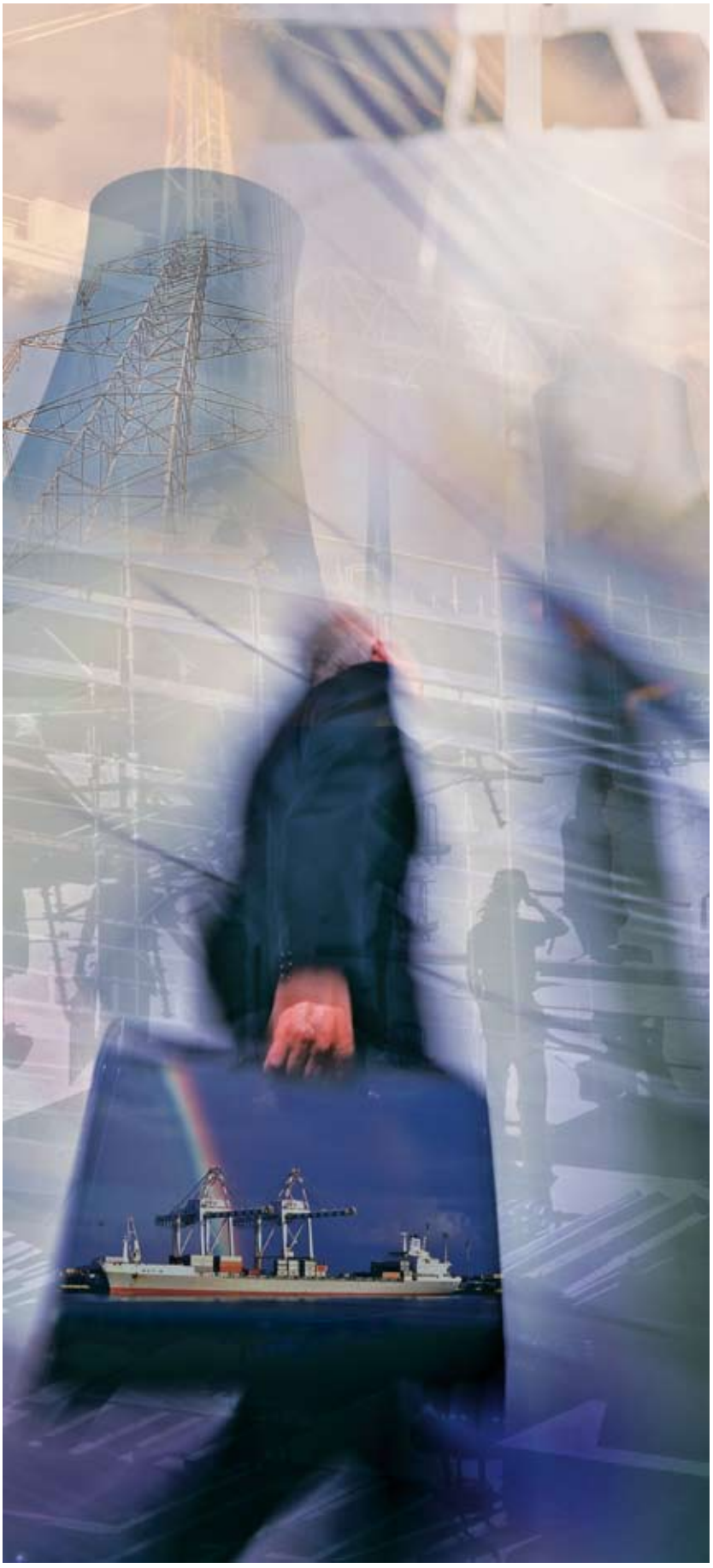


Queensland

The Economic Society of Australia Inc.

Proceedings of the 37th Australian Conference of Economists

Papers delivered at ACE 08

30th September to 4th October 2008

Gold Coast Queensland Australia

ISBN 978-0-9591806-4-0

Welcome

The Economic Society of Australia warmly welcomes you to the Gold Coast, Queensland, Australia for the 37th Australian Conference of Economists.

The Society was formed 83 years ago in 1925. At the time, the Society was opposed to declarations of policy and instead focused on open discussions and encouraging economic debate. Nothing has changed today, with the Society and the conference being at the forefront of encouraging debate.

This year we have a large number of papers dealing with Infrastructure, Central Banking and Trade.

Matters of the greatest global importance invariably boil down to be economic problems. Recent times have seen an explosion of infrastructure spending, after world-wide population growth has seen demand outpace aging supply. The world has become more globalised than at any time since World War I but the benefits of this (and the impact on our climate) has been questioned by some.

At the time of preparing for this conference we could not have known that it would have been held during the largest credit crisis since the Great Depression. The general public and politicians both look to central banks for the answers.

We are also very pleased to see a wide selection of papers ranging from applied economics to welfare economics. An A – Z of economics (well, almost).

Another feature of this conference is that we have gone out of our way to bring together economists from all walks of life, in particular from academia, government and the private sector. We are grateful to all of our sponsors, who are as diverse as the speakers.

The Organising Committee

James Dick
Khorshed Alam (Programme Chair)
Michael Knox
Greg Hall
Allan Layton
Rimu Nelson
Gudrun Meyer-Boehm
Jay Bandaralage
Paula Knight

Published November 2008

GPO Box 1170
Brisbane Queensland Australia
ecosocqld@optushome.com.au

Our Gold Sponsors

QIC

Queensland Treasury Corporation

Reserve Bank of Australia

Queensland Government
Department of Infrastructure and Planning

Keynote Sponsors

Australian Government
Productivity Commission

Australian Government
The Treasury

Queensland University of Technology
Faculty of Business
School of Economics and Finance

The University Of Queensland
AUSTRALIA

Goldman Sachs JBWere

The University Of Queensland
AUSTRALIA

NCER National Centre for Econometric Research

Central Council

Special Session Sponsors

Griffith UNIVERSITY

The Higher Education Academy Economics Network

Unless we have specifically been requested to do otherwise, all the papers presented at the conference are published in the proceedings in full. A small number of papers will have versions that have also been made available for special editions of Journals, Economic Analysis and Policy, and the Economic Record. Authors will retain the right to seek additional publication for papers presented at the conference so long as it differs in some meaningful way from those published here.

The opinions expressed in the papers included in the proceedings are those of the author(s) and no responsibility can be accepted by the Economic Society of Australia Inc, Economic Society of Australia (Queensland) Inc, the publisher for any damages resulting from usage or dissemination of this work.

The Paper following forms part of - *Proceedings of the 37th Australian Conference of Economists*
ISBN 978-0-9591806-4-0

Perceptions of Assessment Demands in Economics – Implications for Economics Education Research

Tommy Tang & Tim Robinson
Queensland University of Technology, Australia

Abstract: *Educational research based on the Student Experience of Learning (SEL) framework (Marton, 1988) argues that while student personological factors have a general influence in determining how the student goes about a learning task, their learning motivations and strategies are significantly influenced by their perceptions of the learning environment and the demands of the learning task. This paper reports an empirical study to develop an instrument to measure students' perceptions of the cognitive abilities required by an assessment in economics. Using the instrument, two groups of introductory and one group of intermediate economics students at a university in a capital city in Australia were surveyed. Based on data collected by using the instrument, students' perceptions of three assessment types - multiple choice question examination, essay assignment and essay examination were analysed both at the group and individual levels. Results show both consistency within a student and variations across assessments in their perceptions of assessment demands. The implications of these results will be discussed in this paper.*

Keywords*: assessment, economics education, perceptions, student learning*
JEL code: A22

1. Introduction

There are conflicting views in economics education as to what kind of cognitive abilities different examination types measure. Walstad (2001) listed a number of examination types for first year economics which were further re-grouped into two categories: construct-response or fixed-response. Without referring to research evidence, he claimed that fixed-response questions, for example multiple choice questions (MCQ) and T/F items, measure low level cognitive skills whereas essays and construct-response questions (such as short answer questions) are more able to tap higher abilities.

