Research Journal of Applied Sciences, Engineering and Technology 4(12): 1805-1814, 2012
ISSN: 2040-7467

Submitted: February 10, 2012 Accepted: March 06, 2012 Published: June 15, 2012

The Development and Validation of Conceptual and Procedural Understanding Test for Integral Calculus

[1]Tuan Salwani Awang Salleh and [2]Effandi Zakaria
[1]Universiti Kuala Lumpur Malaysia France Institute, Malaysia
[2]Faculty of Education, University Kebangsaan Malaysia

Abstract: This study discusses the process of developing and validating a conceptual and procedural understanding test for integral calculus. The aim was to produce a valid and reliable test to measure students' understanding of both types of knowledge, conceptual and procedural, in integral calculus. The test questions were developed in four main stages consisting of eight distinct systematic phases. The validation process started with discussions with content experts followed by a pilot test. Four experts provided inputs for the content validity purposes. Their inputs were used to produce a valid set of Integral Calculus achievement test comprising 15 questions. A pilot test assessed the reliability and fit statistics of the test. The test was administered to 79 students in a higher learning institution selected randomly from a group of students taking Integral Calculus. The data was analyzed using Winstep's software to ensure the quality of the questions. The analysis was done separately for each construct. The results indicated excellent item reliability indices with very good index separation values for both constructs. However, the person reliability indices for both constructs were only fair to good. Likewise, the results revealed only fair index separation values. In terms of fit statistics, three conceptual understanding items were replaced with a new item. Meanwhile, for procedural understanding items, two items were combined and one misfitting item was modified. Based on the experts' views and empirical data, the test questions were modified before they were used in the actual study.

Key words: Achievement, conceptual understanding, integral calculus, procedural understanding, rasch measurement model

INTRODUCTION

Calculus is one of the fundamental courses in engineering mathematics and engineering technology mathematics. As one of the core subjects, it acts as a backbone of the success in any engineering, including engineering technology and science fields (Cheshier, 2006; Kent and Noss, 2000). Integral calculus is one of the two fundamental concepts in this topic. Therefore, it is crucial for students choosing engineering and science fields of study to excel in calculus, including integral calculus. In Malaysia, calculus subject has been introduced into the secondary school curriculum as one of the options in form four and form five mathematics with an aim to provide a basis and a gateway for more advanced mathematics (Tall, 1997). Nevertheless, many studies have shown that students have difficulties understanding the concept of this topic (Crowther *et al*., 1997; Mahir, 2009). The under-preparedness in this topic is evident way before students enroll in engineering or science fields at the university. Most students seem to select an "informal" approach of teaching and learning to avoid the difficulties associated with understanding the foundation needed (Tall, 1992).

The development of engineering skills is related to two mutually-supportive factors, namely conceptual and procedural knowledge (Taraban *et al*., 2007). Therefore, to improve students' understanding of calculus, it is important to determine the level of their conceptual and procedural understanding of this topic. The assessment of their understanding needs to define these two factors clearly. However, the current practice of mathematics assessments lacks a clear definition of conceptual and procedural understanding. Hence, mathematics assessment needs to emphasize these two factors.

Mathematical understanding can be attained when the balance between conceptual and procedural knowledge is practiced in mathematics teaching and learning (Mahir, 2009). However, the real challenge is to find a proper balance between conceptual understanding and procedural fluency (Kulm, 1994). Kulm added that given the difficult balance between these two factors, assessment becomes complicated. Many mathematics educators believe that if students are procedurally fluent, they must understand the underlying concepts. Otherwise, they would not be considered conceptually knowledgeable and therefore they could not proceed to the higher mathematics course. Most of the time, these students were mistreated by drill

Corresponding Author: Tuan Salwani Awang Salleh, University Kuala Lumpur Malaysia France Institute, Malaysia

and practice method utilized to "improve" their understanding (Kulm, 1994). In other words, they remain at a lower level of mathematics. One possible way out of this problematic circumstance is by designing an assessment that could tell us clearly about students' conceptual and procedural understanding.

Conceptual understanding of integral calculus involves more than memorizing and applying procedure. Hiebert and Lefevre (1986) described two ways of developing conceptual knowledge. First, conceptual knowledge consists of relationships constructed and connected internally between existing ideas. Second, it develops when the existing knowledge relates to any new information received. Based on the New York State Education Department standard for mathematics learning, Engelbrecht *et al*. (2005) suggested that students use conceptual understanding of mathematics when they identify and apply principles, know and apply facts and definitions and compare and contrast related concepts.

Procedural understanding of integral calculus involves the fluency in skills of carrying out procedures flexibly, accurately, efficiently and appropriately (Engelbrecht *et al*., 2005). Procedural knowledge is composed of two parts, formal language (symbol representation) of mathematics and algorithms (rules) for completing mathematics tasks (Hiebert and Lefevre, 1986). In this case, procedures should be built based on related conceptual prior knowledge of differential calculus rather than rote memory. Students would be tested on applying the appropriate procedure to solve integral calculus problem. In this study, an alternative assessment of students' comprehension in integral calculus will be developed.

