Electronic Journal of Research in Educational Psychology
E-ISSN: 1696-2095
jfuente@ual.es
Universidad de Almería
España

Baghaei, Purya
The Application of Multidimensional Rasch Models in Large Scale Assessment and Validation: An Empirical Example
Electronic Journal of Research in Educational Psychology, vol. 10, núm. 1, 2012, pp. 233-252
Universidad de Almería
Almeria, España

Aplicación de modelos Rasch multidimensionales en la evaluación y validación a gran escala. Un ejemplo empírico.

Purya Baghaei

English Department, Islamic Azad University,
Mashhad Branch, Mashhad.

Irán

Correspondencia: Dr. Purya Baghaei. English Department, Faculty of Foreign Languages, Islamic Azad University, Mashhad Branch, Ostad Yusofi St., 91886-Mashhad, Iran . E-mail: pbaghaei@mshdiau.ac.ir.
Tel: +985116635064, Fax:+985116634763.

Resumen

Introducción. Las evaluaciones de comptencia y de diagnóstico, por lo general, comprenden varias subpruebas para proporcionar información sobre los diferentes componentes de las habilidades evaluadas. Por razones de practicidad estas subpruebas suelen ser cortas. El principal inconveniente de las subpruebas cortos es su baja fiabilidad y la imprecisión. Este estudio muestra cómo un modelo de Rasch compensatorio multidimensional, es decir, un Modelo Multidimensional y Multinomial de Coeficientes ateatorios (MRCMLM), puede utilizarse para las correlaciones entre las subpruebas para mejorar la estimación y la fiabilidad de las medidas.

Método. La muestra estuvo compuesta por 1021 estudiantes iraníes de secundaria. Una prueba de comprensión del idioma Inglés compuesto por cuatro de 10 puntos subpruebas se adminstretd a los participantes. Los datos fueron analizados con tanto MRCMLM y el modelo unidimensional de Rasch.

Resultados. Los resultados mostraron que las subpruebas que contienen sólo diez elementos pueden tener la fiabilidad de las pruebas de 45 elementos cuando habilidades de los estudiantes se calcula con la MRCMLM. También se demostró que el análisis multidimensional da una mejor estimación de las correlaciones entre verdaderos subpruebas y se puede utilizar como un método de confirmación para estudiar la estructura componencial de constructos de competencia.

Conclusiones. El modelo Multidimensional de Rasch puede mejorar la fiabilidad de los subtests cortos. El modelo también puede ser utilizado como un método confirmatorio para estudiar la estructura componencial de las construcciones de competencia y para mostrar la verdadera asociación entre las dimensiones de profificency.

Palabras claves: Modelo unidimensional y multidimensional, modelo de Rasch, multidimensional modelo aleatorio, coeficiente multinomial (MRCMLM), comprensión de segunda lengua, validación.

Recibido: 04/10/11 *Aceptación inicial: 05/11/11* *Aceptación final: 14/03/12*

The Application of Multidimensional Rasch Models in Large Scale Assessment and Validation: An Empirical Example

Abstract

Introduction. Accountability and diagnostic assessments usually comprise several subtests to provide information on the different components of the abilities tested. For practicality reasons such subtests are usually short. The major drawback of short subtests is their low reliability and imprecision. This study shows how a compensatory multidimensional Rasch model, namely, *Multidimensional Random Coefficient Multinomial Logit Model* (MRCMLM), can utilize the correlations among subtests to improve estimation and the reliability of measures.

Method. The sample was composed of 1021 Iranian high school students. An English language comprehension test composed of four 10-item subtests was adminstretd to the participants. The data were analysed with both MRCMLM and unidimensional Rasch model.

Results. Findings showed that subtests that contain only ten items can have the reliability of 45-item tests when students' abilities are estimated with the MRCMLM. It is also demonstrated that multidimensional analysis gives a better estimate of the true correlations among subtests and can be used as a confirmatory approach to study the componential structure of proficiency constructs.

Conclusions. Multidimensional Rasch model can improve the reliability of short subtests. The model can also be used as a confirmatory approach to study the componential structure of proficiency constructs and to show the true association among dimensions of profificency.

