Examining the Validity of Different Assessment Modes in Measuring Competence in Performing Human Services

Hungi Njora
School of Education, Flinders University of South Australia

I Gusti Ngurah Darmawan
School of Education, Flinders University of South Australia

John P. Keeves
School of Education, Flinders University of South Australia *john.keeves@flinders.edu.au*

This article addresses an important problem that faces educators in assessing students' competence levels in learned tasks.

Data from 165 students from Massachusetts and Minnesota in the United States are used to examine the validity of five assessment modes (multiple choice test, scenario, portfolio, self-assessment and supervisor rating) in measuring competence in performance of 12 human service skills. The data are examined using two analytical theories, item response theory (IRT) and generalizability theory (GT), in addition a prior, but largely unprofitable examination using classical test theory (CTT) was undertaken.

Under the IRT approach with Rasch scaling procedures, the results show that the scores obtained using the five assessment modes can be measured on a single underlying scale, but there is better fit of the model to the data if five scales (corresponding to the five assessment modes) are employed. In addition, under Rasch scaling procedures, the results show that, in general, the correlations between the scores of the assessment modes vary from small to very strong (0.11 to 0.80). However, based on the GT approach and hierarchical linear modelling (HLM) analytical procedures, the results show that the correlations between scores from the five assessment modes are consistently strong to very strong (0.53 to 0.95). It is argued that the correlations obtained with the GT approach provide a better picture of the relationships between the assessment modes when compared to the correlations obtained under the IRT approach because the former are computed taking into consideration the operational design of the study.

Results from both the IRT and GT approaches show that the mean values of scores from supervisors are considerably higher than the mean values of scores from the other four assessments, which indicate that supervisors tend to be more generous in rating the skills of their students.

item response theory, generalizability theory, classical test theory,
self assessment, portfolio assessment, supervisor scaling,
scenario assessment, competences, measurement

INTRODUCTION

The general purpose of this study is to examine the validity of different assessment modes in measuring competence in the performance of human service workers, who supported people with disabilities. The data for this study were collected from 165 students in Massachusetts and Minnesota in the United States. Five assessment modes (to be called Multiple Choice, Scenario, Portfolio, Supervisor and Self-Assessment) were employed to measure the students' skill levels in performing 12 human service skills (to be called Competency 1 through to Competency 12). Except for the Multiple Choice, score values 1 to 4 were used to rate the students' skill level, with a low value denoting a less skilled student and a high value denoting a more skilled student. For the Multiple Choice mode of assessment, 10 items were included in the multiple-choice test to measure each competency, making a total of 120 items in the test. In order to make the scores on the Multiple Choice mode of assessment comparable to the other four modes of assessment, the scores from the multiple-choice test were collapsed into score values of 1 to 4. The multiple-choice items for each competency were checked to determine whether the items could be meaningfully added together, and then only those items with adequate fit were combined prior to the collapsing of the Multiple Choice scores.

In the planning stage of this study, it was recognized that it would be expensive (in terms of money and time) to collect data from each student using all the five assessment modes and for all the 12 competencies. Moreover, it was recognized that with a too extensive response task required of both students and assessor, there would be a serious risk of only partial completion of the assessment schedules. As a way of overcoming these problems, an overlapping design was carefully formulated for data collection. This overlapping design was such that common students linked the five assessment modes and the 12 competencies. Generally, data were collected for a majority of the students using at least two of the assessment modes and for at least three of the 12 competencies.

Table 1 provides a summary of the number of students who were assessed using each of the five assessment modes and the number of common students linking the five assessment modes, and Table 2 presents the corresponding information, but for the 12 competencies. In Table 1, the numbers given in bold are the total numbers of students assessed using each of the assessment modes while in Table 2, they are the total numbers of students assessed for each of the 12 competencies. For example, Table 1 shows that a total of 90 students were assessed using Scenario, a total of 94 students were assessed using Portfolio and so on. Likewise, Table 2 shows that a total of 134 students were assessed in Competency 1, a total of 138 students were assessed in Competency 2, and so on. By way of further examples, the meaning of the second entry in the first column of Table 1 is that a total of 81 students were assessed using both Scenario and Portfolio. The meaning of the corresponding entry in Table 2 is that a total of 121 students were assessed in both Competency 1 and Competency 2, and so on.

Table 1. Number of students assessed using the five assessment modes

	Number of Students				
	Scenario	Portfolio	Multiple Choice	Supervisor	Self-Assessment
Scenario	**90**				
Portfolio	81	**94**			
Multiple Choice	89	87	**153**		
Supervisor	78	86	106	**114**	
Self-Assessment	83	91	103	98	**113**

Table 3 gives the total numbers of students assessed for each of the 12 competencies using each of the five assessment modes. For example, Table 3 shows that the total number of students assessed for Competency 1 using Scenario, Portfolio, Multiple Choice, Supervisor and Self-Assessment

were 26, 33, 48, 82 and 106 respectively. When reading Table 3 it is important to recognize that the same student could be assessed for a particular competency using more than one of the five assessment modes.

Table 2. Number of students assessed in the 12 competencies

	Number of Students											
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Competency 1	**134**											
Competency 2	121	**138**										
Competency 3	126	128	**140**									
Competency 4	123	126	131	**140**								
Competency 5	126	127	132	130	**142**							
Competency 6	120	124	123	122	122	**133**						
Competency 7	124	124	131	128	130	125	**140**					
Competency 8	121	123	127	130	127	127	127	**138**				
Competency 9	125	127	133	130	132	125	131	128	**141**			
Competency 10	125	128	131	129	131	122	132	129	131	**140**		
Competency 11	125	122	127	126	130	120	126	124	124	125	**136**	
Competency 12	127	123	127	124	125	124	123	126	127	128	123	**137**

Note: C1 to C12 - Competency 1 to Competency 12.

Table 3. Number of students assessed in each competency using the five assessment modes

	Mode of Assessment				
	Scenario	Portfolio	Multiple Choice	Supervisor	Self-Assessment
Competency 1	26	33	48	82	106
Competency 2	32	31	50	95	107
Competency 3	29	27	53	96	110
Competency 4	35	28	46	93	109
Competency 5	27	38	49	99	107
Competency 6	27	32	53	52	107
Competency 7	23	35	51	103	111
Competency 8	29	31	50	92	109
Competency 9	30	35	48	93	108
Competency 10	34	25	52	111	109
Competency 11	25	27	56	64	105
Competency 12	36	31	46	72	110

RESEARCH QUESTIONS

The specific research questions addressed in this study within the general investigation of the validity of different assessment modes in measuring competence in the performance of human services are listed below.

1. Can the five assessment modes be used to obtain reliable measures?
2. Do the five assessment modes differ in their mean values and spread of scores?
3. Do the 12 competencies differ in their mean values and spread of scores?
4. Can the data be effectively combined? More specifically, do the data form a single underlying dimension, five underlying dimensions (corresponding to the five assessment modes) or 12 underlying dimensions (corresponding to the 12 competencies)?
5. What are the correlations between (a) the five assessment modes, and (b) the 12 competencies?
6. Are there significant interactions between the assessment modes and the competencies?

METHODS

In order to answer the above research questions, three data analysis theories were considered, namely: (a) classical test theory (see Keats, 1997, pp.713-719) (b) item response theory (see Stocking, 1997, pp. 836-840), and (c) generalizability theory (see Allal and Cardinet, 1997, pp 737-741).

Classical test theory (CTT) involves the examination of a set of data in which scores can be decomposed into two components, a true score and an error score that are not linearly correlated (Keats, 1997).

Under the classical test theory (CTT) approach, only correlations can be calculated between the item-case pairs. Thus, this approach yields a large number of correlations, which makes the results difficult to interpret and difficult to summarize. In addition, the correlations under CTT suffer from the small number of cases. Importantly, under this approach, using the small number of cases on which the correlation is based, there is no test of whether the combination of the items is admissible and no adjustment is made for differences in item difficulties. Moreover, the CTT approach does not take into consideration the operational design of this study (that is, assessment modes nested under competencies, see Figure 1). Consequently, it is found that the results based on the CTT approach do not provide a sound and meaningful picture of the relationships among the assessment modes (or competencies), and consequently this approach is not reported in this article.