However, Wainer and Thissen (1993) in their review of the reliability and correlation between multiple-choice and construct-response tests in economics, found very high correlations between multiple-choice and essay scores which suggests that there is little or no difference in the cognitive skills these two assessment types measure. Becker and Johnston (1999), casting

doubt in the validity of Wainer and Thissen's findings, carried out a more detailed investigation. Using a two-stage least-squares regression, they found that multiple-choice and essay response questions in the VCE (Victoria Certificate Examination) economics papers measured different dimensions of student knowledge.

While it is important to determine the cognitive abilities measured by an assessment piece, in investigating student learning it may be more important to look at students' perceptions regarding assessment. This is the position taken in the Student Experience of Learning (SEL) literature (Marton, Hounsell, & Entwistle, 1984). There is now a large amount of educational literature (Campbell et al., 2001; Marton & Svensson, 1982; Ramsden, 1987; Scouller, 1998; Scouller & Chapman, 1999) which shows that how students learn is greatly influenced by what they perceive is being assessed in a graded task. For example, Scouller (Scouller, 1998), in a study of students' perceptions about assessments and their choice of learning approaches in first year psychology, found that the MCQ examination is perceived by students as assessing low level intellectual skills (factual recall and simple application) and that students are more likely to employ a surface learning approach than in the essay assignment context which is perceived to assess higher order ability (viz. analytical and research skills).

The study reported in this paper is part of a larger research program which investigates the learning process in economics. The objective of the research program is to investigate what students perceive as being required in three common, but different, assessment types in economics (MCQ exam, essay assignment and exam essay), and how their perceptions influence their learning approaches in different assessment contexts. This paper will firstly document the development of an instrument for measuring students' perceptions of the cognitive skills being assessed in assessments in economics. It will go on to report a survey using the instrument to collect data on students' perceptions about three common types of assessment in introductory and intermediate economics courses – MCQ examination, essay examination and essay assignment. The implications of findings from the survey will be discussed.

2. The instrument

Scouller (1998) developed a 12-item Assessment Questionnaire to measure student's perceptions of intellectual processing required by different types of assessment in a psychology class. According to Scouller, half of the 12 items describe an assessment as

assessing a low level of intellectual skills, and the other half a high level of intellectual skills. In the present study, the instrument follows the same conceptual framework as Scouller. It consists of 13 items; seven of them are adapted from Scouller's Assessment Questionnaire and the remaining 6 items are based on ideas provided by economics students in focus groups and written surveys conducted in 2003. Appendix 1 provides the sources and labels for each item.

2.1 Method

The 13-item instrument (called Perceptions of Assessment Demands) was administered to three groups of economics students in Economics 1, Economics 2 and Intermediate Macroeconomics in business degrees in a capital city university in Australia. Economics 1 and 2 are introductory economics units. The course is organised in such a way that both microeconomics and macroeconomics are taught in Economics 1, and both are dealt with in greater depth in Economics 2. The surveys were conducted in the last (revision) week in semester 1 2004. By that time, students had already done their mid-semester MCQ examination and essay assignments, but not their final examination. In the survey they were asked to describe the cognitive ability required in the MCQ examination and essay assignment, and to anticipate the cognitive ability to be assessed in the essay questions in the final exam, the nature of which had already been revealed to them in the unit document and sample/past exam papers. They responded to each item on a 5-point scale ('1' indicating strong disagreement and '5' strong agreement with the statement) in relation to each of the three assessment regimes.

2.2 Result

The data were factor analysed for each assessment, using Principal Axis method of extraction and Varimax rotation[1] with Kaiser Normalisation to achieve simple structure. For cases with missing data, the method of listwise deletion was used. The number of valid responses was 648 consisting of 453 Economics 1, 114 Economics 2 and 81 Intermediate Macroeconomics students. Referring to Table 1, the Kaiser-Meyer-Olkin (KMO) indices of the instrument in the three assessments of 0.747 to 0.805 indicate the data are suitable to very suitable for factor analysis (Graetz, 2003).

[1] Oblique rotation was also conducted. In the oblique model, as expected the factor correlations were found to be very small and the two factorial structures obtained by the two methods of rotation were very similar.

Table 1 Scales commonality and reliability (n=647)

	MCQ	ASS	Exam
KMO index	0.747	0.793	0.805
Cronbach alpha	0.643	0.697	0.693
Cronbach alpha (minus du1)	0.686	0.718	0.728

The Cronbach alpha values of 0.643 to 0.697 are above 0.6, which shows the instrument has sufficient internal consistency in all three assessment contexts. Based on theoretical consideration, a two-factor solution was imposed for each assessment context. The factor matrices are presented in Table 2.

Table 2 Factor structures of perceptions of assessment demands (n=647)

	Factor	
	1	2
rw2_mcq	.709	
crit_mcq	.603	
du2_mcq	.560	
intg_mcq	.528	
rhi_mcq	.483	
rw1_mcq	.461	
fdbk_mcq	.387	
rlow_mcq	.333	.311
mem3_mcq		.792
mem1_mcq		.685
mem2_mcq		.570
mem4_mcq	.231	.306
du1_mcq		.222

MCQ Exam

	Factor	
	1	2
rw2_ass	.756	
du2_ass	.649	
intg_ass	.576	
crit_ass	.568	
rw1_ass	.543	
fdbk_ass	.525	
rhi_ass	.501	
rlow_ass	.379	.341
mem3_ass		.759
mem1_ass		.641
mem2_ass		.515
mem4_ass	.261	.357
du1_ass		

Essay Assignment

	Factor	
	1	2
rw2_ex	.657	
du2_ex	.586	
crit_ex	.545	
rw1_ex	.513	
intg_ex	.497	
fdbk_ex	.494	
rhi_ex	.417	.322
mem3_ex		.711
mem2_ex		.706
mem1_ex		.587
rlow_ex	.288	.548
mem4_ex	.236	.347
du1_ex		

Essay Exam

(loading < 0.2 are not shown)

The two factor solutions extracted have almost identical factorial structures for all three assessment types. It is evident from the pattern matrices in Table 2 that Factor 1 represents a perception of high level of intellectual processing, and Factor 2 low level of intellectual processing. Factors 1 and 2 are hence labelled as HIGH and LOW, respectively.

There are several important observations concerning the factor loadings. First, it is noted that the item *du1* (reproduced below) does not have loadings greater than 0.2 on either factor for all three assessment types.

(*du1*) This assessment is one that I can do well even if I don't have a deep understanding of the content.

There is thus empirical ground to discard this item from the scale. The internal consistency of the construct (Cronbach alpha) for all three assessments improve substantially when *du1* is deleted; all alpha values are close to or above 0.7 (Table 1).

Second, two items (*mem4* and *rlow*, reproduced below) were found to have cross loadings for all three assessment types.

(*mem 4*) This assessment measures how much you have listened and remembered in classes.
(*rlow*) This assessment assesses your ability to reproduce factual details and knowledge.

Item *mem4* describes an assessment that tests how much you memorise from the lessons, and *rlow* describes an assessment concerned mainly with recall of low level knowledge. Both items were expected to have high positive loadings on the LOW factor, and zero or negative loadings on the HIGH factor. The positive loadings of these two items on the HIGH factor suggests that while students perceive the assessment as assessing high level intellectual skills, they also tend to believe it involves memorisation and recall of basic knowledge. The existence of cross-loading in a measurement model is, from the point of view of instrument construction, not a desirable outcome. However, given the substantive meaning of the cross loadings (to be discussed), these two items were retained for the next stage of analysis.

A third observation concerns the loading of the item *rhi*. This item describes a perception that the assessment tests students' ability to reproduce a higher level of knowledge.

(*rhi*) The assessment assesses your ability to reproduce ideas and viewpoints presented in readings and lectures.

This item was adapted from Scouller's Assessment Questionnaire (1998). Scouller treated this item as a LOW item, but no empirical justification (for example, factorial structure) was provided in her 1998 paper. In the present study, the item *rhi* has fairly large loadings (>0.4) on the HIGH factor for *all* assessment types and a small loading of 0.322 on the LOW factor *only* in essay examination. A more thorough discussion of this finding and its implications will be provided in Section 3 of this paper.

