 71 was a brackets item involving division operation.

The most obvious agreement between the two types of misfit analysis involves the identification of the division items and of item 71 as being misfitting (though it is noted that the BICAL analysis of Form A did not include items 58 and 64 as misfitting). Items 35, 36, and 56 were also identified as misfitting by both methods.

Discussion

The analysis of the residual covariance matrix generated by NOHARM provided a simple method of identifying misfitting groups of items after fitting the Rasch model. The identified pairings of items in most cases exhibited meaningful relationships in terms of their common facet elements and hence provided content validity for being so classified.

These results compare favorably with the findings of Gustaffson and Lindblad (1978), who were able to identify subsets of grammar items that were jointly misfitting by a test of slopes of the item characteristic curves (Gustaffson, 1980). This suggests that the analysis of the residual covariance matrix is a reasonable alternative means of identifying misfitting subgroups. In fact, the pattern of misfit obtained in the NOHARM analysis (and in the BICAL analysis) suggests that the division items should be separated from the other three operations at late third grade and early fourth grade if fitting of the Rasch model to data is to be achieved. Further, the misfit of the following groups of items suggests that they may not be suitable in a test for this level of subject given the assumption of one latent trait and estimation using one parameter:

1. 2-digit minus 2-digit problems expressed in vertical format requiring borrowing;
2. Multiplication expressed in vertical format;
3. Multiplication by zero;
4. Subtraction in horizontal format with unknown in position a; and
5. Multiplication involving multipliers from 6 to 10.

In fact, it becomes clear that responses from late third grade and early fourth grade students to ad-

Table 3
Total-t and Between-t Item Fit Statistics
for Items Identified as Misfitting Using BICAL 3

	Total-t				Between-t			
	Form A		Form B		Form A		Form B	
Test	Item	t	Item	t	Item	t	Item	t
Addition								
			4	1.81	*4	4.48	*4	8.83
					9	2.62	15	4.61
					6	2.62	18	2.95
					17	2.14	7	2.05
Subtraction								
	*35	3.48	31	4.90	*25	4.33	31	4.24
	34	2.72	30	3.36	34	3.43	*25	3.53
	*36	2.67	*35	2.23	26	2.62	30	3.18
	21	1.77	*36	2.03	*35	2.47	*19	2.56
	*32	1.65	*32	1.55	*19	2.45	*35	2.54
Multiplication								
	45	-2.03	53	-2.14	*42	3.94	*42	3.05
	46	-2.12	55	-2.24	43	2.76		
	*56	-2.35	*56	-2.66	46	2.61		
	49	-2.54			45	2.37		
	50	-2.62						
Division								
	*65	-2.36	*60	-2.02	*60	2.03	62	3.27
	*61	-2.45	58	-2.77	*66	2.00	*60	3.15
	*59	-2.57	64	-2.90			65	2.85
	*60	-3.08	*59	-3.52			*66	2.62
	*63	-3.19	*63	-3.56			59	2.45
	*62	-3.27	*65	-3.89				
	*66	-3.46	*61	-4.32				
			*66	-4.45				
			*62	-4.95				
Fractions								
							69	2.26
Brackets								
	70	-2.00	*71	-2.89				
	*71	-2.23						

*Indicates that an item was commonly identified across the
forms by the fit statistic.

dition items, subtraction items without the need for borrowing, and multiplication items involving multipliers from 1 to 5 fit the Rasch model.

The analysis of item residual covariances gave good agreement (75%) between the pairs of items identified as misfitting across the parallel forms. This finding would provide confidence that the items fitting the model will continue to do so, and further,

that there are genuine reasons for such misfit as it occurs (i.e., inappropriateness of items, differential difficulty, lack of instruction), and that the items are genuinely multidimensional as the facet analysis implies.