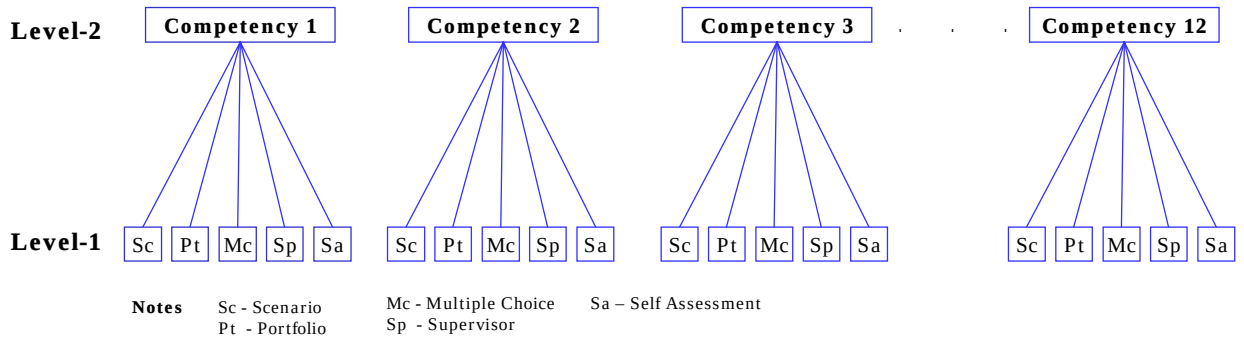

Figure 1. Operational design of the study

Rasch scaling is a procedure within item response theory (IRT) that uses a one-parameter model to transform data to an interval scale with strong measurement properties. It is a requirement of the model that the data must satisfy the conditions of unidimensionality in order for the properties of measurement to hold, namely to be independent of the tasks and the persons involved in providing data for the calibration of the scale (Allerup, 1997).

Under item response theory (IRT), a test is applied to indicate whether it is meaningful to combine the different components of interest in this study (that is modes, competencies and items). For example, under the one-parameter IRT (Rasch) model, the fit of the items and the fit of persons can be examined to test if it is appropriate to combine the data to form a single scale (see Keeves and Alagumalai, 1999, pp.23-42). If a single scale is admissible, then the components (assessment modes or competencies) can be compared and their mean values and spread of scores examined on common (and therefore meaningful) scales.

In addition, under the Rasch model, the scores are adjusted for the differences in difficulty levels of methods and items, which makes it possible to compare the different components. Thus, the IRT approach provides adjusted estimates and larger numbers of cases for the calculation of

correlations. In addition, the IRT approach yields fewer correlations compared to the classical test theory (CTT) approach, which makes the results easier to interpret and summarize.

Despite the advantage of the IRT approach in transforming the scores to an interval scale, the approach does not take into consideration the operational design of this study (that is, assessment modes nested under competencies). Consequently, it is unlikely that the results based on the IRT approach would provide a complete picture of the relationships among the assessment modes (or competencies). However, it should be remembered that, based on the IRT approach, it is meaningful to compare the properties of scores from the different assessment modes (or competencies), and therefore this approach is examined in this article.

An alternative approach uses generalizability theory (GT). Generalizability theory (GT) employs a framework based on analysis of variance procedures to estimate the sizes of effects, variance components, and reliabilities associated with the major sources of variation in a set of measurements made in education and the behavioural sciences (Allal and Cardinet, 1997).

Under the generalizability theory (GT) approach used in this article, the scores are not transformed to an interval scale, but the raw scores can be adjusted for differences in difficulty levels of the modes and competencies. It should be noted that, based on the GT approach, a nested ANOVA analytical procedure is capable of taking into consideration the operational design of the study. However, the complexity and highly unbalanced nature of the design prevents traditional ANOVA analytical procedures being used, but a hierarchical linear modelling (HLM) analytical procedure can be employed.

HLM is designed to analyze nested designs that are unbalanced and provides an empirical Bayes estimation procedure to adjust for imbalance, and for the relatively large number of empty cells or cells with small numbers of cases. There are, however, sufficient numbers of cases in a sufficient number of cells for satisfactory maximum likelihood estimation to be employed where traditional least square estimation procedures would probably fail to provide meaningful estimates. Based on the GT approach and HLM analytical procedures, correlations between the assessment modes are computed taking into account the variability between the competencies. Thus, the HLM analysis is not expected to give identical results to the IRT analysis since the assumptions made and the scales constructed differ.

In the sections that follow, analyses of the data using the IRT (Rasch) and GT (HLM) approaches are described, and the results of the analyses presented and discussed.

IRT APPROACH

In the Rasch analysis, the student scores obtained using the five assessment modes for all the 12 competencies were examined for their fit to the Rasch model. The main aim of the Rasch analysis was to examine whether these data form a single underlying dimension, five underlying dimensions (corresponding to the five assessment modes) or 12 underlying dimensions (corresponding to the 12 competencies).

A preliminary task using the Rasch analysis was to merge the data sets of the five assessment modes and 12 competencies so that they could be analyzed as a single data set. In the combined data set, each of the 12 competencies was represented five times (that is, one time for each assessment mode). Thus, for each student, the number of item slots in the combined data matrix that were to be filled with scores was 60 (that is, 5 assessment modes by 12 competencies), which means that the total number of items in the combined data set was 60. For a particular student, scores were entered in the item slots for the assessment modes and competencies the student was involved in, and blank spaces were left in the item slots for assessment modes and competencies

that the student was not involved in. However, it should be recognized that the assessment modes and competencies are linked together by common students, and therefore, these data can be analyzed together.

The first task in the Rasch analysis was to examine whether it was appropriate to combine the data sets from the five assessment modes and the 12 competencies so as to enable measurement of students' skills on a one-dimension scale (to be called 1-dimension model). For comparison purposes, this task was undertaken using two leading Rasch analysis computer programs: CONQUEST (Wu and Adams, 1998) and RUMM (Andrich et al., 2000). The second and third tasks were aimed at examining whether it was more appropriate to combine these data sets so as to enable measurement of students' skills on a five-dimension scale (to be called 5-dimension model) or a 12-dimension scale (to be called 12-dimension model) rather than on a one-dimension scale. The second and third tasks were undertaken using only CONQUEST because the current version of RUMM did not allow multidimensional modelling of data.

Unidimensional Rasch Analysis

In the paragraphs that follow, the results of the Rasch analysis described above are outlined and discussed. However, for reasons of parsimony, only the results obtained from the 1-dimension model using RUMM have been reported in full detail. In the last part of this section, the deviance statistics obtained using CONQUEST are used to compare the fit of the three models (that is, 1-, 5- and 12-dimension models) to these data.

The outputs generated by RUMM and CONQUEST provide information (in the form of fit statistics) that shows the compatibility of the Rasch model to the data and information (item and person estimates) that shows the location of items and persons on a Rasch measurement scale. For the 1-dimension model, summary fit statistics obtained using RUMM show that this model has ‘good’ fit to these data (based on a separation index of 0.76). For the same model, individual item fit and individual person fit results obtained using both RUMM and CONQUEST indicate that a vast majority of the items and persons have adequate fit. For example, using RUMM, 51 of 60 items have adequate fit (chi-square $p>0.05$) and 154 of 165 cases have adequate fit (residual $<|2.00|$). For the items, the results from the RUMM analysis with the 1-dimension model are discussed in greater details in the paragraphs that follow.

Table 4 presents the results of individual item fit and location following the analysis of the 1-dimension model using RUMM. The first four columns of Table 4 provide information regarding the identity of the item and the number of cases involved in that item (data points). The fifth to the eight columns of Table 4 provide information regarding the fit of the item. In Rasch analysis, 'residuals' are the differences between the Rasch-model-predicted response patterns and the observed response patterns obtained from the data. For items, RUMM allows residuals to be examined through a set of fit statistics: chi-square, degrees of freedom and probability, which are assessed in relation to the number of data points used to obtain the item statistics in order to decide whether the Rasch model fits the item or not. In general, for the small number of data points involved in this study (around 100 or less), the chi-square probability value of 0.05 (or higher) indicates that the model has a sufficient fit. Thus, it can be seen from the results presented in Table 4 that most of the items (51 out of 60) are consistent with the Rasch model.