2.3 Model Calibration and Validation

The next stage of instrument development involves further refinement of the model. This stage of model development consists of two independent steps: (1) model calibration, and (2)

model validation (Byrne, 2001; Kline, 2005). The purpose of model calibration is to determine if a more parsimonious model can be obtained by deleting paths (model trimming) or to see if model fit can be improved by adding paths. This process of model re-specification should be based upon both empirical (for example, based on modification indices generated by AMOS) and theoretical evidence. With regard to theoretical justification for post hoc model modification, Kelloway (1998) pointed out its danger in that SEM researchers had no trouble in giving a legitimate substantive reason for modifying a model in order to achieve a better model fit. He quoted Steiger: "What percentage of researchers would find themselves unable to think up a 'theoretical justification' for freeing a parameter? In the absence of empirical information to the contrary, I assume that the answer … is 'near zero'." (quoted in Kelloway 1998, p. 21) Therefore, while the modification should be substantively meaningful, theoretical justification by itself is insufficient. It is imperative that the modified model must be replicated (or validated) in another sample. Hence, the modified model must be validated with an independent sample (see also Kline, 2005 and Tomarken and Walker, 2005 for same recommendations).

2.3.1 Method

The sample of 647 was split randomly into two roughly equal halves, one for the purpose of calibration and the other for validation.[2] This two-step procedure will minimise the danger of obtaining a well-fitting model that capitalises on chance variation of the specific sample (Chin, 1998). The procedure of calibration and validation conducted for each of the three assessments and detailed results are presented in Appendix 2. A summary of the results is provided below.

2.3.2 Result

The final calibrated model for all three assessment contexts is depicted in Figure 1 below. In the final model, the item MEM4 was deleted for empirical and substantive reasons. The cross loadings on the item RLOW was kept, as it significantly increases the model fit indices and is theoretically meaningful. These empirical and substantive considerations are discussed in Appendix 2. Notice also that covariances between two pairs of error terms (e5/e10 and e11/e12) were included in the final model as freely estimated parameters. The covariance

[2] Ideally the sample for EFA should not be used again for model re-specification as the CFA will very likely confirm the model derived from the EFA. The re-use of the EFA sample for model re-specification here is justified on two grounds. First, the purpose is not to confirm the model per se but for refining the measurement model for later model testing. Second, any modification made based on the calibration sample would be tested again on the validation sample.

between e5 and e10 is justified on the ground that items *rhi* and *rlow* both concern a perception of the assessment as dealing with recall of information.

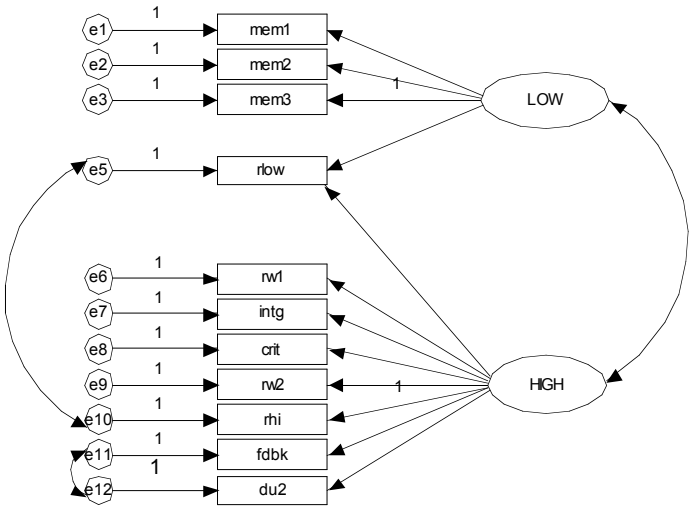


Figure 1 Final model – Perceptions of Assessment Demands

Regarding the covariance between e11 and e12, scrutiny of the corresponding items *fdbk* and *du2* reveals both describe the perception that the assessment is a means to developing deep understanding of the topics. Therefore, the inclusion of the two covariances was empirically (based on modification indices) and conceptually justified.

The goodness of fit indices for the final calibration model for the three assessments are shown in Table 3.

Table 3 Selected goodness of fit statistics – Calibration and Validation Models

	Model	n	χ^2	df	χ^2/df	SRMR	CFI	GFI	RMSEA
Essay assignment	Final Calibration model	328	64.924	40	1.623	0.0438	0.961	0.968	0.044
	Validation model	319	84.410	40	2.110	0.0604	0.952	0.957	0.059
MCQ exam	Final Calibration model	264	56.367	40	1.409	0.0475	0.969	0.966	0.039
	Validation model	302	66.123	40	1.653	0.0487	0.962	0.964	0.047
Exam essay	Final Calibration model	305	72.260	40	1.806	0.0466	0.957	0.961	0.052
	Validation model	342	102.475	40	2.562	0.0729	0.920	0.948	0.068

The following interpretation of these indices is recommended by experienced SEM researchers (such as Byrne, 2001; Kelloway, 1998; Kline, 2005):

	Excellent fit	Reasonable fit
Normed Chi-square (χ^2/df)	1 - 2	2 – 3
Comparative Fit Index (CFI)	> 0.95 (but < 1)	> 0.9
Goodness of Fit Index (GFI)	> 0.95 (but < 1)	> 0.9
Standardised Root Mean Square Residual (SRMR)	< 0.05	0.05 – 0.1
Root Mean-Square Error of Approximation (RMSEA)	< 0.05	0.05 – 0.08

It is evident from Table 3 that all selected indices except one (RMSEA for exam essay) are in the excellent model fit range. The RMSEA value of 0.052 for exam essay is just above the 0.05 limit for excellent model fit.

The final model was subsequently tested on the validation sample for each assessment. The model fit indices (Table 3) are not as good as those for the calibration samples, but are all in the good to excellent ranges. The results of the validation tests provide empirical evidence for the validity of the Perceptions of Assessment Demands construct developed in this study. Here, it must also be pointed out that the above model calibration procedures for the three assessments resulted in the three final models with the same factorial structure and the same added residual covariances. Moreover, they were independently obtained based upon empirical evidence. In a sense the final model presented in Figure 1 was 'validated' three times independently, lending further support for its validity and robustness.

Table 4 Parameter Estimates – Perceptions of Assessment Demands

			Essay Assignment (n=647)		MCQ exam (n=566)		Exam essay (n=647)	
			Std est	p	Std est	p	Std est	p
rlow	<---	LOW	.300	.000	.295	.000	.421	.000
mem3	<---	LOW	.786		.783		.706	
mem2	<---	LOW	.529	.000	.624	.000	.728	.000
mem1	<---	LOW	.639	.000	.701	.000	.637	.000
rhi	<---	HIGH	.473		.459		.438	
rw2	<---	HIGH	.797	.000	.773	.000	.752	.000
crit	<---	HIGH	.575	.000	.583	.000	.574	.000
intg	<---	HIGH	.588	.000	.540	.000	.484	.000
rw1	<---	HIGH	.583	.000	.465	.000	.532	.000
fdbk	<---	HIGH	.451	.000	.237	.000	.387	.000
du2	<---	HIGH	.604	.000	.468	.000	.491	.000
rlow	<---	HIGH	.405	.000	.312	.000	.348	.000

.000 denotes p < 0.001

2.3.3 Parameter Estimates

The final model (Figure 1) in each assessment was run on the Week 13 full datasets[3] (n=647 for essay assignment and exam essay; n=566 for MCQ exam) to obtain the factor loadings. Table 4 presents the factor loadings for all three assessments. Review of the factor loadings shows that all estimates (except for the path from HIGH to fdbk for the MCQ context) are close to or above the level of practical significance of 0.3 (Meyers, Gamst, & Guarino, 2006), with the majority above 0.5. This provides further empirical support for the validity of the model.

3. Discussion of the Factorial Structure

The perceptions of assessment demands construct was empirically tested and confirmed and was found to have the same factorial structure for all three assessment types considered in this study program.

The factorial structure highlights the ambiguous but important role of reproductive learning in the context of learning for assessment as perceived by introductory and intermediate economics students. Contrary to earlier research findings (Scouller, 1998), the ability to reproduce is not perceived by students as a unitary concept. The empirical results show a clear distinction exists between the ability to reproduce basic information (*rlow*) and the ability to reproduce sophisticated ideas and viewpoints (*rhi*). The second type of reproduction (as measured by the item *rhi*) was found in this study to be associated with the HIGH perception in all three assessment contexts, which is inconsistent with Scouller's hypothesised factorial structure.