MATERIALS AND METHODS

An Integral Calculus Achievement test was administered to 79 students. The 15-question test consists of eight conceptual understanding and seven procedural understanding questions. These questions were written in a partial credit-scoring format, which gives 51 partial credit scores. The total score for conceptual understanding items was 23, while the total score for procedural understanding was 28. The samples involved were chosen randomly from the group of students taking Integral Calculus from July to December 2011.

In evaluating the quality of a test, reliability, validity and items bias are critical. This study applied Item Response Theory (IRT) to evaluate the quality of the test developed to measure students' understanding of integral calculus. The Rasch measurement model analysis was applied to evaluate students' responses. The data was analyzed using Winstep's software based on Rasch measurement model.

Test development process: In this study, psychometric properties of a calculus test were investigated using IRT.

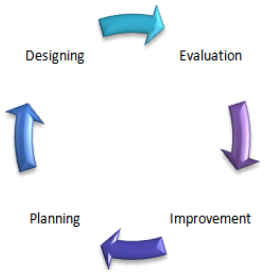

Fig. 1: The four cyclic stages

The test was developed based on four cyclic stages. In theory, the process is an ongoing cycle, but in real application, lecturers involved in designing any tests or assessments may use their experience to decide to end the cycle as soon as they obtain a meaningful and useful test. The four cyclic stages (Fig. 1) are diagrammed as follows:

The four stages are planning, designing, evaluation and improvement. These four cyclic stages can be further detailed out into eight systematic phases. The planning stage is divided into three smaller steps, which are construct analysis, content analysis and formation of Table of Specification (TOS). Designing stage comprises two phases, namely test design and first draft. The evaluation stage involves determining the validity and reliability of test items. Finally, the improvement involves items modification and final draft phases.

RESULTS AND DISCUSSION

Rasch measurement analysis involves two distinct phases of estimation. The first estimation procedure is the calibration of items difficulties and students' ability. The second estimation procedure is the estimation of fit (Bond and Fox, 2007). In this study, respondent-item maps were used to illustrate the relations between items difficulties and students' ability. Fit statistics is discussed based on the mean square values and Z standard values.

In Rasch measurement model, two reliability indices, person and item reliability indices, are provided to help lecturers decide on the sufficient number of items in any test developed based on their students' ability. Person reliability index indicates presumed replicability of person ordering if the same persons were given another parallel set of items measuring the same construct (Bond and Fox, 2007). Item reliability index refers to the items' placements replicability if the same items were given to another group of students with similar behavior (Bond and Fox, 2007). Person reliability in Rasch model is equivalent to Cronbach's Alpha or KR20 measurements (Linacre, 2006; Masters, 1982; Masters and Wright, 1984). Acceptable person and item reliability index ranges from 0.8 to 1.0, whereas the acceptable separation index is greater or equal to 2.0 (Aziz, 2010; Morales, 2009).

Table 1: Characteristics of conceptual and procedural understanding questions

Conceptual understanding	Procedural understanding
Application of prior knowledge to measure the understanding of new information.	Application of prior knowledge.
The knowledge needed for solving the problem was not taught directly in class.	The knowledge needed for solving the problem was taught directly in class.
The lecturer has not discussed the given task.	The lecturer has already discussed the given task.
Question designed in a way to give students flexibility in solving the problem given.	Question requires systematic calculations as taught in class to solve the problem given.
Question requires higher TOS level (in this case, it will be up to level four).	Question requires low TOS level (in this case, it will include level one and two).
Different representations needed to demonstrate a deep understanding of concept.	Mechanical application needed to demonstrate the level of students' knowledge without a deep understanding of concept.

Arslan (2010)

Table 2: Cognitive domains and types of questions

Cognitive domains	Types of questions	Sample verbs
Remember	Questions related to memorizing previously learnt facts	Define, identify list and name
Understand	Questions to show meaning or purpose of something	Convert, explain and summarize
Apply	Questions related to using information and ideas in specific situations	Compute, determine and solve
Analyze	Questions to reveal structure and interrelationships of something	Analyze, differentiate (distinguish) and relate
Evaluate	Questions requiring judgment based on reasoning	Compare, critique evaluate and judge
Create	Questions to combine various elements into a structure	Design, devise formulate and plan

(Aiken and Groth-Marnat, 2006; Green, 2010; Krathwohl, 2002)

Stage 1: Planning: Three phases are involved in the planning stage. The first phase relates to analyzing an appropriate construct to form a meaningful test. In this study, the formation of the test construct was based on the knowledge domain highlighted in the Bloom's Revised Taxonomy Matrix (Green, 2010; Krathwohl, 2002). However, out of four knowledge domains listed in the matrix, which are factual, conceptual, procedural and metacognition, this study focused only on conceptual and procedural knowledge. In this study, the constructs are written to reflect the understanding of integral calculus. Thus, the formed constructs are integral to conceptual and procedural understanding of calculus. These two constructs are used to develop a test to measure students' understanding of the knowledge of integral calculus. The characteristics used to measure these two constructs are based on the characteristics discussed by Arslan (2010). Six characteristics exemplify questions measuring conceptual understanding and procedural understanding. The differences between the characteristics of both construct (Table 1) are as follows:

