CHAPTER 4

METHODOLOGY

4.0 Introduction

This study was designed to investigate Form four science students' representational competence of basic chemical concepts. Specifically, the study attempted to assess students' overall levels of understanding of basic chemical concepts, chemical representations, as well as their representational competence in chemistry. It also sought to compare students' with different levels of understanding of chemical concepts and chemical representations in their representational competence. In addition, an attempt was made to identify their alternative conceptions of chemical concepts, chemical representations, as well as their difficulties when interpreting and using chemical representations. Semi-structured interviews were conducted to gain further insights into selected students' conceptions of chemical representations, as well as their representational competence in chemistry. A further search was undertaken to examine the influence of prior knowledge, working memory capacity, developmental level, learning orientations on representational competence, and subsequently, the best predictor variable for representational competence was determined. A total of 411 Form four science students from seven urban secondary schools in the State of Perak participated in this study. Data for the study was obtained from seven instruments consisting of five paper-and-pencil tests, a questionnaire and interviews. The Statistical Package for the Social Sciences (SPSS) was used to process and analyze quantitative data collected from the study.

In order to achieve the above mentioned objectives, the methodology involved several main procedures. The data for this study were collected using both quantitative and qualitative techniques. Sources of data included paper-and-pencil tests, questionnaires, worksheets, online quiz, interviews, as well as examination of curriculum and instructional materials. Form four students' overall levels of understanding of chemical concepts, as well as their alternative conceptions of chemical concepts, were investigated using a paper-and-pencil test – the Test on Chemical Concepts (TCC). Students' overall levels of understanding of chemical representations, as well as their alternative conceptions of chemical representations, were investigated using another paper-and-pencil test – the Test on Chemical Representations (TCR). A further paper-and-pencil test – the Test on Representational Competence (TRC) was used to assess students' overall levels of representational competence, as well as to identify their learning difficulties when interpreting and using chemical representations. Possible cognitive variables influencing students' representational competence were identified using either modified versions of existing instruments or new instruments designed by the researcher in this study. Students' prior knowledge in chemistry were assessed using the Test on Chemical Concept (TCC) and the Test on Chemical Representations (TCR), developmental level was assessed using the Classroom Test of Scientific Reasoning, CTSR (Lawson, 2000), working memory capacity was measured using the Digit Span Backwards Test, DSBT (Wechsler, 1955), and learning orientations was investigated using the Learning Approach Questionnaire, LAQ (Boujaoude, Salloum & Abd-El-Khalick, 2004). While the CTSR, DSBT, and LAQ are existing instruments, the TCC, TCR, and TRC are new instruments designed by the researcher in this study. This chapter shall discuss the methodology in detail as follows: (i) The sample, (ii) The instruments, (iii) Data collection, (iv) Data analysis

4.1 The Sample

Although the target population of the study was all the Form four science students in Malaysia, only Form four science students in the State of Perak appeared to be the accessible population for the study. The actual sample of this study comprised 411 Form four science students from 13 intact classes of seven urban secondary schools in Perak. Of the seven schools selected, three were National Type Secondary Schools or *Sekolah Menengah Jenis Kebangsaan (SMJK),* three were National Secondary School or *Sekolah Menengah Kebangsaan (SMK)*, and one was a Technical Secondary School or *Sekolah Menengah Teknik (SM Teknik)*. In terms of gender composition, four were co-educational schools, two were all-boys schools and an all-girls school. Table 4.1 shows the number of classes and number of students in the schools selected while Tables 4.2 and 4.3 show the profile of the subjects in terms of gender and ethnic background respectively.

At the time of data collection, the age of the subjects ranges from 15 to 17, with a mean age of 16 years. All the subjects were taking chemistry as a subject for the first time. Other similar key characteristics included: (i) same chemistry curriculum, (ii) same assessment instruments, (iii) same medium of instruction, as chemistry is taught in English, and (iv) all the subjects were using the same chemistry text book.

Working with 5 independent variables, the sample size (n=411) is considered sufficient for the analysis of data and the generalization of findings (Hair, Anderson, Tatham & Black, 2006).