Keywords: unidimensional Rasch model, multidimensional Rasch model, multidimensional random coefficient multinomial logit model (MRCMLM), second languge comprehension, validation

Received: 10/04/11 *Initial acceptance: 11/05/11* *Final acceptance: 03/14/12*

Introduction

Accountability assessment systems are designed to measure statewide or nationwide student progress and evaluate school performance. Since it is diagnostic assessment which helps educational systems to focus on students' needs and weaknesses, accountability assessments are required, by law, to produce individual student diagnostic reports as well (Ligon, 2007). Constructing comprehensive diagnostic instruments to provide information at the subskill level is a challenging task. In order to do so, test developer has to identify the subskills underlying the intended skill and develop several items to measure each subskill. Separate ability measures on the combination of items written to measure each subskill, which is referred to as a subscale in psychometric terms, need to be reported to the examinees and test users. The wide range of skills and subprocesses which constitute educational constructs and constraints in testing time usually force test developers to construct short subtests which suffer from unreliability and imprecision problem.

As stated earlier accountability assessments are also used to diagnose learnres problems. To serve diagnostic purposes, more detailed information is dervived out of accountability assessments by reporting scores on specific content domains. That is, apart from reporting an overall, say, foreign language score several subscores on reading, grammar and vocabulary are also reported. Subscores can be used to evaluate the curriculum and instructional materials, identify students who need remedial instruction and put them in appropriate instructional programmes.

There are a number of approaches to report subscale scores in diagnostic tests or tests composed of multiple subtests. The simplest way is to report number-correct score or percentage score for each subscale. The drawback of this method is that raw scores are test-dependent and are not appropriate interval measures of students' abilities. Moreover, raw scores are not adjusted for form difficulty and do not permit comparisons across forms or populations (Yao & Boughton, 2007).

The other approach is to analyze the tests with a Rasch model or an IRT model. The standard Rasch/IRT models are unidimensional, however. Unidimensionality is a fundamental assumption of standard Rasch and IRT models. That is, all the items in an instrument should

measure one single trait (Lord, 1980). However, detailed diagnostic tests are more complex as such tests cover several dimensions and abilities. In order to capture the complexity of modern assessments, multidimensional models are required. Analyzing a test comprising of multiple subtests with a unidimensional model violates the unidimensionality assumption of these Rasch/IRT models. Moreover, application of composite unidimensional models on multidimensional tests can bias parameter estimation (Folk & Green, 1989) and results in the loss of information on subscales which in turn prevents diagnostic examination of persons' competence. Alternatively, one can use the consecutive approach (Davey & Hirsch, 1991), i.e., analyze each subtest separately, one subscale at a time with a unidimensional model and report Rasch/IRT scale scores for each subtest. The problem associated with this approach is that the reliability of subtests is very low due to the small number of items that cover each subscale, i.e., employing short subtests results in very imprecise measurements. The raw score method mentioned above suffers from this drawback too. Another drawback of the consecutive approach is its failure to use all available data, that is, the information which is contained in the other subscales of the test (Adams, Wilson, & Wang, 1997). The approach uses only the portion of the data which is related to a dimension and ignores the rest. A combined simultaneous analysis with a multidimensional model results in more stable and accurate item and persons parameter estimates because all available data are used.

In order to solve unreliability problem of short subtests a few solutions have been suggested. Yen (1987) proposed incorporating information from total test scores to improve subscale estimation. Wainer, et al. (2001) suggested using information from other subscales to stabilize subscale scores. Other suggested method is using collateral information, i.e., information from other traits, to increase the accuracy of subscale measures (Ackerman & Davey, 1991; Davey & Hirsh, 1991; Kahraman & Kamata, 2004). These methods only increase the precision of subscale scores if the subscales are highly correlated and the number of out-of-scale items is high. Correlations as high as .90 are required to increase subscale scores precision when out-of-scale information is used (Kahraman & Kamata, 2004). The use of collateral information about persons' educational background and demographic variables to improve measurement precision has also been suggested by Mislevy (1987).

The other method is using multidimensional Rasch or IRT models. Multidimensional models simultaneously calibrate several dimensions and capture the complexity of tests that measure several traits (Adams, et al., 1997; Kelderman, 1996; Rost & Carstensen, 2002; Yao

& Schwarz, 2006). These models are employed in situations where unidimensional models are not appropriate. That is, when assuming a single underlying trait to account for the observed responses is not justifiable. When using multidimensional models the items of each subscale are assigned to one or more dimensions and a multidimensional analysis is run. This procedure yields separate ability estimates for examinees on each dimension.