The second phase in the planning stage is content analysis, which determines the contents of the test. This phase is very crucial, as it gives the details about all subtopics that need to be included in the assessment. In this study, the subtopics of integral calculus were based on the mathematics syllabus provided by the university involved in the study. Included are six subtopics in integral calculus, which are introduction to integral calculus, integrating products, integrating quotients, trigonometric integrals, area between the curve(s) and volume of solid of revolution. The last two subtopics involve applications of integral calculus. The number of questions for each subtopic in the test was determined based on the number of contact hours (lecture and tutorial) for each subtopic. The number of questions in the test was 7, 2, 2, 1 and 3 for Introduction to Integral, Integrating Products, Integrating Quotients, Trigonometric Integrals and Applications of Integral, respectively. The first subtopic comprised the greatest number of questions, since the techniques discussed in this subtopic were used in all other subtopics.

A clear assessment guideline, known as Table of Specifications (TOS), helps plan questions related to all subtopics determined in the second phase. Aiken and Groth-Marnat (2006) suggested that the guideline could be built based on Bloom Taxonomy. In this study, TOS was formed based on the Revised Bloom's Taxonomy. Six cognitive domains need to use different keywords, as shown in Table 2:

In this study, the questions were designed based on the university standards set for diploma students, according to which students are only assessed up to level four in TOS, that is, up to analysis level. Table 3, shows the distributions of items in the test according to the main subtopics in Integral Calculus. The first draft included 15 questions in total.

Stage 2: Designing stage: Designing stage comprises two phases, test design and initial draft. This study employs a combination of traditional and alternative assessment types. The traditional type of assessment comprises

Table 3: Distributions of items according to topics and bloom's taxonomy level

	Integral calculus					
	Introduction	Products	Quotients	Trigonometric	Area	Volume
Remember	1 item	1 item				
Understand	5 items	1 item	1 item			
Apply			1 item			1 item
Analyze	1 item			1 item	2 items	
Evaluate						
Create						

multiple choice questions, whereas the alternative type of assessment includes constructed-response format (Chatterji, 2003). Both types of test items were graded using the partial credit-scoring format. In this case, a set of scoring rubrics, or a set of criteria, was developed to ensure the judgments in the marking process are free from lecturers' biases or errors.

The first draft included eight questions measuring conceptual understanding and seven questions measuring procedural understanding of integral calculus. The questions on conceptual understanding were designed to measure the application of topics prior to integral calculus in order to test students' understanding of the newly learnt integral calculus concept. Furthermore, it consists of questions that assess students' higher thinking regarding various concepts learnt in this topic. The questions measuring procedural understanding focus on procedures involved in solving integral calculus problems.

Stage 3: Evaluation stage: Evaluation phase concerns the evaluation of the quality of assessment prior to its use in actual situation. In this study, the validation process was divided into two parts. The first part involved content validation. During content validation process, experts determined whether the items constructed are aligned with the specifications set by the university. The content validation procedure is conducted to ensure that the items are free from potential errors that could decrease the validity and reliability of the outcomes (Chatterji, 2003). Four experts were involved in content validation process. Three of them teach mathematics at the university and all have been teaching Engineering Technology Mathematics for more than ten years. One of the lecturers has been teaching mathematics for 20 years, while the other two lecturers have been teaching the subject for 16 years. Another expert is teaching a technical subject that is highly related to mathematics and he has been teaching the subject for 12 years. All comments from the experts were considered to modify the test items before running the pilot study.

The second part of the evaluation process involved data collection, observations and data analysis, followed by evaluations of the psychometric properties of the test. This part was done in a pilot test study. The main purposes were to determine whether the items constructed are free from errors, fit the purpose of the test, are suitable for the targeted population, contain adequate number of items and provide sufficient test duration. Both parts of the evaluation process provided insights into areas that need to be improved. The discussion of data analysis of the pilot study is divided into two parts, evaluation of conceptual understanding and evaluation of procedural of understanding of test items.

Evaluation of conceptual understanding test items: The first outcome is the person and item reliability and separation, as shown in Table 4. The person reliability value is 0.77. This value is appropriate (Aziz, 2010; Fisher, 2007) and indicates that the variability in the ability of students in this study is sufficient. The items seem to gauge a various levels of students' ability. In other words, the ability of all students in the group was well tested. This fact is supported by the value of person's separation, which equals to 1.83. This value can be round up to 2 separations, which implies that the group of students can be separated into two distinct groups.

Unlike the person reliability, the item reliability value of 0.95 was excellent while item separation value of 4.43 was very good (Aziz, 2010; Fisher, 2007). The item reliability index was excellent, indicating that a similar group of students would respond similarly when given the same items. The item separation is also tally with the number subtopics in the integral calculus.

Instead of analyzing all eight items as a whole, we can also observe the relation between the students' ability with items' difficulty level. Figure 2 shows the relation between the two components. Observe that the items are scattered around the mean ability value (M on the left), which is above the mean value of item difficulty (M on the right). However, one level of item number 5 in part B, specifically, item B5C4, is above the ability of the most capable student. This indicates that this item is the most difficult.