Two explanations are proposed. The first explanation is related to the two types of understanding proposed by Tang and Bain (1994). During their course of study, students are often presented with sophisticated ideas, techniques and viewpoints in class or text, with structures and explanations being provided by the lecturer or author of the text. The ability to reproduce these structures and explanations is known as 'reproductive understanding', which is distinguished from 'transformative understanding'. The latter is close to the constructivist view of knowledge, in that knowledge is personally constructed by making connections of the materials with personal experiences, so that they can analytically and critically apply the materials to novel situations, whereas for reproductive understanding, applications occur only

[3] Data imputation was performed on this dataset and therefore it has no missing value.

in familiar situations. It is speculated that when thinking about an assessment in introductory and intermediate level economics, students tend to confound these two forms of (reproductive and transformative) understanding. In other words, when students perceive that it is the reproduction of these ideas, techniques and viewpoints that is being assessed in any assessment regimes, they tend to associate this ability with other higher order skills, such as critical analysis and evaluation. Hence, the high loading of this item (*rhi*) on the HIGH factor was observed.

The second explanation concerns the importance of effective recall in exam situations. In learning for exam type assessment, some students would believe that understanding is important and that the assessment is testing their ability to critically analyse and to develop personal ideas and viewpoints – transformative understanding. Nonetheless they would also see the importance and need for reproductive learning in an examination condition. This is because in a tense, invigilated environment, and with a time constraint, the ability to memorise and effectively reproduce sophisticated knowledge obtained in their revision is crucial. This is an entirely reasonable position and explains the association of the item *rhi* with the HIGH factor in exam type assessments. Therefore, the loading of *rhi* on the HIGH factor is substantively meaningful.

In the EFA conducted in this study the cross-loadings of the item *rlow* were observed, which means the variance of this item is significantly accounted for by *both* the HIGH and LOW factors. That this item loads on the LOW factor is self-explanatory. However, the association of this item with the HIGH factor in all three assessment contexts is unexpected, but can be explained by co-existence of high and low level questions in introductory and intermediate economics assessments within each of the three assessments. The empirical evidence regarding the cross loading of *rlow* implies that while an economics assessment is perceived as assessing high level skills, a significant proportion of its content involves basic reproduction of 'factual details and knowledge', which explains the loading of *rlow* on the HIGH factor.

4. Assessment Demands in Different Assessments

In order to compare students' perceptions of assessment demands within and across assessment contexts, simple composite scores were computed for the two sub-scales – HIGH and LOW perceptions. While the cross-loading of *rlow* provides insights into the structure of

students' perceptions of assessment demands as discussed in Section 2.2, from a measurement perspective the cross-loading is problematic. Therefore, this item (*rlow*) was omitted in the computation of the composite scores. The composite scores are interpreted in the following way. A large score on the LOW sub-scale means that students agree the assessment requires low cognitive skills. Similarly, a large score on the HIGH sub-scale indicates the assessment is perceived as assessing high level cognitive skills. Students' perceptions of assessment demands within and across assessment types were compared and contrasted.

4.1 Within-assessment Comparisons

First, comparison of the LOW and HIGH scores is made *within* each assessment type. Refer to Table 5. For MCQ exam, as expected, it has a large LOW score and a small HIGH score indicating that MCQ exam is perceived as assessing more low level than high level cognitive skills, and the difference of the two scores is significant at 0.000 (t = 18.985). The opposite is true for essay assignment. Essay assignment is perceived as assessing more high level than low level cognitive skills (t = -31.805, p = 0.000). The difference in mean score of close to 1.3 on a five-point scale is very large.

Table 5 Comparisons of perceptions within an assessment (paired t-tests)

		Mean	n	Std. dev.	t (p-value)
Essay Assignment	LOW	2.5993	647	.81216	-31.805 (.000)
	HIGH	3.8923	647	.60110	
MCQ Examination	LOW	3.9018	566	.79295	18.985 (.000)
	HIGH	3.0590	566	.64089	
Essay Examination	LOW	3.7359	647	.77354	4.009 (.000)
	HIGH	3.5899	647	.54892	

Interestingly, the exam essay has *both* a large LOW and a large HIGH score, with both scores being greater than 3.5. This implies that students believe that this assessment type requires both understanding *and* memorising. Moreover, it is observed that the LOW score (3.736) is slightly greater than the HIGH score (3.590). This shows that in students' eyes, exam essay questions assess more low level than high level skills. Note also that although the difference between the HIGH and LOW scores for exam essay (0.154) is statistically significant, it is a lot smaller in comparison to those of the other two assessment types.

4.2 Between-assessment Comparisons

In order to compare students' perceptions across assessment types, one-way analysis of variance (ANVOA) was conducted The result presented in Table 6 shows that there is significant difference among the three assessment types on both the LOW and HIGH perceptions [F(2, 315.093) = 501.006, p = 0.000; F(2, 103.386) = 299.268, p = 0.000, respectively].

Table 6 One-way ANOVA – between assessment comparisons of assessment demands

		Sum of Squares	df	Mean Square	F	p
LOW	Between Groups	630.186	2	315.093	501.006	.000
	Within Groups	1167.905	1857	.629		
	Total	1798.091	1859			
HIGH	Between Groups	212.771	2	106.386	299.268	.000
	Within Groups	660.137	1857	.355		
	Total	872.908	1859			

Post hoc multiple comparisons were performed to determine which assessments have different means on each perception sub-scale. The Bonferroni adjusted procedure was used for that purpose to avoid the inflation of probability of detecting significant difference by chance due to multiple comparisons. Table 7 reveals significant mean differences between all three assessments on both sub-scales.

Table 7 Post hoc Bonferroni Multiple Comparisons

Perception	(I) Assessment	(J) Assessment	Mean Difference (I-J)	Std. Error	p
LOW	essay assignment	MCQ	-1.30253	.04564	.000
	essay assignment	exam essay	-1.13658	.04409	.000
	MCQ	exam essay	.16595	.04564	.001
HIGH	essay assignment	MCQ	.83328	.03431	.000
	essay assignment	exam essay	.30242	.03315	.000
	MCQ	exam essay	-.53085	.03431	.000

First, essay assignment is perceived as assessing less LOW level skills and more HIGH level skills than exam type assessments (that is, MCQ and exam essay). The differences on LOW scores (essay assignment – MCQ = -1.303; essay assignment – essay exam = -1.137) are large and highly statistically significant at the level of 0.000. The differences on HIGH scores are

not as big (especially when compared to exam essay) but also statistically significant at p = 0.000 (essay assignment – MCQ = 0.833; essay assignment – essay exam = 0.302). Therefore, based on students' perceptions, it is anticipated that essay assignment is most likely to elicit deep learning, and least likely to elicit surface learning.

Second, comparing the two types of exams, MCQ exam has a slightly higher score on the LOW sub-scale (diff = 0.166, p = 0.001), and a much lower score on HIGH sub-scale (diff = -0.833, p = 0.000) than exam essay. This empirical observation implies that, according to students' perceptions, both of the exam type assessments assess a lot of LOW level cognitive ability and to a similar extent, but exam essay assesses much more HIGH level cognitive ability than MCQ exam.

4.3 Summary for this section

According to students' perceptions, essay type assessment requires more deep understanding of content compared with fixed-response assessment, whereas exam type assessment assesses more surface understanding than non-exam type assessment. The above findings also show that the HIGH and LOW perceptions vary indirectly in MCQ exam (large LOW score, small HIGH score) and essay assignment (small LOW score, large HIGH score). Take MCQ exam as an example, students tend to agree that it requires a lot of rote-memorising and reproduction of knowledge without the need to deeply understand the content. However, the perceptions regarding exam essay are somewhat different. Students believe it is assessing *both* high and low level cognitive abilities (both HIGH and LOW scores being greater than 3.5).

5. Variation and Consistency of Perceptions of Assessment Demands

The last section investigates variations of students' perceptions of assessment demands at the group level. To further analyse students' perceptions of assessment demands, distinction is made between variations of perceptions at the individual level and variations at the group level. Other studies (Campbell et al, 2001; Sambell and McDowell, 1998) found variations *between* students in terms of their perceptions within the *same* learning environment. These are variations at the individual level. In this study, to investigate between-student variations (i.e., variations at the individual level), student's LOW and HIGH scores for each assessment type and their correlations are computed. The results of correlation analysis are presented in Table 8.