The ninth to the eleventh columns of Table 4 give the estimated difficulty level of the item (location), the standard error of this estimate (SE) and the order of the item from least difficult to most difficult (rank), respectively. It should be noted that Rasch measurement scales are not ratio scales but interval scales, with a zero point that is commonly and arbitrarily located at the mean difficulty level of the items under consideration.

Table 4. Item characteristics based on 1-dimension model

Mode	Compt	Item Code	Data Points	Residual	DF	Chi-Sq	Prob	Location	SE	Rank
Scenario	1	sc01	26	-0.22	23.59	6.89	0.01 λ	0.37	0.33	38
	2	sc02	32	-0.47	29.03	3.42	0.16	-0.31	0.29	23
	3	sc03	29	-0.21	26.31	1.41	0.48	-0.05	0.24	28
	4	sc04	35	-0.37	31.75	3.32	0.17	0.48	0.27	42
	5	sc05	27	0.65	24.50	4.53	0.08	1.57	0.28	57
	6	sc06	27	-0.03	24.50	0.59	0.74	2.14	0.35	59
	7	sc07	23	0.50	20.87	0.13	0.94	0.51	0.27	43
	8	sc08	29	-0.09	26.31	1.47	0.47	2.03	0.30	58
	9	sc09	30	-0.20	27.22	1.01	0.59	0.39	0.32	39
	10	sc10	34	0.06	30.85	0.37	0.83	1.51	0.38	54
	11	sc11	25	-0.26	22.68	2.42	0.28	1.27	0.40	51
	12	sc12	36	-0.56	32.66	0.62	0.73	0.34	0.34	37
Portfolio	1	pt01	33	1.56	29.94	9.08	0.00 λ	0.32	0.22	36
	2	pt02	31	0.25	28.12	3.80	0.13	-0.51	0.20	17
	3	pt03	27	0.30	24.50	0.54	0.76	1.47	0.43	53
	4	pt04	28	1.02	25.40	1.70	0.41	0.98	0.31	47
	5	pt05	38	0.83	34.47	1.26	0.52	1.09	0.21	48
	6	pt06	32	-0.49	29.03	2.52	0.26	0.97	0.28	46
	7	pt07	35	1.31	31.75	2.01	0.35	0.88	0.23	44
	8	pt08	31	-0.45	28.12	1.19	0.54	2.49	0.30	60
	9	pt09	35	1.32	31.75	3.13	0.19	1.22	0.25	50
	10	pt10	25	0.90	22.68	1.23	0.53	1.57	0.33	56
	11	pt11	27	0.46	24.50	0.41	0.81	1.57	0.37	55
	12	pt12	31	0.13	28.12	5.27	0.05	1.37	0.29	52
Multiple	1	mc01	47	1.19	42.64	2.40	0.28	0.43	0.15	40
Choice	2	mc02	48	0.76	43.55	0.63	0.72	0.25	0.16	35
	3	mc03	53	-0.57	48.08	0.07	0.97	-0.46	0.15	19
	4	mc04	45	1.60	40.83	3.16	0.19	-0.04	0.14	29
	5	mc05	47	-0.28	42.64	0.65	0.72	0.03	0.15	30
	6	mc06	50	0.47	45.36	0.77	0.67	0.07	0.14	31
	7	mc07	49	1.63	44.45	1.34	0.50	-0.17	0.14	25
	8	mc08	49	-0.37	44.45	0.63	0.72	1.14	0.19	49
	9	mc09	47	1.55	42.64	4.19	0.10	-0.92	0.16	13
	10	mc10	51	0.66	46.27	4.59	0.08	0.90	0.20	45
	11	mc11	56	-0.03	50.80	0.81	0.66	-0.46	0.16	20
	12	mc12	47	1.54	42.64	5.78	0.03 λ	-0.48	0.16	18
Supervisor	1	sp01	82	-0.12	74.39	1.95	0.36	-2.09	0.21	4
	2	sp02	95	-0.33	86.19	3.08	0.19	-1.80	0.16	5
	3	sp03	96	-0.27	87.09	4.06	0.11	-1.24	0.15	11
	4	sp04	93	2.33	84.37	5.29	0.05	0.11	0.16	32
	5	sp05	99	-0.60	89.82	6.30	0.02 λ	-2.29	0.16	2
	6	sp06	52	-1.24	47.18	5.03	0.06	-2.35	0.26	1
	7	sp07	103	-0.66	93.44	1.78	0.40	-1.80	0.15	6
	8	sp08	92	0.18	83.47	0.93	0.62	-0.90	0.17	14
	9	sp09	93	2.34	84.37	5.95	0.03 λ	-0.13	0.12	27
	10	sp10	111	-0.40	100.70	1.17	0.55	-1.79	0.14	7
	11	sp11	64	-0.51	58.06	0.62	0.73	-1.47	0.20	10
	12	sp12	72	0.40	65.32	0.25	0.88	-2.18	0.21	3
Self	1	sa01	106	0.39	96.17	2.19	0.32	-1.63	0.17	9
Assessment	2	sa02	107	-0.05	97.07	6.56	0.01 λ	-0.43	0.20	21
	3	sa03	110	4.52	99.80	12.79	0.00 λ	-0.25	0.10	24
	4	sa04	109	-0.02	98.89	2.21	0.31	-0.16	0.18	26
	5	sa05	107	-0.63	97.07	5.99	0.03 λ	-0.55	0.21	16
	6	sa06	107	0.18	97.07	3.06	0.20	-1.09	0.18	12
	7	sa07	111	-0.75	100.70	6.73	0.01 λ	0.43	0.17	41
	8	sa08	109	0.23	98.89	0.15	0.93	0.22	0.16	34
	9	sa09	108	2.23	97.98	4.91	0.06	-0.36	0.15	22
	10	sa10	109	-0.38	98.89	1.71	0.41	0.12	0.17	33
	11	sa11	105	0.75	95.26	3.32	0.17	-0.59	0.17	15
	12	sa12	110	-0.52	99.80	4.03	0.11	-1.75	0.16	8

Note: λ - Item fit is suspect (chi-square probability <0.05).

Items with location value equal to zero or close to zero (e.g. mc05 and sc03), are of average difficulty, items with large positive location values (e.g. pt08 and sc06) are so-called 'difficult' items while those with large negative location values (e.g. sp06 and sp05) are so-called 'easy' items.

Thus, the location values and ranks presented in Table 4 show that most of the supervisor items are relatively easy compared to the items in the other four modes of assessment. In other words, supervisors tend to be lenient in rating the skills of their students.

Table 5 displays descriptive statistics for the items in the five assessment modes and for the 12 competencies. For example, for the competencies, the mean locations are obtained by taking the average of the locations of the items (results in Table 4) in each competency.

Table 5. Descriptive statistics of items by assessment modes and competencies (1-dimension model)

	Mean Location	Standard Error	Standard Deviation
Scenario	0.85	0.23 $^{\xi}$	0.81
Portfolio	1.12	0.21 $^{\xi}$	0.73
Multiple Choice	0.02	0.17	0.59
Supervisor	-1.49	0.24 $^{\xi}$	0.82
Self-Assessment	-0.50	0.20	0.69
Competency 1	-0.52	0.55	1.24
Competency 2	-0.56	0.34	0.75
Competency 3	-0.11	0.44	0.99
Competency 4	0.27	0.21	0.46
Competency 5	-0.03	0.68	1.52
Competency 6	-0.05	0.78	1.75
Competency 7	-0.03	0.47	1.06
Competency 8	1.00	0.61	1.37
Competency 9	0.04	0.36	0.81
Competency 10	0.46	0.62	1.39
Competency 11	0.06	0.58	1.30
Competency 12	-0.54	0.65	1.46

Note: ξ - The mean location (taken in absolute terms) is more than twice its standard error (i.e. significantly different from the scale zero).