Table 4.1
Number of classes and number of students in the schools selected

*School	No. of classes	No. of students
YC	5	142
NH	3	115
SP	1	21
DMK	1	31
TK	1	23
TC	1	37
ST	1	42
Total	13	411

* *Short forms are used for confidentiality*

Table 4.2
Profile of subjects in terms of gender

Gender *School	Male	Female	Total
YC	98	44	142
NH	53	62	115
SP	21	-	21
DMK	10	21	31
TK	7	16	23
TC	-	37	37
ST	42	-	42
Total	231 (56.2%)	180 (43.8%)	411 (100.0%)

* *Short forms are used for confidentiality*

Table 4.3
Profile of subjects in terms of ethnic background

Race *School	Chinese	Malay	Indian	Total
YC	139	1	2	142
NH	112	3	-	115
SP	9	8	4	21
DMK	-	25	6	31
TK	-	22	1	23
TC	4	16	17	37
ST	41	1	-	42
Total	305 (74.2%)	76 (18.5%)	30 (7.3%)	411 (100.0%)

* *Short forms are used for confidentiality*

4.2 The Instruments

For the purpose of data collection, seven instruments were employed. These were:

i The Test on Chemical Concepts (TCC)

ii The Test on Chemical Representations (TCR)

iii The Test on Representational Competence (TRC)

iv The Classroom Test of Scientific Reasoning (CTSR)

v The Digit Span Backwards Test (DSBT)

vi The Learning Approach Questionnaire (LAQ)

vii The Semi-Structured Interviews (SSI)

As the sample size (n = 411) was relatively large, the main instruments used for data collection were paper-and-pencil tests (TCC, TCR, TRC, DSBT and CTSR) and survey questionnaires (LAQ). Interviews were limited to a smaller purposive sample (n=9) to gain further insight and understanding related to Form four students' representational competence of basic chemical concepts.

Preliminary Survey Questionnaires (PSQ) on Chemistry Teachers' and Chemistry Students' Perceptions of Chemical Representations

Subsequent to informal interviews with some chemistry teachers and chemistry students on issues pertaining to representations in chemistry, two instruments were designed to gather some preliminary data on chemistry teachers' and chemistry students' perceptions of chemical representations. These two instruments were:

(i) Chemical Representations: What are Chemistry Teachers' Perceptions?

(ii) Chemical Representations: What are Chemistry Students' Perceptions?

The above instruments (see *Appendix 2* and *Appendix 2a*) were constructed based on feedbacks from the informal interviews conducted earlier, as well as from observation and classroom experience as a teacher researcher in chemical education.

The purpose of administering these preliminary Survey Questionnaires was to collect empirical data to corroborate the information obtained through informal interviews conducted earlier (see Chapter 1 - Section 1.4: Rationale of the Study).

The teacher's version of the PSQ were administered to 40 chemistry teachers throughout the state of Perak during a seminar on chemistry in October 2007 while the student's version of the PSQ were administered to 42 Form four science students taught by the researcher in this study. The student's version of the PSQ was administered between October and November, 2007.

The survey data showed that 95% of the teachers (n=38) indeed had no clear idea what chemical representations are. They tended to relate chemical representations to symbolic representations like symbols of the elements, chemical formulae and chemical and ionic equations only. None of the teacher respondents were aware of the three levels of thinking or representation in chemistry. See *Appendices 2 and 2a* for samples of respondents' questionnaires.

4.2.1 The Test on Chemical Concept (TCC)

The TCC is a two-part paper-and-pencil test used to investigate Form four students' understanding of basic chemical concepts, which was considered the prior knowledge in chemistry of the Form four students in this study. Since chemical representations have a dual nature – they are visual displays as well as conceptual constructs, knowledge of appropriate chemical concepts are required for a conceptual understanding of chemical representations, as well as to be able to interpret and use representations in chemistry. That is: representational competence.

4.2.1.1 Development of the TCC

Content area of the TCC

The subjects of this study comprised only Form four science students. At the time of data collection, they would most probably had completed only 1 ½ semester (or 8 months) of their chemistry course. Therefore, only basic chemical concepts related to matter such as pure substances and mixtures, elements and compounds, atoms, molecules and ions, sub-atomic particles, proton number and nucleon numbers, electron arrangement, valence electron, as well as the idea of a physical change or a chemical change, and chemical bonds were assessed (*Appendix 3a*).

Construction of the TCC

Items in the TCC were generated by the researcher based on classroom experience (see *Appendix 5*). The pilot version of the TCC comprised a total of 30 items in two parts. Part A contains 22 True-False items while Part B contains 8 Multiple-choice items. The initial draft of the TCC was given for validation to two experienced secondary school chemistry teachers and a university Professor in chemical education. Their feedbacks were favourable.

Scoring procedure for the TCC

The TCC is a two-part paper-and-pencil test with 22 True-False items in Part A and 8 Multiple-Choice Questions in Part B. For both Parts A and B, each correct answer was awarded one point and no point was given for an incorrect answer. Hence, total possible test point awarded for the 30 dichotomous items was 30. Test score for the TCC may range from a minimum of 0 to a maximum of 30 points.