Adams et al. (1997) define two different types of multidimensionality: between-item multidimensionality and within-item multidimensionality. Between-item multidimensionality occurs when the item are unidimensional, i.e., each item measures a single latent trait and each item is assigned to one and only one dimension. Several combinations of items, which are referred to as subtests or subscales, are designed to measure multiple dimensions, but each item belongs to only one dimension. Within-item multidimensionality occurs when items or a handful of them are multidimensional, i.e., each item is designed to measure simultaneously more than one dimension. Consider an item which is intentionally written to test both grammar and vocabulary. So, some of the items are designed to belong to more than one dimension. Figure 1 shows that these two different types of multidimensionality graphically. The application of within-item multidimensionality has not become very common in educational testing yet. Another disadvantage of the consecutive approach is that modeling within-item multidimensionality, where one item can measure more than one dimension, is not possible.

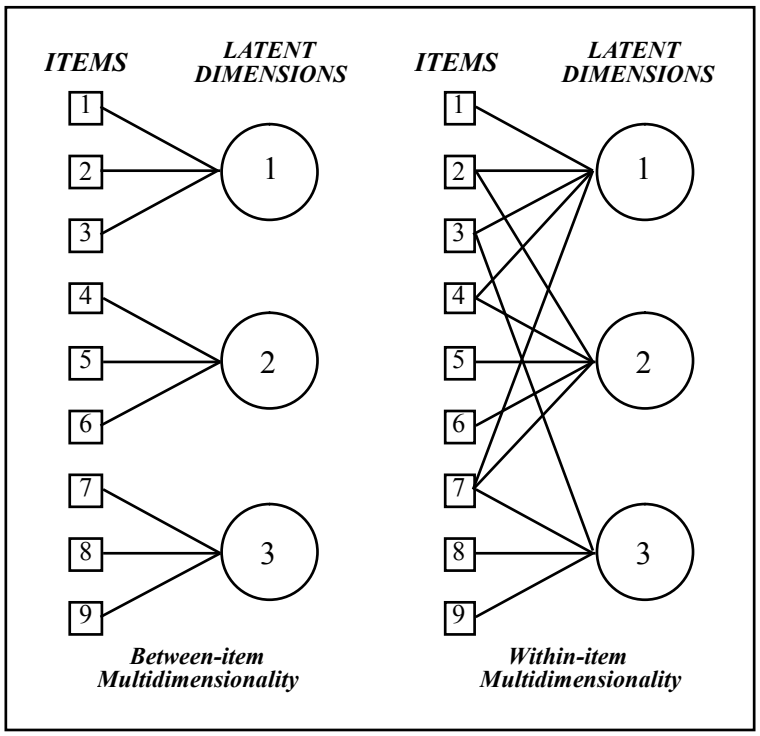

Figure 1. ***Within-item and between item multidimensionality (reprinted from Wu et al., 2007, with permission)***

Multidimensional models yield more precise measures because they estimate all the parameters jointly. In some of these models which are referred to as *compensatory* models, the correlations among dimensions are used as collateral information to provide additional information about examinees and to increase measurement precision. That is a low score on one dimension can be compensated for with a high score on another dimension. The greater the number of subscales and the higher the correlations among them the greater is the precision of measurement when the multidimensional approach is used (Wang, Chen, & Cheng, 2004).

Another advantage of multidimensional models is that because in these models measurement error is taken into account the correlations among dimensions are estimated directly and are free from *attenuation due to measurement error*. Since in the consecutive approach persons' abilities are estimated on the basis of short subtests they contain error and thus the correlations among subtests are underestimated and are not 'true' correlations (Mislevy,

1984). Deriving disattenuated correlations in the consecutive approach is based on the disattenuation correction formula which is a classical test theory concept and is not rooted in IRT (Wang & Chen, 2004).

Multidimensional models have also been implemented in computerize adaptive testing, referred to as multidimensional adaptive testing (MAT) versus unidimensional adaptive testing (UAT). Studies have shown that MAT achieves greater or comparable measurement precision with one third fewer items than UAT, when nine subtests with moderate to high correlations are in an instrument (Segall, 1996).

Method

Participants

A sample of 1021 Iranian high school students grade 1 to 3 (637 female, 384 male, aged 15-17), from different cities of Khorasan Razavi province was used for this study.

Instrument

The instrument was an English comprehension test containing 40 items on two skills of listening and reading comprehension. The instrument was designed as a diagnostic measure to provide information on test-takers' weaknesses and strengths in English comprehension skills. The data belonged to a province-wide project intended to diagnose four areas in second language comprehension in high school students in Khorasan Razavi province in Iran. The project had been funded and carried out by the Department of Public Education in Khorasan Razavi in 2009. The listening comprehension test was composed of two subtests of informational and interactional listening skills and the reading test had two subtests of expeditious and careful reading skills (Hughes, 2003). Each subtest contained ten multiple choice items. Therefore, the instrument contained four subtests of informational listening skills, interactional listening skills, expeditious reading skills and careful reading skills.