From the results presented in Table 5, it would seem that it was much easier for students to get higher scores if assessed by their supervisors than if they were assessed using the other four assessment modes. Interestingly, these results also indicate that supervisor ratings are more lenient that are self-assessment ratings.

It also appears that the Multiple Choice assessment mode has a smaller spread of scores than the other four modes, all of which have similar standard deviations. In addition, it should be noted that there are sizeable differences between the mean scores for the different assessment modes, and as a consequences the different modes would not produce a consistent assessment grade, unless further adjustments were made.

For the competencies, it appears that Competency 5 is of near average difficulty (-0.03), Competency 2 is the easiest (-0.56), and Competency 8 (1.00) is the hardest. However, it is also noted that apart from Competency 4 (with a standard deviation of 0.46), most competencies have similar spreads and, all 12 competencies have mean locations that are not significantly different from zero, which seems to suggest that there are only small differences between the mean values of these competencies. Nevertheless, the distribution of the scores for a mode or a competency is best assessed not only in terms of the mean value but also in terms of the spread of the students' scores associated with that mode or competency which are presented in Figures 2 and 3.

Multidimensional Rasch Analysis

In the paragraphs that follow, the results of CONQUEST runs of the 5- and 12-dimension models are outlined.

Figures 2 and 3 display the item map obtained following CONQUEST runs of the 5- and 12-dimension models respectively. In Figure 2, dimensions '1', '2', '3', '4' and '5' are Scenario, Portfolio, Multiple Choice, Supervisor and Self-Assessment respectively, and in Figure 3, dimensions '1' through to '12' are Competency 1 through to Competency 12 respectively.

