Detecting and evaluating the impact of multidimensionality in language test data[1]

R. J. Adams
Australian Council for Educationai Research
T. F. McNamara
University of Melbourne
and
S. Zammit
Australian Council for Educational Research[2]

Abstract

The paper uses new multidimensional Rasch models to explore the dimensionality associated with the stimulus passages in two foreign language listening comprehension tests designed for beginning and intermediate level students of foreign languages in Australian secondary schools. The data were analysed twice, once fitting a model assuming independence of items (a 'one-factor' or unidimensional analysis) and once fitting a model which took into account possible dependency of items on the passages with which they were associated (an 'actual factor' or multidimensional analysis). While the results show clear evidence of passage-related multidimensionality in the data, the consequences of ignoring this dimensionality in the reporting of test results is found to be modest when compared to the error introduced by even limited amounts of unreliability in the test. It is concluded that while exploration of multidimensionality in language test data is worthwhile and an interesting avenue for research, it should not take precedence over efforts to improve test reliability.

1. Introduction

The question of dimensionality in language test data has been debated for over a decade. On the one hand, applied linguists

---

[1]This paper was presented at the Language Testing Research Colloquium, Monterey, CA, March 9-12, 1998.

[2]*Address for correspondence*: Department of Linguistics and Applied Linguistics, The University of Melbourne, Parkville, Victoria 3052, AUSTRALIA *e-mail*: T.McNamara@linguistics.unimelb.edu.au

Melbourne Papers in Language Testing 1998 Volume 7.2 pp. 1-14. The Language Testing Research Centre, The University of Melbourne.

mindful of the complex factors underlying language test performance (Buck 1992, 1994), or the specific skills required in special purpose language tasks (Skehan 1984, 1986), have found the simplifying unidimensionality assumption of standard psychometric procedures problematic. Others have explained or defended the procedures (McNamara 1991, 1996; Henning 1992), insisting on the distinction between psychological constructs and measurement dimensions, the latter involving a deliberate simplification of reality. This sort of simplification is familiar in all modelling - street maps, for example, are not three dimensional, even in a city as hilly as San Francisco, in defiance of the reality of streetscapes and the experience of runaway cars or of walkers and runners, particularly those short of breath or otherwise out of condition. Unidimensionality, so the argument goes, is a matter of convenience (or even of principle as the advocates of Rasch modelling would argue) and in any case is a property of the data matrix rather than of the evidently complex psychological realities underlying performance; and tests for unidimensionality (that is, that a single pattern *is* detectable in the scores) exist, even if they are too seldom applied (de Jong and Stoyanova 1994). The debates have sometimes been triggered by the use of Rasch measurement in language testing research (e.g. Nunan 1988, 1989; Hamp-Lyons 1989), although the undimensionality assumption is a less frequently recognized feature of classical true score analyses also. One persistent context in which dimensionality is likely to be an issue is the use of passages in comprehension tests, with sets of items (item bundles or testlets) written around each passage; it has long been suspected that passage dependency effects would be a likely source of multidimensionality in the data, performance on each passage forming a dimension.

Recent developments in psychometric models and advances in computational power have reopened the debate. Multidimensional models, including multidimensional Rasch models, now abound, and their implementation in computer programs is now routine. Should we not be using them to explore and, if necessary, to model and control for multidimensionality in language test data? Recent papers (McNamara and Adams 1997; in press) have explored the issue of multidimensionality in performance data in language tests (speaking and writing) and have tentatively concluded that the impact on ability estimates of modelling or not modelling multidimensionality in the data are modest at best. However, these

studies used small data sets and called for further research using simpler designs and larger data sets. The present paper is offered as part of that research effort. Here, language data sets with candidature in excess of 2000 on multiple choice tests of 25 items are used to explore further the issue of multidimensionality associated with passage dependency.