Statistical Analysis

The multidimensional model which is employed in this study is Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) (Adams, Wilson, & Wang, 1997). It is a compensatory multidimensional variant of unidimensional Rasch model which estimates all model parameters jointly on a common logit scale. Since MRCMLM is a mem-

ber of the family of Rasch models, it enjoys the measurement properties of these models including a sufficient statistic for parameter estimation. However, unlike unidimensional Rasch models, subtests' raw scores are not sufficient statistics for person estimates on the subtests. Vectors of raw scores which contain the raw scores over all subtests are sufficient statistic to estimate person measures over subtests (Cheng, Wang, & Ho, 2009). MRCMLM is a very flexible model and can accommodate a range of Rasch models including the dichotomous model (Rasch, 1960/1980), rating scale model (Andrich, 1978), partial credit model (Masters, 1982) and the facets model (Linacre, 1989) among other models. MRCMLM is implemented in ConQuest (Wu, Adams, & Haldane, 2007) programme and the Expected A Posteriori (EAP) estimation (Bock & Aitken, 1981) which utilizes correlations among dimensions to improve estimation and measurement precision is implemented in the programme.

Three models were fitted to the data and their overall fits were compared. First, the unidimensional form of Rasch's (1960/1980) dichotomous model was fitted. That is, all 40 items were modeled to load on a single dimension. Second, a two-dimensional dichotomous model, where the listening items were modeled to load on one dimension and the reading items were modeled to load a second dimension, was fitted. Third, a four-dimensional dichotomous model, where the subtests of informational listening skills, interactional listening skills, expeditious reading skills, and careful reading skills were modeled to load on separate dimensions, was fitted. ConQuest (Wu, et al., 2007) was used for the analyses.

Results

Model-data fit

Competing models are compared by comparing the likelihoods of their solutions. Because the multidimensional models are hierarchically related to the unidimensional model, that is, the models are nested, the fit of competing models can be compared with the change in their deviance (G^2). The greater the likelihood, the closer is the fitted model to the true model. The negative loglikelihood (logarithm of likelihood) or deviance is an index of the difference between the estimated model and the true model. Therefore, we expect the deviance to be small and models with smaller deviances are selected (Janssen & De Boeck, 1999). The deviance difference between two models is approximately chi-square distributed with the difference between the number of parameters degrees of freedom (Briggs &Wilson, 2003).

Table 1. ***Global fit statistics and information criteria for the models***

Model	G^2	Change in G^2	# of parameters	AIC	BIC
1-Dimensional	48395.20	-	41	48477	48679
2-Dimensional	48225.01	170.19	43	48311	48523
4-Dimensional	47900.84	324.17	50	48000	48246

Table 1 shows that the two-dimensional model has a smaller deviance than the unidimensional model and the four-dimensional model has a smaller deviance than the two-dimensional model. The change in the deviance from the unidimensional model to the two-dimensional model is statistically significant, $\chi^2(2) = 170.19$, $p<.01$. The change in the deviance from the two-dimensional model to the four-dimensional model is also statistically significant, $\chi^2(7) = 324.17$, $p<.01$. Therefore, the four-dimensional model which has the smallest deviance has the best fit. Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (*BIC*) (Schwarz, 1978) which are based on the deviance and also incorporate the number of estimated parameters and sample size also show that the four-dimensional model has the best fit as it has the smallest AIC and BIC.

Reliability, measurement error and sample statistics

It was stated in the introduction that MRCML model is a compensatory multidimensional Rasch model which uses the correlation among subtests to improve measurement precision and reliability. In this section the reliabilities of subtests, which contain only ten items, are compared in the multidimensional analysis and the consecutive analysis, i.e., when subtests are analyzed one at a time. Tables 2 and 3 show the reliabilities of the four subtests when they are analyzed one at a time, the so called consecutive analysis and when they are analyzed simultaneously in the multidimensional analysis.