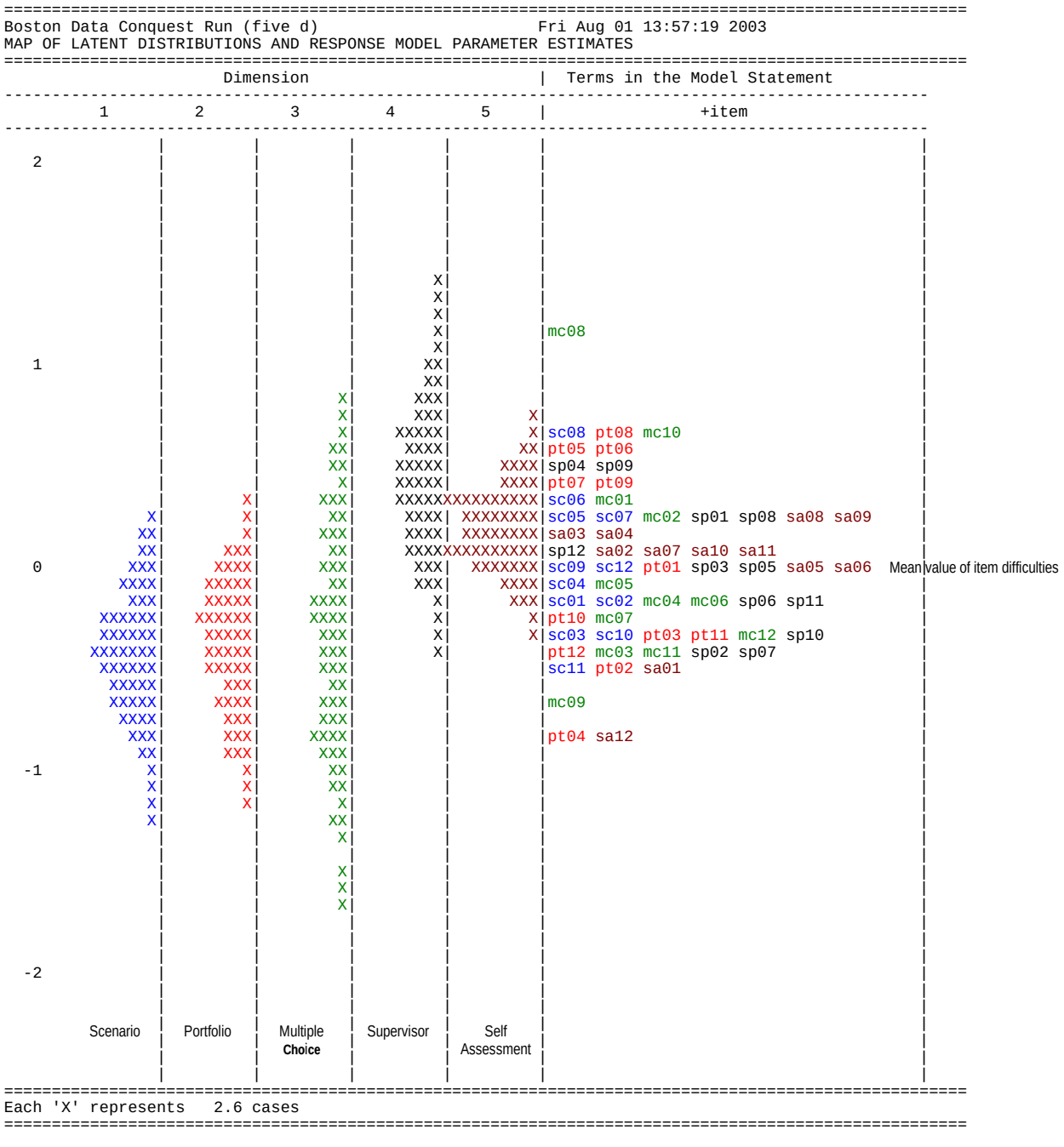

Figure 2. Item map based on the 5-dimension model

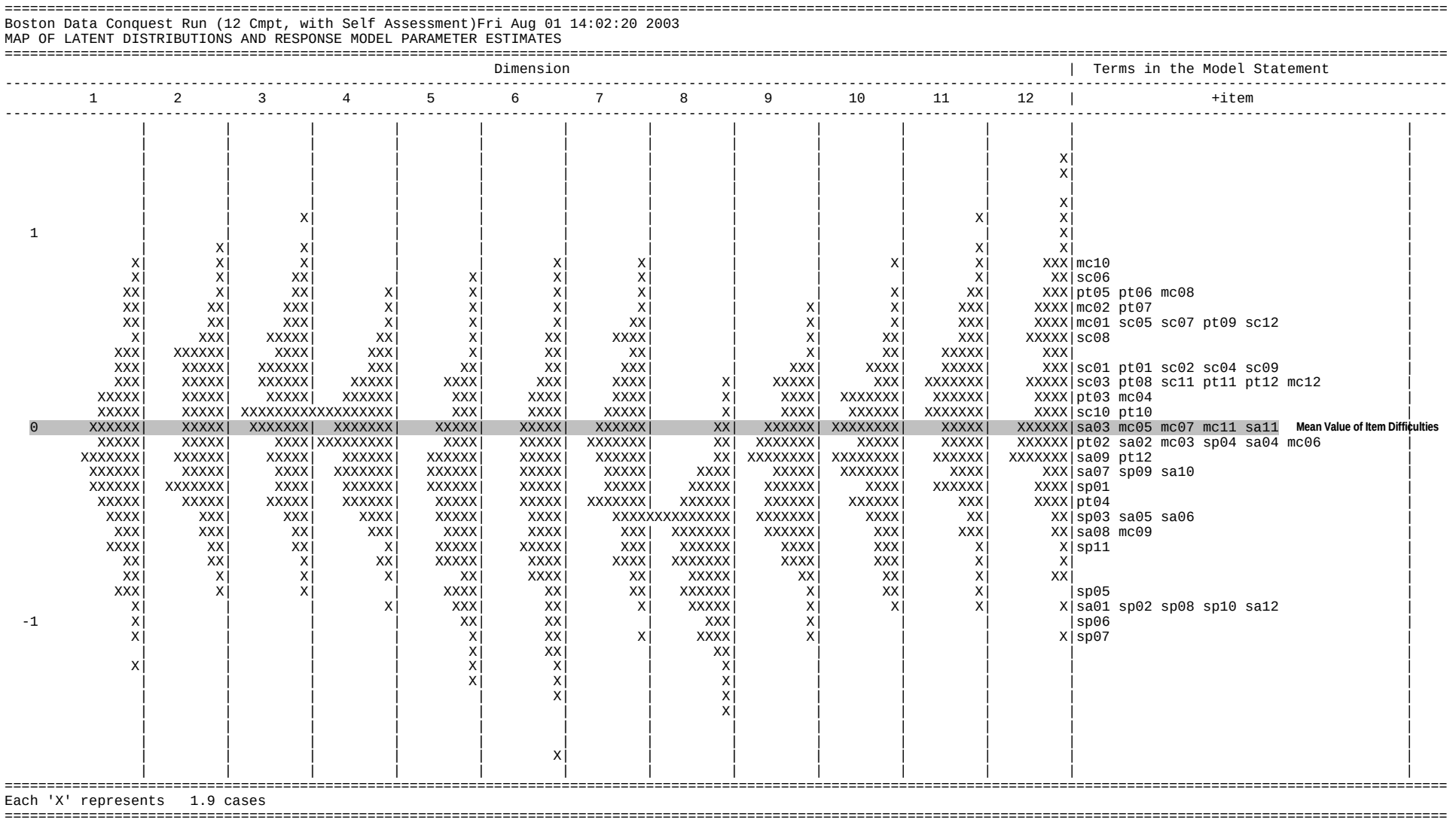

Figure 3. Item map based on the 12-dimension model

From Figure 2, it is evident that the students are more likely to be rated as above average when assessed by their supervisors than when assessed using the other four assessment modes. The spread of the ratings for each dimension gives a graphical indication of the size of the differences in the mean value between the five different modes.

Thus, Figure 2 seems to confirm what is found using the 1-dimension model, that is, supervisors tend to be more generous when rating their students in performing human service skills. In addition, Figure 2 indicates that, in general, Scenario, Portfolio and Multiple Choice yield scores that have almost equal means, and that Supervisor and Self-Assessment yield scores that have similar means.

The item map from the CONQUEST run of the 12-dimension model (Figure 3) indicates that, except for Competency 8 which has a low mean value, all the other competencies yield scores that generally do not differ markedly in terms of mean and spread. Again, this seems to confirm what is suggested following the 1-dimension analysis: there are very small differences between the means of the competencies. It should be noted that the scores plotted in Figures 2 and 3 have been adjusted for the differences between the means of the assessment modes and competencies.

Consequently, Tables 6 and 7 show that the correlations between the students' scores obtained for the five assessment modes and for the 12-competencies from the CONQUEST analyses of the 5- and 12-dimension models respectively. The values of the correlations differ from those that would be obtained with the raw data under classical test theory (CTT), because the differences between the modes and competencies have been removed, prior to the calculation of the correlations that are shown.

Table 6. Correlations of individual scores between assessment modes based on the 5-dimension model

	Scenario	Portfolio	Multiple Choice	Supervisor	Self-Assessment
Scenario	1.00				
Portfolio	0.48	1.00			
Multiple Choice	0.80	0.48	1.00		
Supervisor	0.34	0.42	0.52	1.00	
Self-Assessment	0.17	0.43	0.11	0.23	1.00

Table 7. Correlations between competencies based on the 12-dimension model

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Competency 1	1.00											
Competency 2	0.39	1.00										
Competency 3	0.33	0.42	1.00									
Competency 4	0.28	0.29	0.28	1.00								
Competency 5	0.48	0.50	0.41	0.31	1.00							
Competency 6	0.44	0.49	0.32	0.36	0.59	1.00						
Competency 7	0.41	0.50	0.44	0.34	0.51	0.51	1.00					
Competency 8	0.38	0.35	0.34	0.27	0.45	0.46	0.37	1.00				
Competency 9	0.24	0.25	0.35	0.21	0.36	0.30	0.27	0.33	1.00			
Competency 10	0.38	0.37	0.33	0.30	0.46	0.41	0.39	0.38	0.32	1.00		
Competency 11	0.39	0.42	0.32	0.31	0.49	0.46	0.41	0.38	0.27	0.40	1.00	
Competency 12	0.42	0.39	0.33	0.34	0.47	0.46	0.49	0.38	0.29	0.42	0.44	1.00

Except for Self-Assessment, the results in Table 6 show moderate to strong correlations (Cohen, 1992; p.157) between the scores obtained using the other four modes of assessment. Thus, with the exception of self-assessment, it appears that the ranking of students based on scores obtained from any of the assessment modes does not differ markedly from the ranking obtained using

scores from the other assessment modes. For Multiple Choice and Scenario, the correlation is very strong (0.80), which suggests a high degree of agreement between the ranks obtained using these two assessment modes after allowance has been made for systematic differences between the modes and competencies.

In addition, the results in Table 6 show mostly small correlations between Self-Assessment and the other assessment modes, which suggest small agreement between the ranks obtained using Self-Assessment and the ranks obtained using the other four assessment modes.

For the competencies, the results in Table 7 show mostly moderate correlations between the scores for the 12 competencies. However, it should be noted that a few of the correlations are small (for example, those involving Competency 9) and some are strong (for example, those between Competencies 2, 5, 6 and 7). Nevertheless, it can be concluded that a considerable number of students who are rated highly on one of the competencies are in most cases also rated highly on the other competencies.

In Table 8, the fit of 1-, 5- and 12-dimension models are compared using the deviance statistics (obtained from output generated by CONQUEST) and chi-square tests. In Table 8, the fit of the 12-dimension model is compared to the fit of the 1-dimension model, and the fit of the 1-dimension model is compared to the fit of the 5-dimension model. Chi-square tests presented in Table 8 indicate better fit of the 1-dimension model compared to the 12-dimension model and better fit of the 5-dimension model compared to the 1-dimension model. Therefore, the 5-dimension model has the best fit to these data.

Table 8. Comparison of model fit using difference in deviance statistics

Model	Deviance Statistic	Number of Parameters	Chi-square Statistic	Degrees of Freedom
12-dimension	9201.77	138		
1-dimension	9011.92	61	189.86	77
5-dimension	8950.91	75	61.01	14

GENERALIZABILITY THEORY APPROACH

The design of this study provides data that have a multilevel structure, that is, five assessment modes are nested beneath 12 competencies (see Figure 1). No allowance for this aspect of the design is made using IRT. Consequently, this section examines the relationships between the assessment modes and the competencies taking into consideration the operational design of the study. The computer package used for the multilevel analyses in this study is HLM5 developed by Raudenbush, Bryk, Cheong and Congdon (2000).

The main task before the HLM analysis was to construct dummy variables for the assessment modes and competencies. For the assessment modes, five dummy variables (scenario, prtfolio, mchoice, supvisor and selfasmt) were constructed to denote Scenario, Portfolio, Multiple Choice, Supervisor and Self-Assessment. In coding of the data a ‘1’ was used to indicate a student’s score obtained using that assessment mode and a ‘0’ was used to indicate a student’s score obtained using the other four assessment modes. Similarly, for the competencies, 12 dummy variables (compt01, compt02, compt03, . . . , compt12) were constructed to denote the 12 competencies.

Specification of HLM models

It should be noted that, with the operational design described above and using HLM, only a maximum of four variables denoting assessment modes can be included in an analysis simultaneously, leaving the fifth variable as a dummy for balancing the analysis. For the purposes of this study, it was considered important to examine the relationship between Supervisor and the

other four assessment modes. Consequently, a decision was made to run two models: one model with Self-Assessment as the dummy for balancing the analysis (to be called Model-M), and the other model with Multiple Choice as the dummy for balancing the analysis (to be called Model-S). That is, for Model-M, Multiple Choice was included in the analysis and Self-Assessment excluded from the analysis while for Model-S, Self-Assessment was included in the analysis and Multiple Choice excluded from the analysis.

For example, following the notations and arguments presented by Raudenbush and Byrk (2002), Model-M can be described as follows.