Translation of the TCC

The researcher in this study believed that the English version of the TCC was sufficient for the purpose of the study. This was because English is the medium of instruction in Form four Chemistry. Furthermore, the subjects in this study have been taught science and mathematics in English since they were in Form one. The year 2008 marked the full Implementation of the Teaching and Learning of Science and Mathematics in English or *Pengajaran dan Pembelajaran Sains dan Matematik dalam Bahasa Inggeris (PPSMI).* Therefore, no translation of the TCC was done and only the English version was administered to the subjects.

Pilot study of the TCC

For the TCC, specifically, the objectives of pilot-testing the instrument were to estimate the time required for the subjects to complete the test, find out the difficulty level of each item, as well as the discriminating power of the item. Item analysis had been done by computing the difficulty index (p value) and the index of discrimination (ID) for each item (see *Appendix 6*).

Reconstruction of the TCC

Based on *Appendix 6*, items with undesirable difficulty index or discrimination index were either reconstructed or discarded in order to increase the reliability of the test score. Such fine-tuning procedures were necessary to increase the reliability of the new instrument. The final version of the TCC used in the actual study contained 15 True-False items in Part A and another 15 Multiple-choice questions in Part B (*Appendix 5a*).

4.2.1.2 Validity of the TCC

Since the TCC is a test, evidence for face validity and content-related validity need to be gathered to help establish the validity of the new instrument.

To check for face and content-related validity, the draft version of the TCC had been reviewed by two experienced chemistry teachers from two different premier secondary schools in Perak, and a university Professor in chemical education. One of the chemistry teachers has a Bachelor Degree in Science with Education, majoring in Chemistry. This teacher has taught chemistry for the past 25 years and is currently teaching chemistry for Forms four to six. Another teacher reviewer is also a major in chemistry and has a Masters Degree in Science Education (Chemistry). This teacher has taught chemistry for the past 15 years and is currently a chemistry lecturer for matriculation classes. The items in the TCC appeared to be relevant for testing students' understanding of basic chemical concepts.

To establish the content validity of the instrument, the content area of interest had been determined (*Appendices 3 & 3a*). A table of specification was also constructed (*Appendix 4*).

Although a total of 12 chemical concepts were tested in the TCC (*Appendix 3*a), a look at the Table of Specification (*Appendix 4*) shows that 20 of the 30 items (or two-thirds) of the items tested on the first five chemical concepts.   The rationale is these five concepts are the most basic or fundamental concepts in chemistry as the subjects in this study were Form four students. Besides, the students had in fact learned these concepts (except the concept of `ion') in their lower secondary science lessons in Form one and Form two.

4.2.1.3 Reliability of the TCC

Each item in the TCC was scored dichotomously. Hence, the test score reliability was estimated using the Kuder-Richardson formula (Kuder & Richardson, 1937, cited by Mehrens & Lehmann, 1973). Pilot test of the 30-item TCC with a similar sample of students (n=57) gave a KR-20 of 0.56 (*Appendix 7).* This indicates the new instrument has moderate reliability. After the pilot study, steps were taken to reconstruct the test items to further increase test score reliability of the TCC.

In the actual study, a KR-20 of 0.59 was recorded for the 30-item reconstructed version of the TCC for a sample of n=383. Additionally, test-retest reliability was also estimated. The statistical procedure used to examine test-retest reliability is correlation. Since total test scores for the TCC is continuous, the Pearson correlation coefficient is the statistic used to reflect test-retest reliability.   A high correlation coefficient indicates the instrument is stable over time. For the actual study, test-retest with a smaller sample (n=45) after a 3-week interval gave a correlation coefficient of r=0.84. This figure suggests that the scores on the TCC are stable over time. *Appendix 7a* shows the correlation coefficient and a scatter plot of test-retest scores of the TCC for the actual study.

4.2.2 The Test on Chemical Representations (TCR)

The TCR was used as the instrument to collect data for research questions (i) (b) and (iii) (b) in this study. The TCR served a dual purpose: (i) the total test score was used as a measure of students' overall levels of understanding of chemical representations, (ii) it was also used to identify students' alternative conceptions of chemical representations,

For the pilot study, the TCR is a 50-item, `true' or `false' format paper-and-pencil test (*Appendix 11*). Table 4.4 below shows the composition of the items in the TCR (pilot study).