Table 2. ***Statistics for the four dimensions in the multidimensional analysis (EAP estimation)***

	Dim. 1	Dim. 2	Dim. 3	Dim. 4
Reliability	.85	.85	.85	.86
Mean (Ability)	.43	-.53	-.52	.41
SD (Ability)	.98	.93	.91	1.53

Note: Dim. 1= informational listening, Dim.2= interactional listening, Dim. 3= expeditious reading, Dim. 4= careful reading

Table 3. ***Statistics for the four dimensions in the consecutive analysis (EAP estimation)***

	Dim. 1	Dim. 2	Dim. 3	Dim. 4
Reliability	.67	.58	.57	.78
Length Increment	280%	434%	450%	180%
Cronbach's Alpha	.66	.58	.57	.80
Mean (Ability)	.42	-.53	-.49	.42
SD (Ability)	1.04	.87	.82	1.57

Dimension 4 has the highest standard deviation, i.e., it has spread the subjects more widely compared to other subtests. The standard deviations of the other three scales are very similar, meaning that the persons were quite similar in their informational listening skills, interactional listening skills and expeditious reading skills but very different in their careful reading skills. The consecutive approach indicates more variability in the four subtests among the students as the standard deviations of the four subtests are more varied (Table 3). The reason is that the consecutive approach is not compensatory and students' measures on one subscale is not compensated by their measures on other subtests, therefore, more variability is observed in consecutive subtest ability estimates.

Dimension 1, 2, 3 and 4 have reliabilities of .85, .85, .85 and .86 respectively in the multidimensional analysis. Their reliabilities are .67, .58, .57 and .78 when these dimensions are analyzed separately with the (unidimensional) consecutive approach. Spearman-Brown Prophecy formula shows that if we use the consecutive approach, these subtests should be lengthened by 280%, 434%, 450% and 180% respectively to have the reliability indices that they have in the multidimensional analysis. The largest improvement in reliability was observed for Dimension 3 and the lowest improvement was observed for Dimensions 4. The reason for the greater improvement of Dimension 3 reliability is its higher correlation with other dimensions and the reason for rather smaller improvement of Dimension 4 reliability is its somewhat lower correlation with other dimensions.

Two-dimensional analysis

Wang and Chen (2004) argue that the more the number of subtests, the higher the correlations among them and the more scoring levels are in the items, i.e., when polychotomous items where subjects can score on a continuum depending on the quality of the answer they have provided rather than dichotomous items where there are only two response levels either right or wrong are used, the more efficient the multidimensional approach is. In order to test if having only two subtests is worth conducting a multidimensional analysis a two dimensional model was fitted to the data with reading and listening items as two dimensions and the precision of the two subscales of listening and reading in the multidimensional analysis and the consecutive analysis were compared. The two-dimensional analysis revealed that the direct correlation between listening and reading subtests is .83. The multidimensional reliability of the listening subtest was .79 and that of the reading subtest was .83. The reliabilities of listening and reading subtests in the consecutive analysis were .78 and .81 respectively and their raw score Cronbach's Alpha reliabilities were .76 and .81, respectively. The multidimensional reliabilities of the two subscales of listening and reading are higher than their corresponding consecutive reliabilities, although the improvement is not substantial. Multidimensional reliabilities were higher when we had four subtests albeit the subtests of the four-dimensional analysis had half the items that the two dimensional subtests had. It is observed that the improvement in reliabilities, which is brought about by the multidimensional model, when there are only two subtests is not very high, although the correlation between the subtests are quite high. It depends on the test developer to make a decision if the augmentation in reliability is worth utilizing a multidimensional model when there are only two subtests. It is important to note that if the correlation between the two subtests were higher (and if there were more subtests, of course) there would be more gain in measurement precision when the two-dimensional approach was used.

Item-person map and correlations

Item-Person map depicts the distribution of item difficulty estimates and person ability estimates on one scale and allows the direct comparison of items and persons in terms of targeting of the test and coverage of the scale. Figure 1 shows the distribution of subjects along the four dimensions. The subjects have generally performed better on the first dimension as the bulk of subjects are clustered more towards the upper part of the scale. Dimension 4 has distributed the subjects more widely as the subjects have covered a wider range of the scale.

Generally speaking the distributions of subjects do not match the distribution of item difficulties very well in this test. There are lots of persons who are below and above the hardest and easiest items. The mismatch is more pronounced for Dimensions 1 and 4 and less so for Dimensions 2 and 3.