Level-1 model

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{scenario})_{ij} + \beta_{2j}(\text{prtfolio})_{ij} + \beta_{3j}(\text{mchoice})_{ij} + \beta_{4j}(\text{supvisor})_{ij} + r_{ij}$$

Level-2 model

$$\beta_{0j} = \gamma_{00} + \gamma_{0j}C_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{1j}C_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{2j}C_j + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{3j}C_j + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + \gamma_{4j}C_j + u_{4j}$$

[Equation 1]

where:

Y_{ij} is the score (skill level) of student *i* for Competency *j* (C_j);

β_{0j} is the mean score of Competency *j*;

β_{1j}, β_{2j}, β_{3j}, and β_{4j}, are the regression coefficients associated with Competency *j* for Scenario, Portfolio, Multiple Choice and Supervisor assessment modes respectively;

r_{ij} is a random error or 'student effect', that is, the deviation of the student mean from the competency mean;

γ_{00} is the grand mean;

$\gamma_{0j}C_j$ is the direct effect of Competency *j* on the mean score of the students;

$\gamma_{1j}C_j$, $\gamma_{2j}C_j$, $\gamma_{3j}C_j$, and $\gamma_{4j}C_j$, are cross-level interaction effects between Competency *j* and the assessment modes (i.e. Scenario, Portfolio, Multiple Choice and Supervisor respectively);

u_{0j} is a random 'competency effect', that is, the deviation of the competency mean from the grand mean; and

u_{1j}, u_{2j}, u_{3j}, and u_{4j} are random effects associated with the interaction between Competency *j* and the assessment modes (i.e. Scenario, Portfolio, Multiple Choice and Supervisor respectively).

The indices *i* and *j* denote students and competencies where there are

i = 1, 2, . . . , n_j students assessed for competency j; and
j = 1, 2, . . . , J competencies (in this study J=12).

In simple terms, Equation 1 shows that at Level-1, the student score is modelled as a function of a competency mean, assessment modes and a random error, and at Level-2, each competency mean, β_{0j}, is viewed as an outcome varying randomly around some grand mean. For purposes of parsimony, C_j in Equation 1 is used to model the direct and cross-level interaction effects associated with all the 12 competencies. However, it should be noted that, in the actual analyses, only those competencies that have significant effects ($p<0.05$) are included in the model.

The equation for Model-S is similar to the equation for Model-M (i.e. Equation 1). However, it should be remembered that Model-S has the variable selfasmt (Self-Assessment) instead of the variable mchoice (Multiple Choice).

HLM analysis

The two models described above (Models M and S) are estimated using a four step procedure. The first step involves running a null model in order to obtain the amounts of variance available to be explained at each level of the hierarchy (Bryk and Raudenbush, 1992). The null model is the simplest model because it contains only the dependent variable (for this study, student score) and no predictor variables are specified at any level.

In the second step, the four dummy variables that represent the assessment modes (i.e. scenario, prtfolio, mchoice, and supvisor for Model-M, and scenario, prtfolio, selfasmt, and supvisor for Model-S) are included in the analysis at Level-1 simultaneously. At this second stage, no predictors are specified at Level-2, and therefore, Raudenbush et al. (2000) have referred to this type of model as 'unconditional' at Level-2. It is considered important to keep all four dummy variables in each model in subsequent stages of the analysis regardless of whether or not the variable makes a significant contribution overall because the variables may have significant cross-level interaction effects with Level-2 variables (competencies). Moreover, it is necessary to include these four dummy variables because a major aim of this study is to examine the mean scores of the students obtained using all the assessment modes employed in this study.

The third step of the analysis involves building up the Level-2 intercept model through adding the significant ($p<0.05$) competency-related dummy variables into the model. At this stage, the exploratory analysis sub-routine available in HLM5 is employed for examining the inclusion of potentially significant dummy variables that represent the 12 competencies (i.e. compt01 to compt12) in successive HLM runs. In addition, at this stage, a so-called ‘step-up’ approach is followed to examine which of the competency-related variables have a significant influence on student scores. Bryk and Raudenbush (1992) have recommended the step-up approach for inclusion of variables into the model to the alternative approach, referred to as ‘working-backward’ where all the possible predictors are dumped into the model and then the non-significant variables are progressively eliminated from the model.

The fourth step, which is the final step, involves building up the Level-2 slope models through adding the competency-related variables that have significant cross-level interaction effects using the Level-2 exploratory analysis sub-routine and the step-up strategy.

HLM results

For both Model-M and Model-S, the results of the HLM analysis described above provide reliability estimates at Level-1 of the model for each of the four assessment modes included in that model and the correlations between these assessment modes. The results also provide the estimations of the fixed effects for each variable in the equation, the estimations of the variance components and the deviance statistics of the models. These results are discussed in separate sub-sections below.

Reliability estimates

Table 9 displays the reliability estimates of the assessment modes involved in Models M and S at three stages in the development of the model. The three stages in Table 9 refer to (a) the unconditional model (Unconditional Stage) (b) the final model without cross-level interaction effects (Final Stage 1), and the final model with cross-level interaction effects (Final Stage 2).

For both Models M and S, the results in Table 9 indicate that all assessment modes have high reliability estimate (>0.80) regardless of the stage of the model that is considered. Thus, if the skill level of a student were to be measured based on the five assessment modes, equal degrees of confidence could be placed on the scores obtained using any of the five assessment modes.

Table 9. Reliability estimates at different stages of model development

	Unconditional Stage (with Level-1 variables only)	**Final Stage 1** (without interaction effects)	**Final Stage 2** (with interaction effects)
Model-M			
Scenario	0.825	0.825	0.828
Portfolio	0.868	0.869	0.810
Multiple Choice	0.939	0.939	0.939
Supervisor	0.916	0.915	0.874
Model-S			
Scenario	0.812	0.812	0.816
Portfolio	0.898	0.898	0.863
Supervisor	0.946	0.946	0.892
Self-Assessment	0.939	0.939	0.939

Fit of the model

Table 10 present results of deviance statistics and the chi-square tests carried out to compare model fit in progressive stages in the development of Model-M. In Table 10, the Null Stage (null model) is compared to the Unconditional Stage, Unconditional Stage is compared to the Final Stage 1 (i.e. final model without interaction effects), and the Final Stage 1 is compared with the Final Stage 2 (i.e. final model with interaction effects).

The information presented in Table 10 indicates better fit of the model at Final Stage 2 compared to all the other stages. Therefore, the inclusion of cross-level interactions between competencies and assessment modes improves the overall fit of the model. Importantly, the results in Table 10 appear to warrant the inclusion of the cross-level interaction effects because there is better fit of the model at Final Stage 2 compared to the fit of the model at Final Stage 1.

Table 10. Comparison of model fit in successive HLM runs for Model M

	Deviance Statistic	**Number of Parameters**	**Chi-square Statistic**	**Degrees of Freedom**	**p-value**
Null	9881.95	3			
Unconditional	8646.06	21	1235.90	18	0.00
Final Stage 1 (without interactions)	8602.81	27	43.24	6	0.00
Final Stage 2 (with interactions)	8579.46	31	23.36	4	0.00

The corresponding results for Model-S are basically similar to the results in Table 10, and therefore, it is concluded that the model at the final stage with interaction effects has a better fit to these data.

Fixed effects

Fixed effects estimated from the Unconditional Stage, Final Stage 1 and Final Stage 2 are presented together in Table 11 for Model-M and Table 12 for Model-S. Both the standardized as

well as the metric regression coefficients of the variables are presented in Tables 11 and 12. The metric regression coefficients are obtained from HLM runs using raw scores of the variables while the standardized regression coefficients are obtained from separate HLM runs using standardized scores of the variables. These results at the various levels of hierarchy are discussed next.

In Tables 11 and 12, for a given stage of the models, the metric intercept is an estimate of the overall (grand) mean score (skill level) of the students on the original outcome scale of 1 to 4.

On the other hand, the value of fixed effect for a particular assessment mode is an estimate of the score points (on the original scale) that should be added to (or subtracted from) the student score so as to adjust for the advantage (or disadvantage) associated with being assessed using that assessment mode. Similarly, the value of the fixed effect for a particular competency is an estimate of the score points that should be used to adjust the student score in order to cater for the advantage (or disadvantage) associated with being assessed for that competency.