However, by looking at the Point Measure Correlation (PTMEA CORR) and INFIT and OUTFIT Mean Squares (MNSQ), item B5C4 can be used in the actual study.

Table 4: Reliability and separation indices for integral calculus conceptual understanding items

						Infit		Outfit	
		Total score	Count	Measure	Model error	MNSQ	ZSTD	MNSQ	ZSTD
Summary of 79 Measured Person									
Mean		13.1	23.0	36	0.55	0.98	0.0	1.05	0.2
S.D.		4.2	0.0	1.21	0.09	0.23	0.9	0.90	0.8
Max.		21.0	23.0	3.38	0.97	1.59	2.1	6.40	4.2
Min.		3.0	23.0	-2.67	0.49	0.54	-1.6	0.15	-0.7
Real	RMSE	0 .58	True SD	1.06	Separation	1.83	Person reliability		0.77
Model	RMSE	0.56	True SD	1.07	Separation	1.92	Person reliability		0.79
S.E. of Person Mean = 0.14									
Person Raw Score –to-Measure Correlation = .99									
Cronbach Alpha (KR-20) Person Raw Score Reliability = .79									
Summary of 23 Measured Items									
Mean		45.0	79.0	0.00	0.33	0.99	-0.1	1.05	0.0
S.D.		18.5	0.00	1.74	0.16	0.28	2.1	0.52	1.9
Max.		74.0	79.0	5.40	1.02	1.78	6.1	2.05	4.6
Min.		1.0	79.0	-2.93	0.25	0.70	-2.4	0.37	-2.0
Real	RMSE	0.38	True SD	1.70	Separation	4.43	Item reliability		0.95
Model	RMSE	0.37	True SD	1.70	Separation	4.65	Item reliability		0.96
S.E. of Item Mean = 0.37									
UMEAN = 0.0000 USCALE = 1.0000									
Item Raw Score-to-Measure Correlation = -.98									
1817 Data Points. Log – Likelihood Chi-Square: 1666.33 with 1716 d.f.p = 8009									

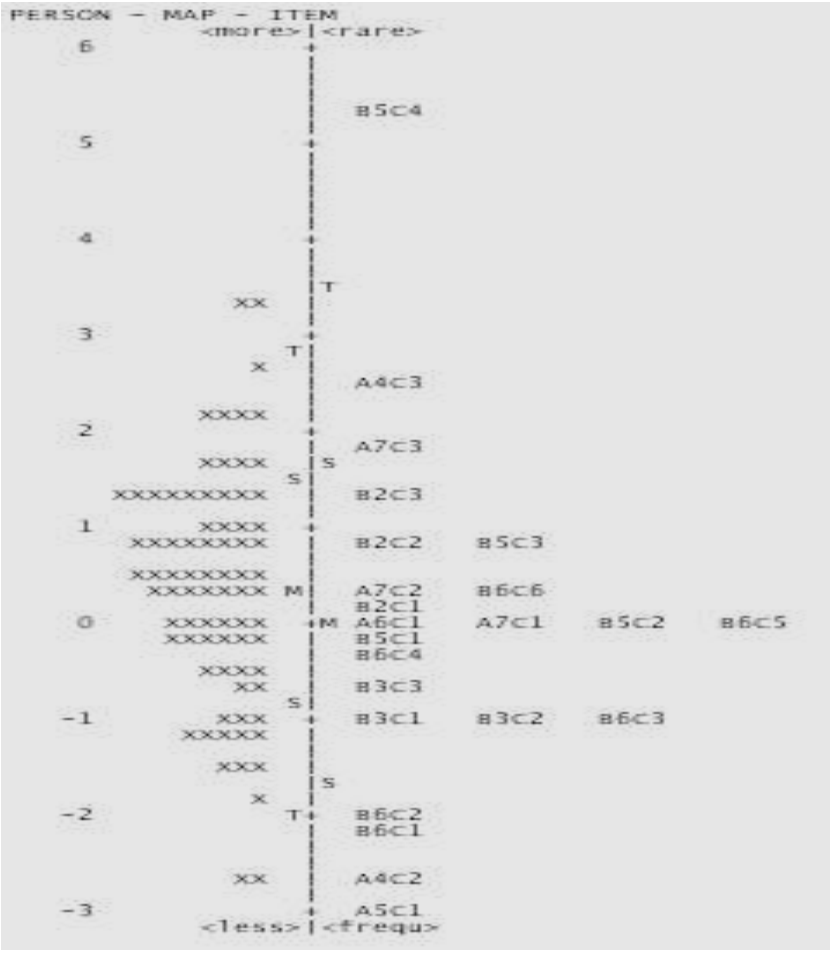

Fig. 2: Person-item map for conceptual understanding items

Table 5: Point measure correlation, infit, outfit and Z standard values for B5C4

					Infit		Outfit		PT-Measure		Exact Match		
Entry number	Total score	Count	Measure	Model S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	Item
17	1	79	5.40	1.02	1.07	0.4	1.03	0.6	0.04	0.13	98.7	98.7	B5C4