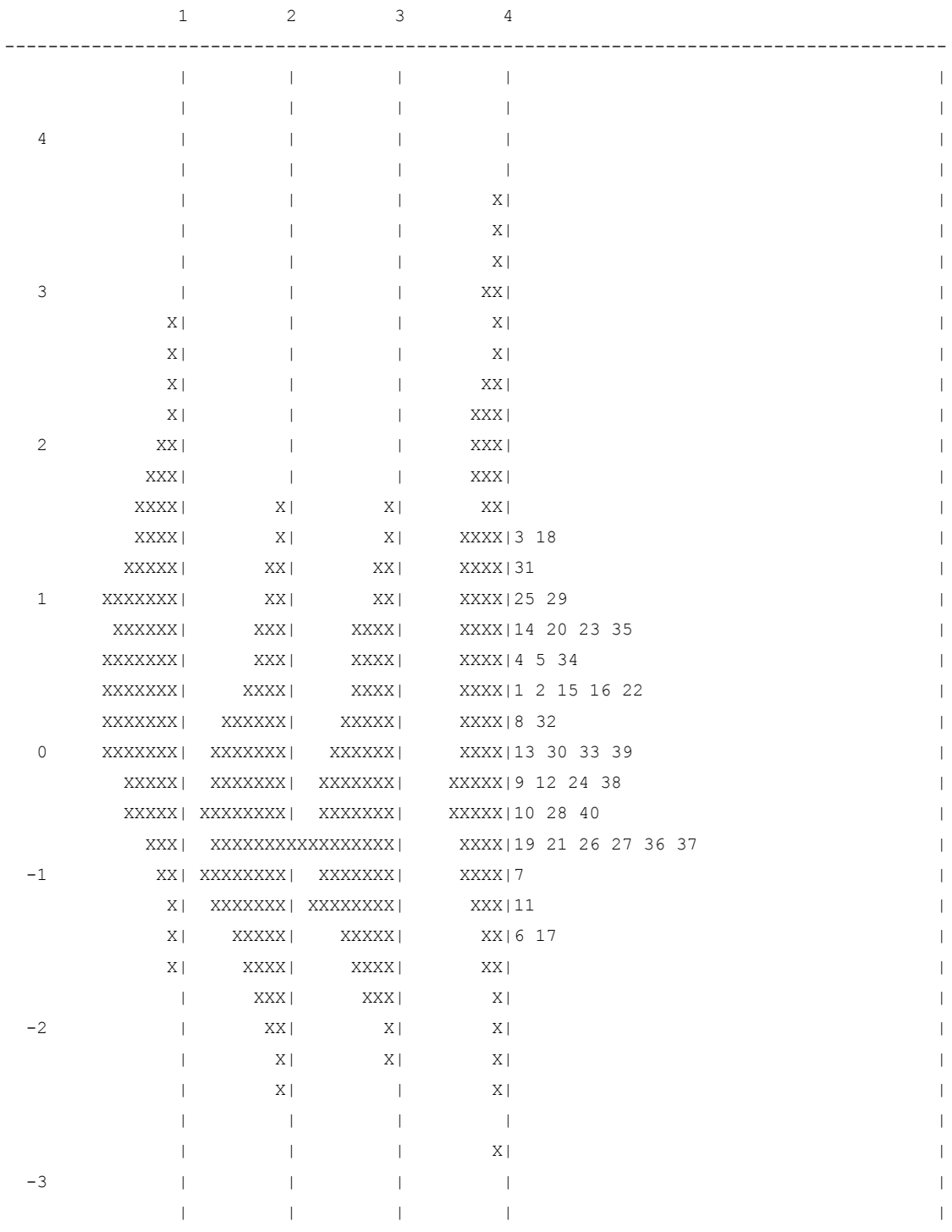

Figure 2. ***Map of latent distribution for the four dimensions***

Table 4 shows that the coefficients of correlations between the dimensions are much higher in the multidimensional model. Direct estimates of correlations in the multidimensional analysis are between .68 and .91, whereas, in the consecutive approach they are between .49 and .58. The correlations of raw scores are also given in parenthesis in the upper triangle of Table 4. They are very close to the consecutive correlations and in one case even slightly higher.

Table 4. ***Correlations between the four dimensions***

	Dim. 1	Dim. 2	Dim. 3	Dim. 4
Dim. 1	**1**	.58(.57)	.52(.52)	.49(.48)
Dim. 2	.91	**1**	.52(.53)	.49(.48)
Dim. 3	.86	.89	**1**	.54(.53)
Dim. 4	.68	.73	.81	**1**

Note: Consecutive correlations above diagonal; multidimensional correlations below diagonal; raw score correlations in parenthesis

Compensation for person estimates

As was stated before MRCMLM is a compensatory multidimensional Rasch model. This means that test-takers' low measures on one dimension can be compensated for with high measures on other dimensions.