The measures of item misfit (Between-t and Total-t) provided by BICAL are also very simple to use to identify misfitting items. A number of the

items identified across both forms could be grouped meaningfully as measured by the facet elements they possessed. The level of agreement of misfitting items across forms (46% and 29%, respectively), was, however, quite low for the two measures of misfit (Total-t and Between-t) employed in BICAL. This finding leads to questioning the use of these measures in a single analysis. If the measures of item fit used in BICAL are so unstable when using parallel forms and matched subjects, then caution is needed in using the BICAL indices as a basis for editing items to form a so-called "more comfortable set" (Wright & Panchapakesan, 1969). This is of even more concern when it is realized that the linking procedure used in setting up a scaled item bank (Wright, 1977) often requires the scaling of different groupings of items with unmatched subjects. In addition, any analysis of the characteristics of a misfitting item (see Wright & Stone, 1979) when using such unstable measures of misfit may be questioned.

The results of this study indicated that the use of a cutoff level for the residual covariances resulted in a relatively stable identification of misfitting items when applied to parallel forms containing items defined to map the behavioral domain of arithmetic skills appropriate to Grade 3-4 students compared to that observed with the BICAL indices of misfit. Further, the misfit of these items is easily understood because of their precise definition in terms of the facets of the behavioral domain. Logical clusters of misfitting items are evident, yielding significant educational implications concerning the reason for misfit and leaving logical subsets of items that fit the model. The use of facet analysis provides a helpful complement to item analysis. This procedure could easily provide a basis for comparative studies of fit with models of more than one parameter or studies involving multidimensional items.

References

Andersson, C. G., Christoffersson, A., & Muthén, B. (1974). *FADIV. A computer program for analysis of dichotomized variables* (Research Report 74-1). Uppsala, Sweden: Uppsala University, Statistics Department.

Balla, J. R. (1982). Developments in the areas of item banking and latent trait measurement as related to the solution of problems associated with common testing. In J. Veness (Ed.), *Mathematics—A universal language* (pp. 201–202). Sydney: The Australian Association of Mathematics Teachers.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

Foa, U. G. (1965). New developments in facet design and analysis. *Psychological Review, 72,* 262–274.

Foa, U. G. (1968). Three kinds of behavioral changes. *Psychological Bulletin, 70,* 460–473.

Fraser, C. (1984). *NOHARM II—A Fortran program for fitting unidimensional and multidimensional models of latent trait theory.* Tamworth, Australia: University of New England, Center for Behavioral Studies in Education.

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

Gustaffson, J. E., & Lindblad, T. (1978). *The Rasch model for dichotomous items: A solution of the conditional estimation problem for long tests and some thoughts about screening procedures* (Report No. 67). Göteborg, Sweden: University of Göteborg, Institute of Education.

Guttman, L. (1954). An outline of some new methodology for social research. *Public Opinion Quarterly, 18,* 395–404.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1–47.

Hively, W. (1974). Introduction to domain referenced testing. *Educational Technology, 14,* 5–10.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A universe defined system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275–290.

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

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

McDonald, R. P. (1982). Fitting latent trait models. In D. Spearitt (Ed.), *The improvement of measurement in education and psychology* (pp. 213–237). *Proceedings of the 1980 Invitational Seminar for the Fiftieth Anniversary of the Australian Council for Educational Research.* Melbourne: Australian Council for Educational Research.

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 Computerized Adaptive Testing Conference* (pp. 127–148). Minneapolis: University of Min-

nesota, Department of Psychology, Psychometrics Program.

Martuza, V. R. (1977). *Applying norm-referenced and criterion-referenced measurement in education.* Boston: Allyn and Bacon Inc.

Rentz, R. R. (1978, March). *Monitoring the quality of an item-pool calibrated by the Rasch model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Wood, R. (1978). Fitting the Rasch model—A heady tale. *British Journal of Mathematical and Statistical Psychology, 31,* 27–32.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Wright, B. D., Mead, R. J., & Bell, S. R. (1979). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23B). Chicago: Chicago University, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29,* 24–48.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Author's Address

Send requests for reprints and further information to Roderick P. McDonald, School of Education, Macquarie University, North Ryde, N.S.W. 2113, Australia.