2. Data

The data were drawn from performance on the listening comprehension sections of tests for the National Australia Bank Language Certificates. The National Australia Bank Language Certificates program is designed to encourage students learning a foreign language at school. The program, which started in 1990, has continued to grow, with over 680,000 participants in 1200 schools registered in 1997.

Under the program, students complete listening and reading tasks at school, and their answer sheets are scored centrally at the Australian Council for Educational Research. Feedback takes the form of detailed school reports and student certificates describing the levels achieved by individual participants. Students benefit from receiving positive recognition of their language learning efforts. Each participating student receives a certificate describing the level or 'band' of their personal achievement.

Certificates for secondary school students are available at Beginners and Intermediate levels, and in six languages, Mandarin Chinese, French, German, Indonesian, Italian and Japanese.

Chinese (Beginners)
French (Beginners and Intermediate)
German (Beginners and Intermediate)
Indonesian (Beginners)
Italian (Beginners and Intermediate)
Japanese (Beginners and Intermediate)

Secondary school students who have studied the target language for between 80 and 200 hours are eligible to take part in the Beginners level. These students are most likely to be in their second year of language study. For the Intermediate level, students should have

received between 200 and 300 hours of language instruction, and are likely to be in their third or fourth year of language study.

Two listening tests designed for the 1997 administration were chosen as the source of data for this study: Indonesian Beginners level and French Intermediate level. Tables 1 and 2 below show the topic or setting of each of the stimulus passages on the two tests and the number of multiple choice comprehension questions (item bundles) associated with each passage.

Passage No.	Topic/Setting	Dialogue/ Monologue	No. of items
1	New student, Classroom	Dialogue	6
2	Shopping, Market	Dialogue	6
3	Chatting with neighbour	Dialogue	4
4	Making a reservation, airline office	Dialogue	3
5	Checking in, airport	Dialogue	6

Table 1: Indonesian (Beginners' Level), N=3819; 5 stimulus passages

Passage No.	Topic/Setting	Dialogue/ Monologue	No. of items
1	Radio interview with Youth Hostel manager	Dialogue	7
2	Horoscopes	Monologue	8
3	Radio advertisements	Monologue	5
4	Weather report	Monologue	3

Table 2: French (Intermediate level), N=4789; 4 stimulus passages

3. Research questions

The study attempts to answer the following questions:

1. How good are Rasch methods at detecting multidimensionality?

2. Can multidimensionality associated with stimulus passages be detected in this data?

3. What difference does modelling or not modelling such dimensionality have on scores and score reporting?

4. Method

In order to compare the effect on person measures of modelling or failing to model the multidimensionality of item bundles associated with particular stimulus passages, the following steps were taken for each language.

Step 1
The data were analysed twice, once fitting a model assuming independence of items (a 'one-factor' or unidimensional analysis) and once fitting a model which took into account possible dependency of items on the passages with which they were associated (an 'actual factor' or multidimensional analysis). All analyses were conducted using the *ConQuest* computer program (Wu, Adams and Wilson 1998). The analyses yielded:

1.1 estimates of the *reliability* of the tests;

1.2 *item fit statistics;*

1.3 a summary measure of the overall misfit in each analysis, the *deviance;*

1.4 from the multidimensional analysis, a *covariance-correlation matrix,* providing:

- *correlations* between each of the dimensions, corrected for error;
- estimates of the *variance* on each dimension, corrected for error.

Step 2
The fit of the two models was investigated:

2.1 *item fit statistics* for individual items from the one-factor analysis were examined to explore patterns of misfit in items associated with particular passages;

2.2 the *overall misfit on deviance* from the two analyses was compared using a chi-square test.

The question being asked here was, 'Do the data overall show evidence of dimensionality associated with passage dependency of items? That is, does an analysis allowing for passage dependency of items produce a significantly better fit to the data?'

Step 3
Evidence for the size and nature of dimensionality associated with stimulus passages effects was explored in the covariance-correlation matrix from the 'actual factor' analysis:

3.1 correlations between the dimensions were explored;

3.2 the variance on particular dimensions was considered.