Table 6: Conceptual understanding items statistics: Measure order

					Infit		Outfit		PT-Measure		Exact Match		
Entry number	Total Score	Count	Measure	Model S.E.	MNSQ	ZSTD	NSQ	ZSTD	CORR.	EXP	OBS%	EXP%	Item
17	1	79	5.40	1.02	1.07	0.4	1.30	0.6	0.04	0.13	98.7	98.7	B5C4
2	12	79	2.50	0.34	1.23	1.1	1.95	1.8	0.14	0.36	83.5	85.9	A4C3
7	19	79	1.82	0.29	1.25	1.6	2.05	2.7	0.15	0.41	77.2	78.8	A7C3
10	25	79	1.34	0.27	0.93	-0.5	0.88	-0.4	0.49	0.43	73.4	73.9	B2C3
16	32	79	0.86	0.26	0.89	-1.1	0.88	-0.6	0.53	0.45	74.7	70.7	B5C3
9	33	79	0.79	0.26	0.88	-1.2	0.82	-1.0	0.55	0.45	75.9	70.3	B2C2
23	39	79	0.40	0.25	0.78	-2.4	0.70	-2.0	0.63	0.46	78.5	69.6	B6C6
6	40	79	0.34	0.25	1.63	5.4	1.91	4.4	-0.02	0.46	44.3	69.6	A7C2
8	43	79	0.14	0.26	1.10	1.0	1.05	0.4	0.40	0.46	64.6	70.2	B2C1
4	44	79	0.08	0.26	1.06	0.6	1.23	1.3	0.40	0.46	65.8	70.4	A6C1
5	45	79	0.01	0.26	1.78	6.1	2.01	4.6	0.12	0.46	41.8	70.6	A7C1
15	45	79	0.01	0.26	0.85	-1.4	0.77	-1.4	0.58	0.46	77.2	70.6	B5C2
22	45	79	0.01	0.26	0.80	-2.0	0.75	-1.5	0.61	0.46	77.2	70.6	B6C5
14	48	79	-0.19	0.26	0.86	-1.3	0.77	-1.3	0.57	0.46	78.5	71.8	B5C1
21	49	79	-0.26	0.26	0.77	-2.2	0.66	-2.0	0.63	0.46	79.7	72.3	B6C4
13	54	79	-0.62	0.27	0.85	-1.2	0.79	-0.9	0.56	0.45	81.0	75.3	B3C3
11	58	79	-0.93	0.29	0.86	-1.0	0.69	- 1.2	0.55	0.43	81.0	77.9	B3C1
12	58	79	-0.93	0.29	0.80	-1.4	0.64	- 1.4	0.59	0.43	83.5	77.9	B3C2
20	59	79	-1.01	0.29	0.70	-2.2	0.55	- 1.7	0.65	0.43	88.6	78.6	B6C3
19	68	79	-1.92	0.36	0.70	-1.4	0.40	- 1.5	0.59	0.37	88.6	87.0	B6C2
18	70	79	-2.19	0.38	0.70	-1.2	0.37	- 1.3	0.57	0.34	91.1	89.2	B6C1
1	73	79	-2.71	0.45	1.34	1.0	1.73	1.1	0.02	0.30	92.4	92.4	A4C2
3	74	79	-2.93	0.49	1.04	0.2	1.20	0.5	0.23	0.28	93.7	93.7	A5C1
Mean	45.0	79.0	0.00	0.33	0.99	-0.1	1.05	0.0			77.9	77.7	
S.D.	18.5	0.0	1.74	0.16	0.28	2.1	0.52	1.9			13.5	8.8	

Based on Table 5, item B5C4 can be retained in the actual study due to the following outcomes:

- PTMEA CORR is positive
- INFIT and OUTFIT mean square values fall within the acceptable range, i.e., $0.5 \leq MNSQ \leq 1.5$, with the values of 1.07 and 1.30 for INFIT and OUTFIT mean square values, respectively
- The Z standard values are also within the acceptable range, i.e., $-1.9 < Z < 1.9$, with the values of 0.4 and 0.13 for INFIT and OUTFIT Z standards, respectively

The positive Point Measure Correlation indicates that the item measured students' integral calculus conceptual understanding. The INFIT value is more sensitive to the targeted persons' responses pattern or vice-versa. The value between 0.5 to 1.5 is considered productive for measurement (Linacre, 2002). Similarly, OUTFIT value within the range 0.5 to 1.5 is considered productive for measurement. The OUTFIT value is sensitive to the items with difficulty far from the person, or vice-versa. In this case, even though the item is above students' ability, it is not considered far from students' actual ability. Data for which standard Z value falls within the -1.9 to 1.9 range have a reasonable predictability value (Linacre, 2002).

However, three elements related to trigonometric integrals need to be replaced with new item(s). The items are A7C1, A7C2 and A7C3. Based on Table 6, we can observe that the items are not fit to be included in the test.