	LOW_mcq	LOW_ex	LOW_ass
LOW_mcq	1 (578)		
LOW_ex	**.472**** (560)	1 (613)	
LOW_ass	**.105*** (576)	**.178**** (613)	1 (641)

	HIGH_mcq	HIGH_ex	HIGH_ass
HIGH_mcq	1 (565)		
HIGH_ex	**.329**** (543)	1 (600)	
HIGH_ass	**.093*** (562)	**.499**** (599)	1 (630)

** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).
Legend: mcq = MCQ examination; ex = essay examination; ass = essay assignment

Table 8 Pearson correlations (sample size in brackets)

Referring to the figures in the table on the left, the results show small to large effect size correlations of the LOW scores among the three assessment types. The correlations are significant at the level of 0.05 or above. This means when a student perceives an assessment as assessing low level skills, he also tends to see the other two assessments as assessing low level skills in comparison to other students in the sample. There is a similar observation in relation to the HIGH scores. This suggests consistency in students' perceptions of assessment demands.

This consistency of perceptions within the individual is in contrast, but not in contradiction, to the variations of perceptions across assessment types at the group level presented earlier. The between-assessment variability at the group level is due to the differences in content and format of these assessment types. On the other hand, the within-individual consistency can be explained by the existence of factors in the individual student that shapes their perceptions of assessment demands in a consistent way.

5.1 Implications for economics education research

The last observation has two implications for modelling learning in economics. The first is self-evident – the model should not ignore student input factors that potentially explain individual differences in their perceptions and learning of economics. The second is related to the first – what student factors should be chosen? As discussed earlier, with very few exceptions, in previous research studies in economics education conducted so far the search for explanatory variables has been confined to general variables such as student's aptitude, age, gender, maths background, language proficiency and socioeconomic background (Becker, 1997). In two independent review studies (Becker, 1997b; Shanahan et al., 1997) inconsistent results were found regarding the effects of these student input variables on examination performance (with aptitude being the only exception). For example, there is no consistent

evidence to support the notion that the performance of one gender is better than the other (Shanahan et al, 1997). On the effect of prior economic knowledge, the research findings range from a significant positive impact, slight positive impact, to no effect at all. In a study by McCosker (2000), the effect of prior economic knowledge on performance in introductory university economics was found to be negative.

These inconsistent research findings suggest that between-student variations may not be fully accounted for by these *general* student input factors. It is commonly accepted that vital differences exist in the methods of creation and acquisition of knowledge in different disciplines (Kolb, 1984). Recognising these disciplinary differences and drawing on work by Eley and Meyer (2000), and Meyer and Cleary (1998) (cited in Meyer and Shanahan 2002), Meyer and Shanahan (2002) argued we should 'seek additional sources of variation that are perhaps conceptually unique" within economics (p. 204). Their research effort prompted the authors (Tang & Robinson, 2004) to develop an instrument to measure students' misconceptions about economics and naïve economic thinking in order to investigate how these discipline specific factors might relate to students' learning approaches and their academic performance.

6. Conclusion

In this paper, the instrument developed to measure students' perceptions of assessment demands was found to have good to excellent psychometric properties. Examination of the factorial structure reveals several important observations. Recall of knowledge such as 'ideas and viewpoints' as presented by the lecturer and in text is regarded as a *high* level intellectual skill from the students' perspective. Moreover, the cross-loading of the item RLOW suggests that an assessment perceived as dealing with higher level skills, often also includes items that assess low level intellectual skills.

In this study, variations of perceptions across assessments types at the group level were demonstrated and the patterns of variations were found to be largely consistent with earlier studies. Our analysis at the individual level shows there is consistency of perceptions within an individual. Since perceptions of assessment impact on learning approaches, consistency of perceptions would imply a certain degree of consistency of learning approach within an individual student. This paper also argues that besides general student input variables, we

should include discipline-specific variables such as students' economic beliefs and thinking in modelling learning in economics. To extend this argument, it can be speculated that generalised learning inventories such as Bigg's SPQ (1987) and Ramsden's ASI (1987) is unlikely to fully capture important variations in learning activities that are specific to the discourse of a discipline. Therefore, to model learning in a discipline, a discipline-specific learning inventory is recommended.

According to the SEL framework, students' perception of the learning context has great influence on their approaches. Given students' perceptions of assessment demands in this study, it is anticipated that students will utilise more deep and less surface learning activities in essay assignment than in MCQ exam. With regard to exam essay, since it is perceived as assessing both high and low level cognitive abilities, it is unclear how such perceptions influence students' learning approaches for this assessment type.

Moreover, if students' perceptions of assessment demands truly reflect what is actually being assessed in each of the three assessment types, there will be important implications concerning the associations of learning approaches and academic performance in different assessment contexts. For example, it would imply that if essay assignment requires deep understanding of the content, then there should be positive association between deep learning approach and essay assignment mark. On the other hand the deep learning approach will have low or no association with MCQ mark if students' perceptions of assessment demand in MCQ exam are correct. Therefore, the development of the Perceptions of Assessment Demands instrument represents an important step towards investigating the so-called "black box" in economics education that is the process of learning in economics (Shanahan et al., 1997). The next stage of the research program is to investigate students' approaches to learning for assessments and the causal relationships between perceptions of assessment demands and learning approaches for assessments, and the associations of learning approaches and academic achievement in different assessment contexts.

References

Becker, W. E. (1997b). Teaching Economics to Undergraduates. *Journal of Economic Literature, 35*(3), 1347-1373.

Becker, W. E., & Johnston, C. (1999). The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding. *The Economic Record, 75*, 348-357.

Biggs, J. B. (1987). *Study Process Questionnaire Manual. Student Approaches to Learning and Studying*: Australian Council for Educational Research, Hawthorn.

Byrne, B. M. (2001). *Structural Equation Modelling with AMOS*. New Jersey: Lawrence Erlbaum Associates Inc., Publishers.

Campbell, J., Smith, D., Boulton-Lewis, G., Brownlee, J., Burnett, P. C., Carrington, S., et al. (2001). Students' Perceptions of Teaching and Learning: The Influence of Students' Approaches to Learning and Teachers' Approaches to Teaching. *Teachers and Teaching: Theory and Practice, 7*(2), 173- 187.

Chin, W. W. (1998, Feb 20, 1998). *Issues and Opinion on Structural Equation Modeling*. Retrieved 1/06, 2006, from http://www.misq.org/archivist/vol/no22/issue1/vol22n1comntry.html

Entwistle, N. J., Marton, F., & Entwistle, A. (1993). *'Knowledge Object' Constituted through Intensive Academic Study.* Paper presented at the Paper presented at the EARLI Conference at Aix, 1993 - Symposium on Awareness.

Graetz, B. (2003). *Principal Components and Factor Analysis*: Course notes for ACSPRI Winter Program 2003.

Kelloway, E. K. (1998). *Using LISREL for Structural Equation Modeling: A Researcher's Guide*: Thousand Oaks: Sage Publications.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*: The Guilford Press.

Kolb, D. A. (1984). *Experiential Learning*: Prentice-Hall, N.J.

Marton, F. (1988). Describing and Improving Learning. In R. R. Schmeck (Ed.), *Learning Styles and Learning Strategies*. New York: Plenum Press.

Marton, F., Hounsell, D., & Entwistle, N. (1984). *The Experience of Learning*. Edinburgh: Scottish Academic Press.

Marton, F., & Svensson, L. (1982). *Towards a Phenomenography of Learning. II: A Relational View of Study Skill. 1982:07*: Goteborg Univ , Molndal (Sweden) Dept of Education.

Meyer, J. H. F., & Shanahan, M. P. (2002). On Variations in Conceptions of 'Price' in Economics. *Higher Education, 43*, 203-225.

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied Multivariate Research - Design and Interpretation*: Sage Publications, Inc.

Ramsden, P. (1987). Improving Teaching and Learning in Higher Education: The Case for a Relational Perspective. *Studies in Higher Education, 2*(3), 275 - 286.