The question being asked here was, 'Can we see patterns of dependency around particular passages? Or are our estimates of ability derived from performance on items associated with each of the passages essentially stable across sets of items?'

Step 4
Evidence of the impact of the dimensions discovered on overall ability estimates was sought, using two measures:

4.1 correlations between ability estimates derived from the unidimensional analysis and ability estimates on each of the dimensions;

4.2 the proportion of individuals classified into different reporting bands under the two analyses.

5. Results

5.1. Individual item misfit (Step 2.1)

Three items showed significant misfit in the Indonesian test, two in the French test. For neither test was any passage-related pattern of fit observed in the unidimensional analysis; there was a scatter of fit across the passages.

5.2. Comparison of overall model-data fit (Step 2.2)

A chi-square test was applied to compare the deviances (overall fit) yielded by the two analyses (Tables 3 and 4), and evaluated in the light of its degrees of freedom.[3] The results show evidence of multidimensionality associated with passage effects; the multidimensional analysis shows significantly better fit to the data in the case of each language: Indonesian, chi-square = 239.2, df = 14, p <.01; French, chi-square = 82.3, df = 9, p <.01.

Analysis	Deviance	chi-square	df	p
One-factor	101322.9	239.2	14	<.01
Actual factor	101083.7			

Table 3: Comparison of overall data-model fit, unidimensional ('one-factor') and multidimensional ('actual factor') analyses - Indonesian data

Analysis	Deviance	chi-square	df	p
One-factor	118096.7	82.3	9	<.01
Actual factor	118014.4			

Table 4: Comparison of overall data-model fit, unidimensional ('one-factor') and multidimensional ('actual factor') analyses - French data

5.3. Exploration of dimensions (Step 3)

The covariance-correlation matrix from each of the analyses is reproduced in Tables 5 and 6. Values below the diagonal (in bold)

[3]For the Indonesian data, with 5 stimulus passages, the unidimensional analysis estimated 26 parameters (25 item parameters and one parameter for the latent variable); the multidimensional analysis yielded 40 parameters (25 item parameters and the 15 parameters associated with the covariance-correlation matrix - 5 variances and 10 correlations); the degrees of freedom were thus 40-26=14. For the French data, with 4 stimulus passages, the unidimensional analysis estimated 24 parameters (23 item parameters and one parameter for the latent variable); the multidimensional analysis yielded 33 parameters (23 item parameters and the 10 parameters associated with the covariance-correlation matrix - 4 variances and 6 correlations); the degrees of freedom were thus 33-24=9.

are correlations (the values above are covariances). The dimensions are the particular stimulus passages and associated items.

	Dimension				
Dimension	1	2	3	4	5
1		1.113	0.903	0.755	0.840
2	**0.857**		1.131	0.902	1.005
3	**0.763**	**0.810**		0.746	0.946
4	**0.789**	**0.800**	**0.725**		0.705
5	**0.806**	**0.818**	**0.845**	**0.779**	
Variance	1.102	1.532	1.272	0.830	0.986

Table 5: Covariance/correlation matrix, Indonesian data

	Dimension			
Dimension	1	2	3	4
1		0.788	0.834	0.627
2	**0.893**		0.877	0.644
3	**0.853**	**0.850**		0.756
4	**0.712**	**0.694**	**0.735**	
Variance	0.837	0.932	1.144	0.926

Table 6: Covariance/correlation matrix, French data

The correlations reported in the lower half of the matrix (in bold) are correlations of the latent variables in the analysis, that is, they have already been corrected for measurement error. The default expectation (the null hypothesis) is that the correlations will be 1.0 if the test data are unidimensional; that is, that no matter on which passage the ability estimate is made, the estimates will be identical. In fact the range of correlations between the dimensions for the Indonesian test is from 0.725 to 0.857; for French, the range is 0.694 to 0.893. The variances are also not constant; we would expect them to be similar across dimensions (stimulus passages), but in fact they show variability. For example, in the Indonesian test, passages 2 and 5 (each with the same number of items) show

variances of 1.532 (Passage 2) and 0.986 (Passage 5). Thus both in terms of the correlations of the latent variables, and in terms of equality of variances, the analysis shows evidence of multidimensionality in the data.