A7C1 and A7C2 have negative Point Measure Correlation values, indicating that they are not measuring the same concept as other items; in other words, they are not measuring students' conceptual understanding of Integral Calculus. Item A7C3 needed to be replaced with a new item even though it has a positive Point Measure Correlation because this item follows up on items A7C1 and A7C2. Furthermore, the OUTFIT and standard Z values for item A7C3 are not within the acceptable range.

The OUTFIT values for items A4C2 and A4C3 are also not acceptable. Nevertheless, the items will be retained in the real study since their INFIT values are acceptable. Item A4C3 is considered difficult for the sample in this pilot study, but it will be maintained in the

Table 7: Reliability and separation indices for integral calculus procedural understanding items

					Infit		Outfit	
	Total score	Count	Measure	Model error	MNSQ	ZSTD	MNSQ	ZSTD
Summary of 79 Measured Person								
Mean	15.4	28.0	0.42	0.53	0.99	-0.1	0.99	0.0
S.D.	5.6	0.0	1.50	0.10	0.32	1.2	0.84	1.0
Max	27.0	28.0	4.47	1.06	1.81	2.9	4.69	3.1
Min.	4.0	28.0	-2.71	0.48	0.53	-2.1	0.16	-1.5
Real	RMSE	0 .57	True SD	1.39	Separation	2.45	Person Reliability	0.86
Model	RMSE	0.54	True SD	1.40	Separation	2.60	Person Reliability	0.87
S.E. of Person Mean = 0.17								
Person Raw Score –to-Measure Correlation = 0.99								
Cronbach Alpha (KR-20) Person Raw Score Reliability = .87								
Summary of 28 Measured Items								
Mean	43.4	79.0	0.00	0.33	1.00	-0.1	0.99	-0.1
S.D.	20.2	0.00	1.80	0.09	0.23	1.4	0.60	1.1
Max	77.0	79.0	3.21	0.73	1.72	3.7	3.63	3.0
Min.	9.0	79.0	-4.08	0.27	0.66	-2.6	0.31	-1.9
Real	RMSE	0 .35	True SD	1.77	Separation	5.01	Person Reliability	0.96
Model	RMSE	0.34	True SD	1.77	Separation	5.20	Person Reliability	0.96
S.E. of Person Mean = 0.17								
UMEAN = .0000 USCALE = 1.0000								
Item Raw Score-to-Measure Correlation = -0.99								
2212 Data Points . Log-Likelihood Chi-Square: 1833.04 with 2106 d.f.p = 1.0000								

Table 8: Procedural understanding items statistics: Measure order

					Infit		Outfit		PT-Measure		Exact Match		
Entry Number	Total Score	Count	Measure	Model S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.	OBS%	EXP%	Item
6	9	79	3.21	0.41	1.72	2.3	3.63	2.6	0.00	0.43	84.8	90.2	A3P2
21	12	79	2.77	0.36	0.90	-0.4	0.70	-0.4	0.52	0.46	87.3	87.2	B7P7
4	15	79	2.41	0.33	1.27	1.4	1.69	1.4	0.31	0.48	82.3	84.2	A2P3
28	16	79	2.30	0.33	0.95	-0.2	0.67	-0.7	0.54	0.49	79.7	83.3	B8P7
14	20	79	1.90	0.30	0.98	-0.1	0.89	-0.2	0.52	0.51	84.8	80.6	B4P5
20	22	79	1.72	0.30	0.73	-1.9	0.55	-1.6	0.67	0.52	86.1	79.3	B7P6
19	23	79	1.63	0.29	0.68	-2.4	0.52	-1.8	0.69	0.52	87.3	78.8	B7P5
18	24	79	1.55	0.29	0.66	-2.6	0.51	-1.9	0.71	0.52	86.1	78.3	B7P4
27	32	79	0.92	0.27	0.88	-0.9	0.74	-1.1	0.61	0.54	79.7	74.7	B8P6
17	34	79	0.77	0.27	0.75	-2.2	0.70	-1.4	0.67	0.54	79.7	74.2	B7P3
26	35	79	0.70	0.27	0.84	-1.3	0.71	-0.3	0.63	0.54	81.0	74.0	B8P5
3	36	79	0.63	0.27	1.51	3.7	1.81	3.0	0.23	0.54	59.5	73.9	A2P1
13	40	79	0.34	0.27	1.15	-1.3	1.10	0.5	0.45	0.53	69.6	73.5	B4P4
25	45	79	-0.02	0.27	0.89	-0.9	0.77	-0.9	0.59	0.52	79.7	74.1	B8P4
16	48	79	-0.24	0.27	1.03	0.3	0.87	-0.4	0.51	0.51	72.2	74.8	B7P2
2	50	79	-0.39	0.27	1.31	2.3	1.36	1.2	0.33	0.50	64.6	75.4	A1P3
24	50	79	-0.39	0.27	0.90	-0.8	0.86	-0.4	0.56	0.50	79.7	75.4	B8P3
9	52	79	-0.54	0.28	1.08	0.6	1.08	0.4	0.45	0.50	77.2	76.1	B1P3
5	54	79	-0.70	0.28	1.07	0.6	1.22	0.7	0.43	0.49	75.9	76.7	A3P1
15	55	79	-0.78	0.28	1.02	0.2	0.88	-0.2	0.48	0.48	77.2	77.1	B7P1
12	59	79	-1.11	0.30	0.91	-0.6	0.81	-0.3	0.51	0.45	79.7	78.8	B4P3
8	63	79	-1.49	0.32	0.94	-0.3	0.79	-0.3	0.46	0.42	81.0	81.5	B1P2
11	66	79	-1.81	0.34	1.01	0.1	1.02	0.2	0.38	0.39	83.5	84.3	B4P2
22	68	79	-2.05	0.36	0.97	-0.1	0.85	-0.1	0.38	0.36	86.1	86.6	B8P1
23	68	79	-2.05	0.36	0.97	-0.1	0.85	-0.1	0.38	0.36	86.1	86.3	B8P2
10	69	79	-2.18	0.37	1.02	0.2	1.18	0.5	0.33	0.35	87.3	87.7	B4P1
7	74	79	-3.05	0.49	1.05	0.3	0.71	-0.1	0.25	0.26	93.7	93.6	B1P1
1	77	79	-4.08	0.73	0.88	0.0	0.31	-0.5	0.25	0.17	97.5	97.4	A1P2
Mean	43.4	79.0	0.00	0.33	1.00	-0.1	.99	-0.1			81.1	80.7	
S.D.	20.0	0.00	1.80	0.09	0.23	1.4	0.60	1.1			7.8	6.4	