Scouller, K. (1998). The Influence of Assessment method on Students' Learning Approaches: Multiple Choice Question Examination versus Assignment Essay. *Higher Education, 35*, 453-472.

Scouller, K., & Chapman, E. (1999). *What Students Learn When they Write Essays*. Paper presented at the HERDS Annual International Conference, Melbourne, 12-15 July 1999.

Shanahan, M. P., Findlay, C., Cowie, J., Round, D. K., McIver, R., & Barrett, S. (1997). Beyond the 'Input-Output' Approach to Assessing Determinants of Student Performance in University Economics: Implications from Student Learning Centred Research. *Australian Economic Papers, Special Issue*, 17 - 37.

Tang, T., & Bain, J. (1994). *Repetitive Learning, Understanding and Examination Performance.* Paper presented at the Phenomenography: Philosophy and Practice Conference, 7-9 November 1994.

Tang, T., & Robinson, T. (2004). The Effects of Introductory Economics Course on Students' Beliefs and Aptitudes in Economics. *Australasian Journal of Economics Education, Vol. 1*(No. 2).

Wainer, H., & Thissen, D. (1993). Combining Multiple-choice and Constructed Response Test Scores. *Applied Measurement in Education, 6*(2), 103-118.

Walstad, W. B. (2001). Improving Assessment in University Economics. *Journal of Economic Education, 32*(3), 281-295.

Appendices

Appendix 1 Perceptions of Assessment Demand

	Your perceptions of EACH assessment type in this unit	Scouller	Own	Label
	I feel that this assessment, …			
1.	is one that good memory power can help get a good mark.		✓	MEM1
2.	has content that is mostly relevant to the real world.		✓	RW1
3.	is one that I can do well even if I don't have a deep understanding of the content.		✓	DU1
4.	is one in which I have to concentrate on memorising or reproducing a good deal of what I have to learn, if I want to do well.		✓	MEM2
5.	assesses my ability to integrate information from a variety of sources.	✓		INTG
6.	tests how well I memorise pieces of information.	✓		MEM3
7.	assesses my ability to analyse and evaluate the unit content.	✓		CRIT
8.	assesses my ability to reproduce factual details and knowledge.	✓		RLOW
9.	tests my ability to apply economic theories and principles to issues in the real world.	✓		RW2
10.	assesses my ability to reproduce ideas and viewpoints presented in readings and lectures.	✓		RHI
11.	measures how much I have listened and remembered in classes.	✓		MEM4
12.	provides me feedback about how well I understand the topic(s).		✓	FDBK
13.	gives me an opportunity to develop a deep understanding of the topic(s).		✓	DU2

'Own' refers to items developed by the authors based on student survey data.

Legend:

MEM = memorising; RW = (relevant/related to) real world; DU = (requiring/developing) deep understanding; INTG = (ability to) integrate; CRIT = (ability to analyse/evaluate) critically; RLOW = recall of low level knowledge; RHI = recall of high level knowledge; FDBK = (provides) feedback

Appendix 2 Model Calibration & Validation

The Perceptions of Assessment Demands model has to be calibrated to improve model fit for future model testing. The calibrated model will then be validated by an independent sample. This procedure was conducted for each of the three assessments. This appendix describes the procedures and results of calibrating and validating the model of Perceptions of Assessment Demands with independent samples for each of the three assessment contexts.

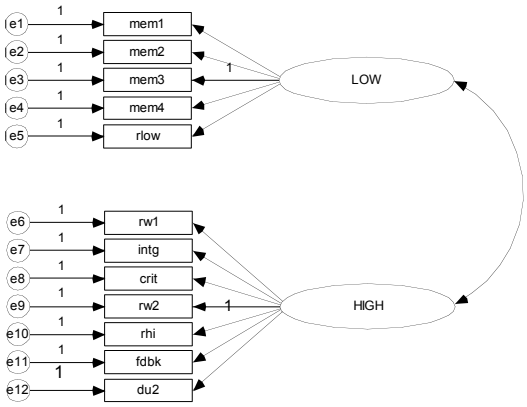
(1) Essay Assignment

Model Calibration

Method

The dataset used was the Week 13 samples (n = 647) from all three economics units. The data was split into roughly two equal halves. The calibration sample contains 328 cases. The model based upon the factorial structure obtained by EFA in the study is labelled Model 1A and is depicted in Figure 2.1. It was tested on the calibration sample.

Figure 2.1 Model 1A (un-refined model)



Results

MODEL 1A

The model hypothesised two perceptions – HIGH and LOW – that have no cross loadings. Items *mem1 to 4* and *rlow* load on the LOW factor and the remaining items load on the HIGH factor.

Model	χ^2	df	Normed Chi-square	CFI	RMSEA
1A	231.186	53	4.362	0.746	0.101
1B	154.188	52	2.965	0.857	0.078
1C	91.103	42	2.169	0.923	0.060
1D	76.421	41	1.864	0.944	0.051
1E	64.924	40	1.623	0.961	0.044
Validation	84.410	40	2.110	0.952	0.059

Table 2.1 Selected goodness of fit indices (Essay Assignment)

Table 2.1 presents selected goodness of fit statistics associated with this model. It is clear that Model 1A fits the sample data poorly as reflected by the CFI value of 0.746, normed Chi-square of 4.362 and RMSEA of 0.101.

To locate the misfit, the modification indices (MIs) were examined. Table 2.2 presents the MIs for the regression weights (factor loadings) for Model 1A. It is noted that only MIs greater than 10 are shown in the table. MIs represent the minimum improvement (reduction) in the Chi-square statistic if the relevant parameter (measurement error covariance or regression weight) is freely estimated. There is a large MI for the path: HIGH→*rlow* (48.8, in bold). It is evident that the model would be substantially improved if *rlow* is allowed to load on the HIGH factor. The path implies that when students see the assessment as assessing high level ability, there is also a need to reproduce low level knowledge. This re-specified model is labelled Model 1B.

			M.I.	Par Change
mem4_ass	<---	HIGH	14.434	.656
mem4_ass	<---	LOW	23.392	.392
mem4_ass	<---	fdbk_ass	17.483	.213
mem4_ass	<---	crit_ass	16.699	.255
mem4_ass	<---	rhi_ass	20.625	.285
mem4_ass	<---	mem3_ass	20.303	.258
mem4_ass	<---	rlow_ass	42.276	.362
crit_ass	<---	rlow_ass	11.233	.149
rhi_ass	<---	mem4_ass	12.614	.161
rhi_ass	<---	rlow_ass	18.686	.198
rlow_ass	**<---**	**HIGH**	**48.788**	**1.176**
rlow_ass	<---	mem4_ass	33.353	.311
rlow_ass	<---	du2_ass	22.043	.252
rlow_ass	<---	fdbk_ass	12.046	.172
rlow_ass	<---	crit_ass	31.690	.343
rlow_ass	<---	rw2_ass	32.032	.389
rlow_ass	<---	rhi_ass	38.408	.379

Table 2.2 Modification indices for regression weights – un-refined model

MODEL 1B

The re-specified model (Model 1B) was re-estimated. Referring to the second row in Table 2.1, it is clear that the goodness of fit indices for MODEL 1B improve a lot. The normed Chi-square and RMSEA are both just within the adequate level but CFI still does not reach the level of adequate fit. For further model re-specification, the standardised residual covariances matrix was examined. The matrix shows large correlation residuals associated with item *mem4*. Eight of the 11 residual correlations are greater than 2.58[4]. The item was intended to represent a perception that it is the factual knowledge remembered that is assessed. The result of EFA presented in the main study also indicates that this item loads on both the HIGH and LOW factor.

(*mem4*) I feel that this assessment measures how much I have listened and remembered in classes.

Scrutiny of the item (reproduced above) shows that it does not indicates what is being "remembered" – it can be basic facts and details or highly structured knowledge object (Entwistle, Marton, & Entwistle, 1993). The item can thus be subjected to different interpretations. This provides an explanation why it loads on both HIGH and

[4] It is noted that 2.58 is the critical z-value associated with the conventional level of significance of 0.05.

LOW factors. Based upon empirical evidence (large covariance residuals and cross loadings) and its ambiguous content, *mem4* was deleted from the instrument. The revised model is labelled Model 1C.