5.4. Impact on scores (Step 4)

If we were to model this multidimensionality, what impact would this have on the reported abilities?

Two methods were used to answer this question.

First, correlations between ability estimates based on the unidimensional and multidimensional analyses were calculated. Maximum likelihood estimates were used and the correlations are reported in Tables 7 and 8.

	Estimates based on passage-related dimensions				
	1	2	3	4	5
Estimates from 'one-factor' analysis	0.82	0.84	0.80	0.78	0.83

Table 7: Correlations (*r*) between ability estimates under unidimensional and multidimensional analyses - Indonesian test

	Estimates based on passage-related dimensions			
	1	2	3	4
Estimates from 'one-factor' analysis	0.82	0.83	0.79	0.69

Table 8: Correlations (*r*) between ability estimates under unidimensional and multidimensional analyses - French test

Although the correlations appear relatively high, it must be remembered that the estimates going into the calculation have already been corrected for measurement error. In other words,

ignoring the passage dependency effects does seem to have an impact on ability measures.

In fact, precise ability estimates are not reported in this assessment scheme; rather, candidates are allocated to bands of ability. In order to consider the question of the impact on the allocation to bands of ability, a type of classification analysis was carried out.

The cohort was classified into quintiles, and the classification on the basis of the unidimensional analysis was compared with the classification (one dimension at a time) resulting from the multidimensional analysis. An example is provided in Table 9, showing a comparison of the quintiles to which individuals would be classified using a unidimensional analysis and those to which they would be allocated according to performance on the first dimension (first stimulus passage).

			Quintile classification on dimension 1				
			1	2	3	4	5
Quintile classification on 'one-factor' analysis	1	%	**12.46**	4.50	1.96	0.86	0.13
		n	476	172	75	33	5
	2	%	4.77	**6.60**	5.47	2.23	0.60
		n	182	252	209	85	23
	3	%	1.99	5.21	**5.76**	5.13	2.30
		n	76	199	220	196	88
	4	%	0.63	2.64	5.00	**6.76**	4.87
		n	24	101	191	258	186
	5	%	0.10	0.68	2.04	5.05	**12.23**
		n	4	26	78	193	467

Table 9: Classification analysis, unidimensional and multidimensional analysis (Dimension 1) - Indonesian test

It can be seen from Table 9 that there is a good deal of misclassification. The figures in bold are frequencies of correct classification at each level; misclassification by one level is almost as common. Table 10 shows the frequencies of misclassification for Table 9 by extent of misclassification.

How classified in two analyses	%
Same quintile	43.81
1 quintile apart	40.00
2 quintiles apart	13.16
3 quintiles apart	2.77
4 quintiles apart	0.23

Table 10: Comparison of classification in uni- and multi-dimensional analysis - Dimension 1, Indonesian data

Table 11 provides the same summary information about classification and misclassification using results from the other dimensions in the Indonesian data. It shows a similar pattern of misclassification.

No of levels apart in 2 analyses	Dimension 2	Dimension 3	Dimension 4	Dimension 5
0	45.04	42.65	40.11	44.89
1	40.41	40.05	39.07	38.85
2	12.52	14.34	16.23	13.15
3	2.02	2.70	4.27	2.88
4	0.03	0.23	0.32	0.23

Table 11: Classification analysis, uni- and multi-dimensional analyses - Dimensions 2-5, Indonesian data

While this amount of misclassification may appear extensive, and problematic, it needs to be kept in perspective. Not dissimilar orders of misclassification are routine. In order to establish a framework for evaluating the extent of the increase in misclassification resulting from ignoring multidimensionality in the data, an analysis was done in which two classifications of the same candidates were done using data from the same test, for example by using items from different parts of the test, as in the split half reliability method. The reliability of the Indonesian test was 0.805.