actual study in order to see the improvement of students' conceptual understanding after being introduced to a new intervention strategy in learning Integral Calculus. Since items A4C2 and A4C3 were related to each other, they will be combined to form one level of partial credit score.

Evaluation of procedural understanding test items: Person and item reliability and separation values are shown in Table 7.

Table 7 shows that the person reliability index value of 0.86 with 2.45 separations is acceptable. Likewise, the

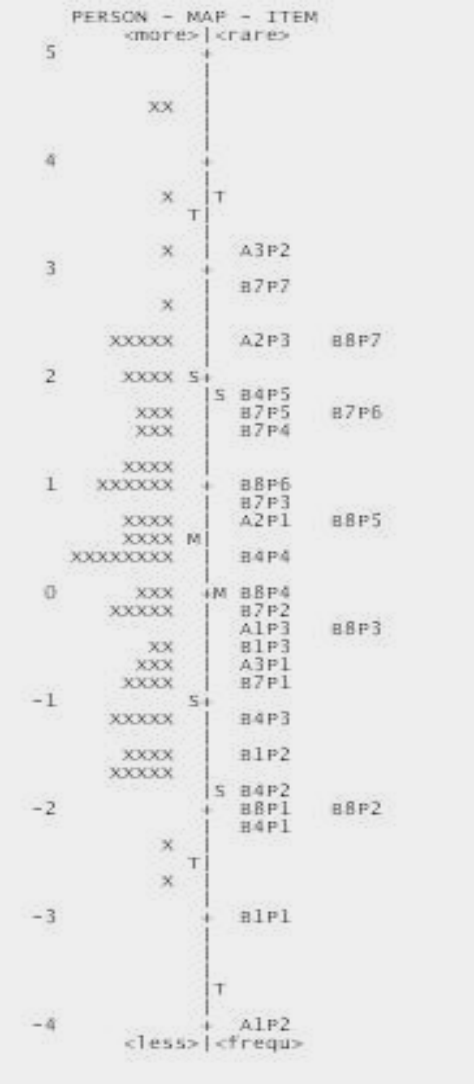

Fig. 3: Person-item map for procedural understanding items

item reliability value of 0.96 with 5.01 separations is also acceptably high. These values indicate that the items in this part are reliable and cater to various range of students' ability. However, few items need to be modified and removed, as indicated in Table 8.

Item A3P2 is a misfit item; therefore, it had to be modified in order to be used in the actual study. Items A3P2 and A3P1 are related, thus in the actual study, they will be combined to form one level question so that question A3 will only have one mark. The marking scheme for items B7P4 and B7P5 were reviewed since their standard Z values implied that the items were too predictable.

Figure 3 shows that number of respondents and the items difficulty are above two logits scale. This implies that the items were well spread out towards the students' ability (Siti *et al*., 2010). Difficult items are located above the scale, while less difficult items are located below the scale. The most difficult item in this part is item A3P2, whereas the easiest item is A1P2. All items will be used again in the actual test except for item A3P2, which was modified and combined with A3P1 to provide only one mark.

Stage 4: Improvement stage: The improvements were done based on the Stage 3 outcomes. The feedbacks from experts along with the pilot validation test provided information on how to modify the items. The modifications took place before using the test in the actual setting. The final draft was developed after removing bias from items and obtaining acceptable validity and reliability of items.

The items were modified in the final step. To measure conceptual understanding of Integral Calculus questions

accurately, A7 was replaced by a new item. The old item, which assessed trigonometric substitution in solving a quotient of functions, was replaced with a new item measuring trigonometric integral question while the marking scheme for question A4 was reduced from two marks to one mark. To assess procedural understanding of Integral calculus questions, item A3P2 was removed. Thus, question A3 had only one mark. The marking scheme for question B7 also changed from two to one mark.