Thus, from Tables 11 and 12, it can be observed that the grand mean score of the students is estimated to be 2.58 regardless of the model considered (Model-M or Model-S) and regardless of the stage of the model considered (unconditional, final without interaction effects or final with interaction effects). This grand mean score (2.58), when considered on the original scale of 1 to 4, means that the average score is 0.08 points above the average (2.50) of the original scale used to measure the students' skills.

From the results in Tables 11 and 12, it can be observed that, within the same model, the values of the fixed effects for the assessment modes remain unchanged regardless of the stage considered, which shows that inclusion of the competencies whose mean scores are significantly different from the grand mean and significant cross-level interactions in the analysis do not affect these values.

The following examples illustrate the impact of the model coefficients displayed in Tables 11 and 12 on student score.

If all other factors are equal and based on the Final Stage 2 of Model-M, a student of average skill level would be expected to get a score of 1.82 (that is, grand mean plus coefficient associated with the assessment mode, 2.58 + [-0.76]) if assessed using Scenario, and scores of 1.78, 2.06 and 2.97 if assessed using Portfolio, Multiple Choice and Supervisor respectively. The same student but based on Model-S would be expected to get scores of 2.34, 2.31, 3.10 and 3.49 if assessed using Scenario, Portfolio, Self-Assessment and Supervisor respectively. Although the scores based on Model-M do not match exactly those based on Model-S, they nevertheless follow a similar general pattern and seem to confirm what is found in the Rasch analysis. That is, the students are more likely to be rated as above average when assessed by their supervisors than when assessed using the other assessment modes.

For Scenario and Portfolio, the results in Tables 11 and 12 indicate that, regardless of the model considered, the mean of the scores obtained using Scenario follow closely the mean of the scores obtained using Portfolio, which is consistent with what is found using Rasch analysis. In addition, when interpreting the results in Table 12 (Final Stage 2), it should be noted that the estimated values of the fixed effect (-0.24 and -0.27 respectively) are not significantly different from zero at p=0.05. This means that, based on Model-S, the advantages (or disadvantages) associated with being assessed using Scenario or Portfolio are negligible. In other words, using either Scenario or Portfolio and based on Model-S, a student of average skill level would be expected to get a score roughly equal to the grand mean (2.58) that is predicted by the model.

Table 11. Fixed effects estimates at three stages in the development of Model-M

		Unconditional Stage (with Level-1 variables only)				**Final Stage 1** (without interaction effects)				**Final Stage 2** (with interaction effects)			
		Coefficient				Coefficient				Coefficient			
		Std'zed	Metric	S.E	P-value	Std'zed	Metric	S.E	P-value	Std'zed	Metric	S.E	P-value
Level-1	Intercept	2.58	2.58	0.06	0.00	2.58	2.58	0.01	0.00	2.58	2.58	0.01	0.00
	Scenario	-0.23	-0.76	0.11	0.00	-0.23	-0.76	0.11	0.00	-0.22	-0.76	0.11	0.00
	Portfolio	-0.24	-0.79	0.13	0.00	-0.24	-0.79	0.13	0.00	-0.24	-0.80	0.10	0.00
	interaction with Competency 2									0.05	0.59	0.19	0.01
	Multiple Choice	-0.19	-0.52	0.15	0.01	-0.19	-0.52	0.15	0.01	-0.19	-0.52	0.15	0.01
	Supervisor	0.18	0.39	0.11	0.01	0.17	0.39	0.11	0.01	0.17	0.39	0.08	0.00
	interaction with Competency 4									-0.11	-0.86	0.19	0.00
	interaction with Competency 9									-0.10	-0.78	0.19	0.00
	interaction with Competency 12									-0.07	-0.54	0.19	0.02
Level-2	Competency 1					0.04	0.15	0.05	0.03	0.04	0.14	0.05	0.04
	Competency 2												
	Competency 3												
	Competency 4					-0.07	-0.25	0.05	0.00	-0.07	-0.24	0.05	0.00
	Competency 5												
	Competency 6												
	Competency 7												
	Competency 8					-0.14	-0.47	0.05	0.00	-0.14	-0.48	0.05	0.00
	Competency 9					-0.06	-0.22	0.05	0.01	-0.06	-0.21	0.05	0.01
	Competency 10					-0.04	-0.14	0.05	0.03	-0.05	-0.15	0.05	0.03
	Competency 11												
	Competency 12					0.10	0.34	0.05	0.00	0.10	0.33	0.05	0.00

Notes: - The standard errors (SE) and p-values presented are those obtained using unstandardized (metric) variables.
- Self-Assessment is used as the fifth dummy for balancing the analysis.

Table 12. Fixed effects estimates at three stages in the development of Model-S

		Unconditional Stage (with Level-1 variables only)					**Final Stage 1** (without interaction effects)					**Final Stage 2** (with interaction effects)				
		Coefficient					Coefficient					Coefficient				
		Std'zed	Metric	S.E	P-value		Std'zed	Metric	S.E	P-value		Std'zed	Metric	S.E	P-value	
Level-1	Intercept	2.58	2.58	0.06	0.00		2.58	2.58	0.01	0.00		2.58	2.58	0.01	0.00	
	Scenario	-0.07	-0.24	0.11	0.04		-0.07	-0.24	0.12	0.07	ξ	-0.07	-0.24	0.12	0.07	ξ
	Portfolio	-0.08	-0.27	0.15	0.10	ξ	-0.08	-0.27	0.16	0.12	ξ	-0.08	-0.27	0.14	0.07	ξ
	interaction with Competency 2											0.05	0.60	0.19	0.01	
	Self-Assessment	0.25	0.52	0.15	0.01		0.25	0.52	0.15	0.01		0.25	0.52	0.15	0.01	
	Supervisor	0.41	0.91	0.16	0.00		0.41	0.91	0.17	0.00		0.41	0.91	0.12	0.00	
	interaction with Competency 4											-0.11	-0.87	0.23	0.01	
	interaction with Competency 9											-0.10	-0.77	0.22	0.01	
	interaction with Competency 12											-0.07	-0.53	0.23	0.04	
Level-2	Competency 1						0.04	0.15	0.05	0.03		0.04	0.15	0.05	0.03	
	Competency 2															
	Competency 3															
	Competency 4						-0.07	-0.25	0.05	0.00		-0.07	-0.25	0.05	0.00	
	Competency 5															
	Competency 6															
	Competency 7															
	Competency 8						-0.14	-0.47	0.05	0.00		-0.14	-0.48	0.05	0.00	
	Competency 9						-0.06	-0.22	0.05	0.01		-0.06	-0.21	0.05	0.01	
	Competency 10						-0.04	-0.14	0.05	0.03		-0.05	-0.16	0.05	0.03	
	Competency 11															
	Competency 12						0.10	0.34	0.05	0.00		0.10	0.33	0.05	0.00	

Notes: - The standard errors (SE) and p-values presented are those obtained using unstandardized (metric) variables.

ξ - Variable has no significant effect (p>0.05) but included in the model.

- Self-Assessment is used as the fifth dummy for balancing the analysis.

For the competencies, the results in Tables 11 and 12 show that, after controlling for the differences between the assessment modes, there are advantages associated with being assessed for Competencies 1 and 12, and there are disadvantages associated with being assessed for Competencies 4, 8, 9 and 10. In addition, the results in Tables 11 and 12 show significant interaction effects between Portfolio and Competency 2 and, between Supervisor and Competencies 4, 9 and 12. The interaction effect between Portfolio and Competency 2 mean that there are advantages of being assessed for Competency 2 using Portfolio. On the other hand the interaction effects between Supervisor and Competencies 4, 9 and 12 indicate that there are disadvantages in being assessed on these three competencies by the supervisor.

Despite what has been said above regarding the advantages and disadvantages of being assessed for some competencies, it should be noted that the standardized coefficients for the competencies have small values ($\leq|0.15|$). These small coefficients indicate that any advantages (or disadvantages) that may arise from being assessed for these competencies are very small.

Correlations between assessment modes

The first and the second panels of Table 13 show the correlations between the students' scores from the four assessment modes that are obtained following HLM analyses of the Final Stage 2 of Models M and S respectively.