MODEL 1C

Table 2.1 (third row) shows the goodness of fit indices of the revised Model 1C. All three statistics indicate adequate model fit. Further examination of the MIs for regression weights shows there is no further meaningful factor loading to be added to the model. Moreover, the correlation residuals are mostly less than 2.58. But inspection of the MIs for covariances for Model 1C (Table 2.3) suggests further improvement of model fit can be achieved. Measurement error covariances represent "system rather, than random, measure error in item response" (Bryne 2001, p.107). The systematic error could be due to respondent characteristics (such as social desirability), overlap in item content (Byrne op cit.) or common method effect (Arbuckle 2006).

	M.I.	Par Change
e11 <--> e12	**13.214**	**.156**
e7 <--> LOW	6.748	-.084
e8 <--> LOW	4.831	.079
e10 <--> LOW	4.418	.077
e10 <--> e6	4.146	-.065
e2 <--> e8	6.670	-.108
e3 <--> e11	4.583	.088
e3 <--> e8	5.241	.074
e5 <--> e10	10.934	.123
e5 <--> e2	4.665	.102

Table 2.3 Modification indices for covariances – Model 1C

Referring to Table 2.3 the MI for the residual pair: e11 and e12 (in bold) has the largest value of 13.2. It indicates that by allowing the covariance between two residuals to be freely estimated the Chi-square will improve by at least 13.2. This is the empirical justification. The e11 and e12 are residuals of the two items *fdbk* and *du2*. Scrutiny of these two items reveals that both describe that the assessment deals with higher level cognitive abilities. That explains why they load on the same factor: HIGH. However, there is another commonality. It is that both items explicitly mention that the assessment is a device for developing understanding

(*du2*) This assessment develop a deep understanding of the topic(s).
(*fdbk*) This assessment provides me feedback how well I understand the topic(s).

It is this commonality that provides theoretic justification for the residuals of these two items to be correlated. In Model 1D, e11 and e12 are correlated.

Model 1D

Refer to Table 2.1 for the goodness of fit statistics for Model 1D. The Chi-square is reduced by 14.7 (more than the MI of 13.2 indicates). The model fit statistics now indicate good to very good model fit. Inspection of the

MIs for Model 1D (Table 2.4) shows further improvement of model fit is possible by correlating residuals: e5 and e10. MI for this error covariance is 10.75 (in bold).

Both items – *rlow* and *rhi* – start with "I feel that this assessment tests our ability to *reproduce* …". This indicates a possible commonality about the perception that the assessment is about reproduction of knowledge, the difference being the *types* of knowledge to be reproduced. Therefore this commonality of reproduction for assessment, which is not fully captured by the HIGH and LOW factors, suggests the residuals of the two items should be correlated. The covariance of e5 and e10 is included in the final model (Modal 1E).

	M.I.	Par Change
e7 <--> LOW	6.373	-.082
e8 <--> LOW	5.300	.082
e10 <--> LOW	4.759	.080
e10 <--> e6	4.625	-.069
e2 <--> e8	6.652	-.107
e3 <--> e11	5.190	.091
e3 <--> e8	5.624	.077
e5 <--> e10	**10.751**	**.123**
e5 <--> e2	4.660	.102

Table 2.4 Modification indices for covariances – Model 1D

Model 1E (Final model)

The goodness of fit statistics presented in Table 2.1 show that the final model has excellent model fit statistics (normed Chi-square =1.623, CFI = 0.961 and RMSEA = 0.044). Review of the MIs finds no further meaningful re-specification. Model 1E presented in Figure 2.2 represents the final Perceptions of Assessment Demands model for essay assignment.

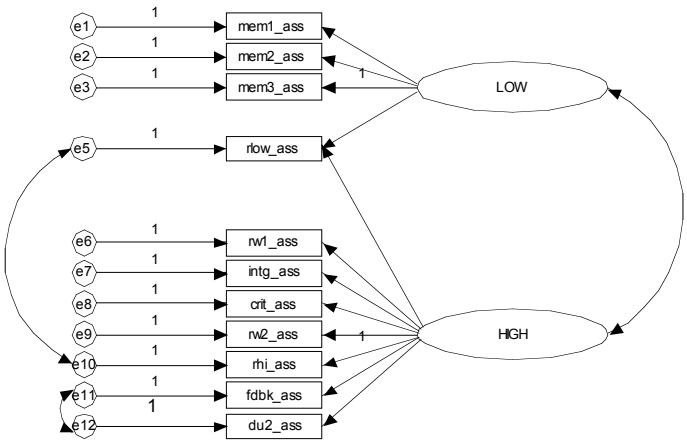


Figure 2.2 The Final Model - Essay Assignment

Model Validation

The final model (Model 1E) was tested on the validation sample – the other half of the Week 13 data (n = 319). The model fit statistics presented in Table 2.1 all indicate good model fit. Moreover, notice that the factor

loading of *rlow* on the factor HIGH is of medium size (0.39) and the re-specified covariance for e11 and e12 is significant at the level of 0.000 and the covariance for e5 and e10 significant at the level of 0.01. Overall, the refined final model fits the data on an independent sample reasonably well.

(2) MCQ Exam

Model Calibration

Method

The Perceptions of Assessments Model was calibrated for the MCQ exam data following a similar procedure. The Week 13 sample (n = 566) was randomly split roughly into two equal halves; the calibration sample has 264 cases and the validation sample 302 cases. The smaller sample size is because only Econ 1 and Econ 2 have MCQ exam in their assessment. The un-refined model which was same as the model for essay assignment (Figure 2.1) was tested on the calibration sample for the MCQ context.

Result

Model 2A

The goodness of fit indices for the un-refined model (labelled Model 2A) are given in Table 2.5. All statistics except the normed Chi-square indicate very poor fit.

Model	χ^2	df	Normed Chi-square	CFI	RMSEA
2A	173.630	53	3.276	0.781	0.093
2B	146.096	52	2.810	0.829	0.083
2C	114.449	42	2.725	0.862	0.081
2D	63.620	41	1.552	0.957	0.046
2E	56.367	40	1.409	0.969	0.039
Validation	66.123	40	1.653	0.962	0.047

Table 2.5 Selected goodness of fit indices (MCQ Exam)

Review of the MIs for the regression weights (Table 2.6, below) shows the largest MIs is for the path from *fdbk* to *du2* (MI = 42.5).

			M.I.	Par Change
mem4_mcq	<---	du2_mcq	18.314	.241
mem4_mcq	<---	fdbk_mcq	11.639	.179
du2_mcq	<---	mem4_mcq	15.144	.235
du2_mcq	**<---**	**fdbk_mcq**	**42.522**	**.340**
fdbk_mcq	<---	du2_mcq	32.643	.369
rlow_mcq	**<---**	**HIGH**	**25.116**	**.599**
rlow_mcq	<---	rw2_mcq	25.458	.285
rlow_mcq	<---	rhi_mcq	25.230	.276

Table 2.6 Modification indices for regression weights – Model 2A

This path indicates the two variables are association in some way that is not captured by the common factor HIGH. But as discussed earlier, this association is due to content overlap of the two items which can be dealt

with by correlating their residuals. This is confirmed by the large MIs for covariances of the residuals (e11 and e12) of these two items (MI = 45.8, Table 2.7). This modification will be dealt with in Model 2D. The next highest MI in Table 2.6 is for the factor loading of *rlow* on the factor HIGH. This re-specification is substantively meaningful as explained before and is reflected in the revised model (labelled Model 2B).

	M.I.	Par Change
e12 <--> e4	14.802	.233
e11 <--> e12	**45.799**	**.468**
e5 <--> HIGH	25.445	.186
e5 <--> e9	10.203	.150
e5 <--> e10	11.021	.186

Table 2.7 Modification indices for covariances – Model 2A

Model 2B

The goodness of fit statistics (Table 2.5) for Model 2B have improved substantially. But the CFI value of 0.829 is still smaller than the acceptable level of 0.9 and RMSEA of 0.083 is outside the upper limit for acceptable fit of 0.08. The source of misfit is located in the item *mem4*. The loading for *mem4* on the factor LOW has a small value of 0.17. It is also noted that the standardised residual covariances for the item *mem4* has two large values greater than 2.58 – 4.33 and 3.55. Further model revision involves the deletion of this item. The revised model is labelled Model 2C.