CONCLUSION

In summary, the process involved in developing a test aimed at measuring conceptual and procedural understanding of integral calculus consisted of eight systematic phases. These eight phases can be categorized into four major stages. Planning Stage comprised three phases, which are construct analysis, content analysis and formation of table of specifications. Designing Stage consisted of test design and first draft phases. Evaluation Stage involved the examination of the validity and reliability of items. Finally, Improvement Stage involved items modification and final draft. These four main stages may be cyclic, meaning that additional improvements can be implemented continuously. Theoretically, there is no finite end to the process. However, based on ones' experience, the final decision can be made as soon as a meaningful test is obtained. In this study, a pilot study was conducted to validate a utility test. Four content experts validated the test through the content validation process, using Rasch analysis to determine the reliability of all items and fit statistics. Based on the experts' comments and the statistical data from a pilot test, a final draft of the test consists of 15 questions with 47 partial credit scores.

REFERENCES

Aiken, L.R. and G. Groth-Marnat, 2006. Psychological Testing and Assessment. 12th Edn., Pearson Education Group Inc., Boston, USA.

Arslan, S., 2010. Traditional instruction of differential equations and conceptual learning. Teach. Math. App., 29(2): 94-107.

Aziz, A.A, 2010. Rasch Model Fundamentals: Scale Construct and Measurement Structure. Kuala Lumpur: Perpustakaan Negara Malaysia.

Bond, T.G. and C.M. Fox, 2007. Applying The Rasch Model: Fundamental Measurement in the Human Sciences. 2nd Edn., Lawrence Erlbaum Associates, Mahwah, New Jersey.

Chatterji, M., 2003. Designing and Using Tools for Educational Assessment. Pearson Education, Inc., Boston.

Cheshier, S.R., 2006. Studying Engineering Technology: A Blueprint for Success. Discovery Press, ISBN: 0964696932.

Crowther, K., D. Thompson and C. Cullingford, 1997. Engineering degree students are deficient in mathematical expertise-why? Inter. J. Math. Educ. Sc. Tech, 28(6): 785-792.

Engelbrecht, J., A. Harding and M. Potgieter, 2005. Undergraduate students' performance and confidence in procedural and conceptual mathematics. Inter J. Math. Educ. Sc. Tech., 36(7): 701-712.

Fisher, W.P.J., 2007. Rating scale instrument quality criteria. Rasch Meas. Trans., 21(1): 1095.

Green, K.H., 2010. Matching functions and graphs at multiple levels of bloom's revised taxonomy. PRIMUS, 20(3): 204-216.

Hiebert, J. and P. Lefevre, 1986. Conceptual and Procedural Knowledge in Mathematics: An Introductory Analysis. In: Hiebert, J., (Ed.), Conceptual and Procedural Knowledge: The Case of Mathematics. Lawrence Erlbaum Associates, Hillsdale, New Jersey, pp: 1-27.

Kent, P. and R. Noss, 2000. The visibility of models: Using technology as a bridge between mathematics and engineering. Inter. J. Math. Educ. Sc. Tech., 31(1): 61-70.

Krathwohl, D.R., 2002. A revision of bloom's taxonomy: An overview. Theory into Pract., 41(4): 212-218.

Kulm, G., 1994. Mathematics Assessment: What Works in The Classroom. San Francisco, Jossey-Bass Inc., California.

Linacre, J.M., 2002. What do infit and outfit, mean-square and standardized mean? Rasch Meas. Trans., 16(2): 878.

Linacre, J.M., 2006. A User's Guide to Winsteps Ministep Rasch-Model Computer Programs. MESA Press, Chicago.

Mahir, N., 2009. Conceptual and procedural performance of undergraduate students in integration. Inter. J. Math. Educ. Sc. Tech., 40(2): 201-211.

Masters, G.N., 1982. A Rasch model for partial credit scoring. Psychometrika, 47(2): 149-174.

Masters, G.N. and B.D. Wright, 1984. The essential process in a family of measurement models. Psychometrika, 49(4): 529-544.

Morales, R.A., 2009. Evaluation of mathematics achievement test: A comparison between CTT and IRT. Inter. J. Educ. Psy. Assess., 1(1): 19-26.

Siti, R.A, I. Rodiah and M.I. Noriah, 2010. Differential item functioning in Malaysian generic skills instrument (MyGSI). Jurnal Pendidikan Malaysia, 35(1): 1-10.

Tall, D., 1992. Students' Difficulties in Calculus. Paper Presented at the International Congress on Mathematical Education (ICME 7), Quebec, Canada.

Tall, D., 1997. Functions and Calculus. In: Bishop, A.J., K. Clements, C. Keitel and J. Kilpatrick (Eds.), International Handbook of Mathematics Education. Kluwer, Dordrecht, pp: 289-325.

Taraban, R., A. DeFinis, A.G. Brown, E.E. Anderson and M.P. Sharma, 2007. A paradigm for assessing conceptual and procedural knowledge in engineering students. J. Eng. Educ., 96(4): 335-345.