Table 13. Correlations between assessment modes based on HLM final models

Model	Scenario	Portfolio	Multiple Choice	Supervisor
Model-MC				
Scenario	1.00			
Portfolio	0.95	1.00		
Multiple Choice	0.68	0.53	1.00	
Supervisor	0.57	0.68	0.65	1.00
Model-SA				
Scenario	1.00			
Portfolio	0.97	1.00		
Self-Assessment	0.73	0.78	1.00	
Supervisor	0.69	0.83	0.82	1.00

For both Model-M and Model-S, the results in Table 13 show strong to very strong correlations between the scores obtained using the different assessment modes. Thus, it appears that the ranking of students based on scores obtained using any one of the assessment modes do not differ markedly from the ranking obtained using scores from the other assessment modes. For Scenario and Portfolio, the correlation is near unity (≥ 0.95) regardless of the model considered, which suggests a high degree of agreement between the ranks obtained using these two assessment modes.

When interpreting the correlations presented in Table 13, it should be remembered that these correlations are computed taking into consideration the operational design of the study. In other words, these are the correlations between the assessment modes after the variability between the competencies has been controlled for.

Therefore, the results presented in Table 13 (based on GT approach and HLM analytical procedure) must be giving a better picture of the relationship between the assessment modes compared to the results obtained using the IRT approach (Table 6).

Estimation of variance explained

The percentages of variances available and explained based on Model-M follow closely those based on Model-S, and therefore, only the results for Model-M are presented and discussed in this section.

The results of the final estimation of variance components for Model-M at Final Stage 2 and the results of the analyses of the variance components obtained from the null models are presented in Table 14 in rows 'a' and 'b' respectively. From the information in Table 14 rows ‘a’ and ‘b’, the information presented in rows ‘c’ to ‘f' were calculated. A discussion of the calculations involved here is to be found in Raudenbush and Bryk (2002, p.69-95).

The results in Table 14 show that, 96.1 per cent and 3.9 per cent of the variance of student scores are at the Levels 1 and 2 respectively. These percentages of variance of student scores at the various levels of the hierarchy are the maximum amounts of variance available at those levels that could be explained in subsequent analyses. Thus, the results in Table 14 support what is found using Rasch analysis, that is, there are only small differences between the 12 competencies.

Table 14. Percentages of variance explained based on Model-MC

		Level-1 (N=3,960)	**Level-2** (N=12)	**Total**
a	Null Model	0.851	0.035	**0.886**
b	Final Model (with interaction effects)	0.593	0.000	
c	Variance Available	96.1%	3.9%	
d	Variance Explained	30.4%	98.7%	
e	Total Variance Explained	29.2%	3.9%	33.1%
f	Variance Left Unexplained	66.9%	0.0%	66.9%

In addition, the results in Table 14 show that the variables included in the final model explain 30.4 per cent of 96.1 per cent variance available at Level-1 and that is equal to 29.2 per cent (that is, 30.4 × 96.1) of the total variance explained at the Level-1. Similarly, the variables included in the final model explain all of the variance available at Level 2 (3.9 per cent). Thus, the total variance explained by the variables included in the final model is 29.2 + 3.9 = 33.1 per cent, which leaves 66.9 per cent of the total variance unexplained.

In summary, the results in Table 14 row 'f' indicate that the model developed in this study explains all the between-competencies (Level-2) variance but explains only a small amount of the within-competency (Level-1) variance. The large amount of variance left unexplained at Level-1 (66.9%) strongly indicates that there are other important Level-1 factors influencing the students’ scores that have not been included in the models developed in this study. Certain important Level-1 variables that are not available for examination in this study include student background characteristics (e.g. socio-economic status, gender, age and race) and supervisor background characteristics (e.g. academic qualification and professional experience). Therefore, there is a clear need for a further study to develop models that are the most appropriate for explaining students’ scores and which maximize the total variance explained at Level-1.

SUMMARY

In this study, data from 165 students from Massachusetts and Minnesota are used to examine the validity of five assessment modes (multiple choice test, scenario, portfolio, self-assessment and supervisor rating) in measuring competence in performance of 12 human service skills based on different data analytical theories.

It should be noted that the discussions in this article are based on preliminary results of rich and complex data that need further examination before drawing conclusions or making policy recommendations. Nevertheless, this article has shed some light on the general nature of the scores obtained using the five different assessment modes. Supervisors are evidently more generous in rating the skill levels of their students, compared with the alternative assessment modes, and this raises interesting questions which should form the basis for further analyses of these data. It is clearly premature to make recommendations for policy and practice from an initial and incomplete analysis of these data. Nevertheless, it is clear that classical test theory does not provide a meaningful analysis of the data, and that the use of item response theory in its simplest form, namely Rasch scaling, is inadequate to model fully the structure of the data and the manner in which the data were assembled, while generalizability theory would appear to provide a more adequate view. However, generalizability theory does not convert the data to an interval scale. The GT approach to the examination of the data clearly warrants further investigation, while it might be possible to extend the Rasch approach to take into account more adequately the design of the study.

It is of value to summarize the findings of the investigation reported in this article. The six research questions initially proposed in this article form a useful framework for providing a summary.

1. Can the five assessment modes be used to obtain reliable measures?

After allowance is made for the systematic differences between the five modes of assessment, as well as the systematic differences associated with the 12 competencies in a way that takes into consideration the design of the study, the resulting scores show strong levels of reliability ranging from 0.81 to 0.95.

2. Do the five assessment modes differ in their mean values and spread of scores?

Only after a preliminary examination of these data have differences in mean values and spread of scores been reported in this article, and these are given only for the Rasch approach. It is evident that the supervisor's ratings, and to a lesser extend the self-assessment ratings, are more lenient than the ratings obtained using the other three modes. Moreover, the self-assessment ratings show a smaller spread of scores than do the other four modes.

3. Do the 12 competencies differ in their mean values and spread of scores?

From the preliminary examination of the competency scores, the mean values of the scores are similar except for Competency 8 for which the scores are noticeably lower than for the other 11 competencies.

4. Can the data be effectively combined?

The evidence obtained from this investigation using IRT procedures indicates that with the exclusion of some assessments for particular modes on particular competencies a single scale might be employed. Further analysis is required to examine the strength of the five underlying dimensions associated with modes of assessment, and the 12 underlying dimensions associated with the competencies.

5. What are the correlations between (a) the five assessment modes, and (b) the 12 competencies?

After adjusting the scores using both IRT and GT procedures, the extent of correlation between the different pairs of scores indicates that there are noticeable differences between the different

modes of assessment and the different competencies that would appear to warrant their continued separation in the assessment of student performance.

6. Are there significant interactions between the assessment modes and the competencies?

A limited number of significant interactions were detected that warrant further examination. It should be noted that three out of the four significant interactions were associated with the supervisor mode of assessment and one interaction involved the portfolio mode of assessment.

Clearly there are many more questions that could be asked about the relationships between the models of assessment and the competencies for which answers might be expected to be provided by further analysis of this rich body of data. Such questions would have considerable practical significance for the assessment of competencies and performance skills using the different models of assessment available.

REFERENCE

Allal, L., and Cardinet, J. (1997). Generalizability Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp. 737-741). Oxford: Pergamon Press.

Allerup, P. (1997). Rasch Measurement Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp. 863-874). Oxford: Pergamon Press.

Andrich, D., Lyne, A., Sheridan, B., Luo, G. (2000). RUMM 2010 Rasch Unidimensional Measurement Models [Computer Software]. Perth: RUMM Laboratory.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Application and Data Analysis Methods*. Newbury Park: Sage Publication.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, 112 (1), 155-159.

Keats, J. A. (1997). Classical Test Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp. 713-19). Oxford: Pergamon Press.

Keeves, J. P. and Alagumalai, S. (1999). New Approaches to Measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 23-42). Oxford: Pergamon.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). California: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F. and Congdon, R.T. (2000). *HLM5: Hierarchical linear and Nonlinear Modeling* [Computer Software]. Lincolnwood, IL: Scientific Software International.

Stocking, M. L. (1997). Item Response Theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (2nd ed., pp.836-40). Oxford: Pergamon Press.

Wu, M. and Adams, R. (1998). CONQUEST [Computer Software]. Melbourne:ACER.