Model 2C

Although the revised model (Model 2C) after deleting the item *mem4* has only slightly improved model fit, the reduction of Chi-square by 31.6 associated with a degree of freedom of 10 is significant at the level of 0.005. The improvement is thus statistically significant. Review of the MIs for covariances (not shown here) indicates a substantial improvement of model fit if the residual pair – e11 and e12 – are correlated (MI = 45.9). This covariance was discussed earlier when modification indices of Model 2A were scrutinised. This re-specified model is labelled Model 2D.

Model 2D

The goodness of fit indices for Model 2D (Table 2.5) all indicate a very well fitting model. Review of MIs for covariances (not shown here) indicates that the largest MI for covariances is for the residuals: e1 and e6 (8.00). And the next highest MI is for e5 and e10 (6.99). Since there is no substantive reason for correlating e1 and e6. On the other hand as argued before, there is strong theoretical ground for correlating the residuals (e5 and e10) of the items: *rlow* and *rhi*. Thus, this latter residual pair was correlated in Model 2E.

Model 2E

Table 2.5 shows that the improvement of model fit for Model 2E in terms of the Chi-square statistic is slightly more than the MI of 6.99 suggests. All goodness of fit statistics are now in the excellent model fit range. No further meaningful model re-specification was suggested by the MIs tables. And thus Model 1E (Figure 2.3) was accepted as the final model.

Model Validation

The final model for MCQ exam was tested on the validation sample (n = 302). The goodness of fit statistics (Table 2.5) are all in the excellent range. Moreover, the re-specified factor loading (*rlow* on HIGH) and the two residual covariances are all significant at the level of 0.000. This represents further empirical support for the hypothesised final Perceptions of Assessment Demands model for MCQ exam.

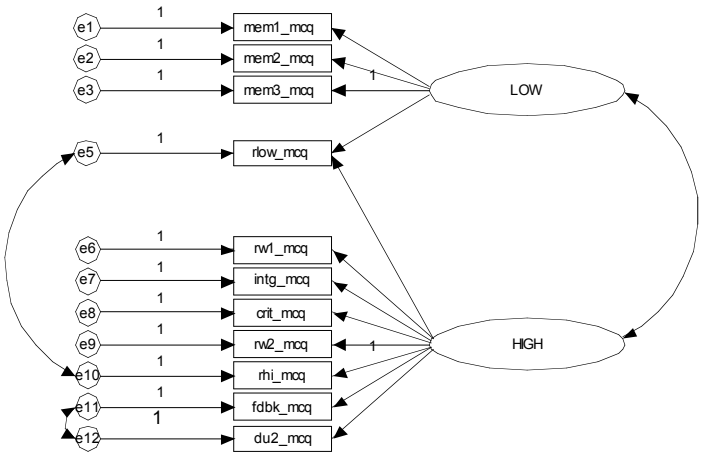


Figure 2.3 The Final Model – MCQ Exam

(3) Exam Essay

Model Calibration

Method

The procedure of calibrating and validating the model for exam essay is the same as for the other two assessments. The calibration sample size for exam essay is 305.

Result

Model 3A

The results of the calibration study are presented below (Table 2.8). The unrefined model (Model 3A) as expected does not fit the data well.

Model	χ^2	df	Normed Chi-square	CFI	RMSEA
3A	211.216	53	3.985	0.800	0.099
3B	158.317	52	3.045	0.865	0.082
3C	130.762	42	3.113	0.881	0.083
3D	83.987	41	2.048	0.942	0.059
3E	72.260	40	1.806	0.957	0.052
Validation	102.475	40	2.562	0.920	0.068

Table 2.8 Selected goodness of fit indices (Exam Essay)

Review of the MIs for regression weights (Table 2.9) shows a large MI (45.5) for the path from HIGH to *rlow* as in the other two assessments. In Model 3B this path is added.

			M.I.	Par Change
mem4_ex	<---	rhi_ex	10.198	.193
du2_ex	<---	fdbk_ex	22.316	.208
fdbk_ex	<---	du2_ex	17.795	.233
rhi_ex	<---	rlow_ex	21.720	.257
mem1_ex	<---	HIGH	10.387	-.481
mem1_ex	<---	crit_ex	11.385	-.201
rlow_ex	**<---**	**HIGH**	**45.465**	**.855**
rlow_ex	<---	du2_ex	14.660	.158
rlow_ex	<---	rw1_ex	14.850	.216
rlow_ex	<---	crit_ex	20.447	.228
rlow_ex	<---	rw2_ex	39.182	.298
rlow_ex	<---	rhi_ex	41.356	.326

Table 2.9 Modification indices for regression weights – Model 3A

Model 3B

Referring to Table 2.8, the model fit indices for Model 3B are still not within the acceptable range. The next re-specification focuses on the item that has the smallest factor loading – *mem4*. Examination of the residual correlation matrix reveals there are two large standardised residual covariances associated with the item *mem4* that are larger than the critical value of 2.58, and a number of smaller residual correlations close to this critical value. As argued above, there is not only empirical evidence but theoretical reason for deleting this item. The deletion of *mem4* is reflected in Model 3C.

Model 3C

Examination of the goodness of fit indices (Table 2.8) shows that the normed Chi-square (3.113[5]) is close to the acceptable limit of 3.0 and the CFI value of 0.881 is also just below 0.9. But the RMSEA of 0.083 is outside the limit of good fit (0.05). Inspection of the MIs for covariances indicates a large value (41.4) for the residual pair: e11 and e12 as for essay assignment and MCQ exam. The revised model (labelled Model 3D) therefore includes a covariance between these two residuals.

Model 3D

The model fit indices for Model 3D (Table 2.8) improve substantially but RMSEA is still outside the excellent model fit range of 0.05. Scrutiny of the MIs for covariances reveals that the highest MI is for the residual pair: e7 and e12 (11.2), and the next highest MI is for the residual pair: e5 and e10 (10.8). Given their small difference and the strong substantive justification for correlating e5 and e10 as discussed earlier, the latter residual pair was correlated in Model 3E.

Model 3E

Referring to Table 2.8, the CFI value of 0.957 and normed Chi-square of 1.806 both indicates excellent model fit. The RMSEA value of 0.052 is just above the benchmark of excellent fit of 0.05. Since further re-specification is not theoretically supported, the Model 3E (Figure 2.4 below) is accepted as the final model for exam essay.

[5] Note: The normed Chi-square for Model 3C actually increased compared with Model 3B, but the increase of Chi-square as a result of the deletion of *mem4* is statistically insignificant. Therefore deleting *mem4* does not represent a worse mode fit.

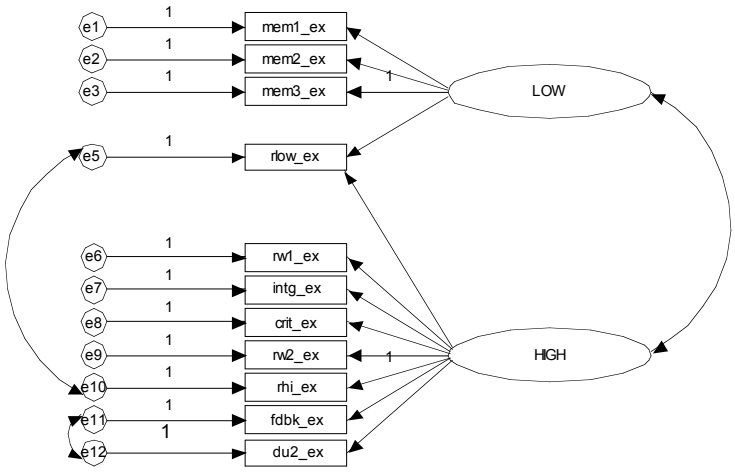


Figure 2.4 The Final Model - Exam essay

Model Validation

The number of cases in the validation sample is 342. When the final model (1E) was fit on the validation sample, the model fit (Table 2.8) is inferior to the calibration sample. However, the goodness of fit indices are all within acceptable range and the three re-specified parameters are all statistically significant at the level of 0.000. So, based on the generally good model fit, the final model (Model 1E) for exam essay was accepted.