Table 5. ***Raw scores, consecutive and multidimensional ability measures for three persons (EAP estimation)***

Student	Raw score				Consecutive Measure				Multidimensional Measure			
	Dim1	Dim2	Dim3	Dim4	Dim1	Dim2	Dim3	Dim4	Dim1	Dim2	Dim3	Dim4
1	8	6	4	9	1.16	.09	-.47	2.69	1.29	.25	.21	2.37
2	5	4	4	4	.13	-.49	-.47	-.31	.20	-.69	-.72	-.23
3	8	4	1	9	1.16	-.49	-1.73	1.93	.67	-.41	-.61	1.30

Table 5 shows that Students 1 and 2 have identical raw scores on Dimensions 3. Under the unidimensional Rasch model items and persons with similar raw scores will have similar difficulty and ability measures. The table shows that Students 1 and 2 have identical measures

of -.47 logit on Dimension 3 in the consecutive analysis, which is a unidimensional Rasch analysis. However, under the multidimensional model Student 1's ability on Dimension 3 is .21 logits and that of Student 2 is -.72 logits; .93 logits difference. The reason is that Student 1 has higher scores on the other three dimensions and his/her low score on Dimension 3 has been compensated for by his/her higher scores on the other dimensions.

Likewise, Student 3 has a very low score on Dimension 3; s/he has answered only one of the ten items of this subtest. However, s/he has scored eight out of ten on Dimension 1 and nine out of the ten on Dimension 4. Under the consecutive approach, where ability measures on subtests are estimated regardless of performance on other subtests, Student 3's ability measure on Dimension 3 is -1.73. Nevertheless, her/his ability measure on this dimension under the multidimensional model is -.61 logits. This ability measure is 1.12 logits higher than the ability measure estimated under the consecutive approach. The low score on this dimension is compensated for by high scores that this student has obtained on other dimensions.

Discussion and Conclusions

Diagnostic assessment of students' competencies requires precise measurement of their underlying constituents. This implies careful analysis of combinations of items that tap into the intended subskills of the intended construct and reporting student abilities at the subtest level. The problem associated with this approach is large measurement error and imprecise estimates of student abilities due to short test length. The other problem is that possible relationships among subskills cannot be accurately estimated, because correlations are attenuated due to large measurement error which is the direct consequence of short subtests. Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM) (Adams, et al., 1997) is a multidimensional variant of Rasch model which draws upon correlations among subtests to improve estimation and increase measurement precision.

The results of the present study showed that subtests which contain only ten items can have the reliability of subtests which have 18 to 45 items if the multidimensional model is utilized. This is a considerable benefit which is gained by just shifting from the common unidimensional analysis to a more recent multidimensional analysis of tests. Comparing this method with the conventional strategy that test developers need to implement to increase reliability, the multidimensional method is more efficient, practical and economical in time and

money. The conventional strategy is, of course, writing more items and lengthening the test which in turn demands longer testing time and more items to be written or exposed from the item bank.

The results also showed that the correlations among the dimensions are much higher when they are estimated directly with the multidimensional model. The consecutive approach yielded correlations which were close to the raw score correlations. This finding indicates that true associations among dimensions can be investigated more thoroughly with the multidimensional model.

Findings showed that students' ability measures on subtests are compensated for. In other words, if a person obtains a low score on a dimension it can be compensated for with high scores on other dimensions. The drawback of this, however, is that it is difficult to explain to test-takers and test-users why ability measures on one subtest should depend on their measures on another subtest. This means that person ability estimates on a dimension depend to a certain extent on ability on other dimensions, assuming the dimensions are correlated, which sounds problematic. One, however, could argue that the unexpectedly high or low estimate of a person on a dimension is the result of measurement error in the case of short tests and these estimates should be amended by means of EAP estimation otherwise we cannot justify why a person who has a high estimate on one subtest should have a low estimate a another subtest when the two subtests are highly correlated. Whether correlations imply unidimensionality is a different issue. Although the absence of a notable correlation implies distinct dimensions, a high correlation does not necessarily mean identicalness of the dimensions. It is possible to have distinct dimensions which are highly correlated. Height and weight are highly correlated in all populations but we do not consider them the same dimensions. Collateral information and information from correlations among subtests should cautiously be used in high-stakes assessments (Cheng et al., 2009).