Table 12 reports the errors in classification which are typically found in such an investigation, in this case for a unidimensional test whose reliability is 0.8.

How classified in repeated analyses	%
Same quintile	48.81
1 quintile apart	39.53
2 quintiles apart	10.77
3 quintiles apart	1.59
4 quintiles apart	0.00

Table 12: Comparison of classification in repeated classification from a unidimensional test whose reliability is 0.8

Ignoring multidimensionality in these data then does not add a different order of magnitude of misclassification. However, the misclassification is somewhat greater in the analyses summarized in Tables 10 and 11.

6. Conclusion

In this paper we have used data from two foreign language comprehension tests designed for beginning and intermediate level secondary school students to investigate the question of dimensionality associated with stimulus passages. New multidimensional Rasch models were used to explore the dimensionality associated with stimulus passages and its impact on scores and score reporting. The results show clear evidence of passage-related multidimensionality in the data. A comparison of classification of students into bands on the basis of the results under two conditions - ignoring or modelling the dimensionality - revealed an increase in error of classification of students when dimensionality is ignored, although the effect size, while real, was not great. While ignoring dimensionality may increase measurement error, the size of the error is relatively modest when compared with the error introduced by even modest amounts of unreliability in the test. Improvement should first go into improving the reliability of tests; but where test reliability is good, and where circumstances and

facilities permit, exploration of multidimensionality in data using methods such as those in this paper may be worthwhile.

7. References

Buck, G. (1992) *The construction of multidimensional data sets.* Paper presented at the 14th annual Language Testing Research Colloquium, Vancouver, February 27th-March 1st.

Buck, G. (1994) The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing 11, 2*: 145-70.

de Jong, J. H. A. L. and F. Stoyanova (1994) *Theory building: Sample size and data-model fit.* Paper presented at the 16th annual Language Testing Research Colloquium, Washington DC.

Hamp-Lyons, L. (1989) Applying the partial credit model of Rasch analysis: Language testing and accountability. *Language Testing 6, 1*: 109-18.

Henning, G. (1992) Dimensionality and construct validity of language tests. *Language Testing 9, 1*: 1-11.

McNamara. T. F. (1991) Test dimensionality: IRT analysis of an ESP listening test. *Language Testing 8, 2*: 45-65.

McNamara T. F. (1996) *Measuring second language performance.* London and New York: Addison Wesley Longman.

McNamara T. F. and R. J. Adams (1997) New approaches to the analysis of task- and rater-related dependencies in performance assessments. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.), *Current developments and alternatives in language assessment - Proceedings of LTRC 96* (pp. 625-635). Jyväskylä, Finland: University of Jyväskylä and University of Tampere.

McNamara T. F. and R. J. Adams (in press) The implications of halo effects and item dependencies for objective measurement. In M. Wilson and G. Engelhard (Eds.), *Objective Measurement, Volume 5.* Norwood NJ: Ablex.

Nunan, D. (1988) Commentary on the Griffin paper. In McNamara T.F. (Ed.) *Language testing colloquium. Selected papers from a Colloquium held at the Horwood Language Centre, University of Melbourne, 24-25 August, 1987. Australian Review of Applied Linguistics 11, 2*: 54-65.

Nunan, D. (1989) Item Response Theory and second language proficiency assessment. *Prospect 4, 3*: 81-93.

Skehan, P. (1984) Issues in the testing of English for specific purposes. *Language Testing 1, 2*: 202-20.

Skehan, P. (1989) Language testing. Part II. *Language Teaching 22, 1:* 1-13.

Wu, M. L, R. J. Adams and M. R. Wilson (1998) *ConQuest* [computer program]. Camberwell, Victoria: Australian Council for Educational Research.