The study implied that multidimensional models are confirmatory in nature and can be adopted, like confirmatory factor analysis, to investigate the componential structure of tests (Baghaei, 2011; Janssen & De Boeck, 1999). Confirmatory factor analysis has gained poularity among psychologists for test validation (see Gadelrab, 2011; Alqaryouti, Abu Hilal, & Ibrahim, 2011; Fernandez, et al, 2010; De la Fuente, Sander, Justicia, Pichardo, & Garcia-Berben, 2010). Mulidimensional Rasch model can be used is a similar fashion for test valida-

tion. This is accomplished by comparing the fits of several competing models with the help of deviance statistic and information criteria. The study showed that a four-dimensional model fits significantly better than a two-dimensional model and a two-dimensional model fits significantly better than a unidimensional model. This is clear evidence of the dimensionality of L2 comprehension proficiency and an indication that a four-dimensional model can better account for the structure of L2 comprehension construct. This concept is directly related to Messick's structural component of construct validity. Structural aspect of construct validity is mainly concerned with the scoring profile. It is highly important to take into account the structure of the test when scoring it. It is not sound to add up the scores of different parts of a test, when each part measures a different dimension. While one single score can summarize the performance of an individual on a unidimensional test, scores on different dimensions must be reported separately. In other words, the scoring models should be informed by the structure of the test (Baghaei & Grotjahn, in press; Messick, 1989).

Multidimensional models can be highly beneficial both for large scale testing purposes and construct validation research. Apart from enormously increasing measurement precision and reliability of tests, which is essential in high-stakes assessments, multidimensional Rasch and IRT models can inform research in componentiality of educational skills and construct validation research (Baghaei & Grotjahn, in press). The model is well worth further exploration by educational measurement researchers and practitioners for application in high-stakes tests and for exploring the nature of educational constructs.

References

Ackerman, T. A., & Davey, T. C. (1991, April). *Concurrent adaptive measurement of multiple abilities.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21,* 1-23.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*: 6, 716-723.

Alqaryouti, I. A., Abu Hilal, M. M., & Ibrahim, M. M. (2011). Validity and reliability of an attention deficit and hyperactivity disorder measure for a sample of Omani children. *Electronic Journal of Research in Educational Psychology, 9*(2), 911-930.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Baghaei, P., & Grotjahn, R. (in press). Establishing the construct validity of conversational C-Tests using a multidimensional Item Response Model. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test*: Current trends. Frankfurt/M.: Lang.

Baghaei, P. (2011, September). Validation of a multidimensional scale of willingness to communicate. Paper presented at the *Meeting of the Methodology and Evaluation Section* of the German Association of Psychology. Bamberg, Germany.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika. 46,* 443-459.

Briggs, D.C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87-100.

Cheng, Y.-Y., Wang, W.-C., & Ho, Y.-H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests. *Educational and Psychological Measurement, 69*, 369-388.

Davey, T., & Hirsh, T. M. (1991, April). *Examinee discrimination as measurement properties of multidimensional tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

De la Fuente, J., Sander, P., Justicia, F., Pichardo, M. C., & Garcia-Berben, A. B. (2010). Validation study of the scale for assessment of the teaching-learning process, student version (ATLP-S). *Electronic Journal of Research in Educational Psychology, 8*(2), 815-840.

Fernandez, M., Benitez, J. L., Pichardo, M. C., Fernandez, E., Justicia, F., & Garcia, T., et al. (2010). Confirmatory factor analysis of the PKBS-2 subscales for assessing social skills and behavioral problems in preschool education. *Electronic Journal of Research in Educational Psychology, 8*(3), 1229-1252.

Folk, V. G. & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

Gadelrab, H. F. (2011). Factorial structure and predictive validity of approaches and study skills inventory for students (ASSIST) in Egypt: A confirmatory factor analysis approach. *Electronic Journal of Research in Educational Psychology*, *9*(3), 1197-1218.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*, 245-268.

Kahraman, N., & Kamata, A. (2004). Increasing the precisions of subscale scores by using out-of scale information. *Applied Psychological Measurement, 28*, 407-426.

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement, 20*, 155-168.

Ligon, G. D. (2007). Why Eva Baker doesn't seem to understand accountability: The politimetrics of accountability. *Third Education Group Review* / Articles, 3(1). Available from http://www.thirdeducationgroup.org/Review/Articles/v3n1.pdf.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Lord, F. (1980). *Application of Item Response Theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Messick, S. (1989) Validity. In R.L. Linn (ed.) *Educational measurement* (pp. 13-103). New York: Macmillan.

Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359-381.

Mislevy, R.J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81-91.

Rasch, G. (1960/1980) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The university of Chicago Press, 1980).

Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement, 26*, 42-56.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461-464.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Wang, W.-C. & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 295-316.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., et al. (2001). Augmented scores: ‘‘Borrowing Strength’’ to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wu, M. L., Adams, R. J., & Haldane, S. A. (2007). *ACER ConQuest*. Australian Council for Educational Research.

Yao, L. & Boughton, K.A. (2007). A multidimensional Item Response Modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83-105.

Yao, L., & Schwarz R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30*, 469-492.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.