Table 4.4
Composition of items in the TCR (pilot study)

True/Falsity of statement	Item No.	Number of items
True	3, 5, 6, 9, 11, 14, 18, 21, 25, 29, 31, 35, 38, 39, 41, 42, 45, 47, 49, 50.	20
False	1, 2, 4, 7, 8, 10, 12, 13, 15, 16, 17, 19, 20, 22, 23, 24, 26, 27, 28, 30, 32, 33, 34, 36, 37, 40, 43, 44, 46, 48.	30

Since each statement was scored dichotomously, that is: one point for a correct response and no point for an incorrect response, total test score ranged from a minimum of `0' to a maximum of `50' points.

The respondents were asked to select `true' or `false' for each statement. The option `uncertain' or `do not know' was not available to avoid the possibility of some respondents selecting `do not know' without thinking or even reading the statements.

A response sheet was provided and respondents were asked to circle either `T' or `F' for each item. The decision to use a response sheet was to save cost and to enable easy scoring and analysis of responses.

The instrument (TCR) was designed to have these characteristics:

i. it could be administered to a large group of students,

ii. the test was relatively simple to score,

iii. scoring was objective,

iv. it contained sufficient items to explore aspects of the three levels of representation of matter and various kinds of chemical representations.

The true-false or alternate response item is essentially a two-response multiple-choice item in which only one of the propositions (answers) is presented and the student judges the truth or falsity of the statement (Mehrens & Lehmann, 1973).

Rationale for choosing the true or false item format for the TCR:

i. The test can cover a large amount of subject matter in a given testing period than any other objective item. Therefore, more questions can be asked. According to Frisbie (1973), cited in Ebel (1993), a student can answer two true-false items for every two multiple choice format items.

ii. The test can be scored accurately, quickly, reliably, and objectively.

iii. Are particularly suitable for testing beliefs in popular misconceptions.

In a study on students' understanding of ionic bonding, Taber (1997) designed an instrument - "The Truth about Ionic Bonding Diagnostic

Instrument" to explore students' understanding of ionic bonding. The instrument contained 30 true-false statements as the only item format.

iv. Students do very little blind guessing on good true-false tests. As cited by Ebel (1993, p.138), "The probability of an examinee achieving a high score on a T-F test by guessing blinding is extremely low. The influence of blind guessing on the scores of a test diminishes as the test increases in length."

4.2.2.1 Development of the TCR

Development of the TCR involved several procedures namely: Defining the content domain of the TCR, construction of the TCR, pilot study of the TCR, and reconstruction of the TCR. These procedures will be discussed in detail as follows:

Defining the content domain of the TCR

The content domain of interest was the three levels of chemical representation of matter (macroscopic, submicroscopic and symbolic). *Appendix 9* shows the content domain for the TCR. The content domain for the TCR was determined after careful examination of some curriculum and instructional materials such as the Curriculum Specifications for Form four Chemistry (Malaysian Syllabus), text book and reference books in chemistry currently used by Form four students, as well as selected college chemistry textbooks (International Editions).

Construction of the TCR

In order to prepare the draft version of the TCR, a review of literature related to science education, in particular chemical education research was conducted. In addition, various curriculum and instructional materials related to chemistry were sought and critically examined. A collection of students' incorrect answers in their written exercises, laboratory reports, revision worksheets, test papers, and examination scripts were rich reference resources for constructing the items in the TCR.

Each item in the TCR consisted of a propositional statement related to the three levels of representation of matter or chemical representations. To investigate students' conception through the `true' or `false' item format, a useful strategy is to create pairs of statements, one true and one false, based on a single idea.

For example:

(i) *H_2 and O_2 are symbols of the element hydrogen and oxygen respectively.*

(ii) *H and O are symbols of the element hydrogen and oxygen respectively.*

While statement (ii) is a true statement, statement (i) is a false statement. In this item format, the intended correct answer should be obvious only to those who have good command of the concept being tested, whereas the wrong answer should be made attractive to those who lack the desired command.

Regardless of the type of interpretation to be made of the scores, it is believed that the job of a test item is to discriminate between those who have and those who lack command of some element of knowledge. Those who have achieved command should be able to answer the question correctly, while those who lack it should find a wrong answer attractive. Therefore, to enhance item discrimination in the TCR, several other measures have also been taken. These included:

i. Using more false statements than true statements

   In the TCR (pilot study), there are 30 false statements and 20 true statements. It is believed that when in doubt, student seem more inclined to accept than to challenge propositions presented in a true-false test. False statements also tend to be more highly discriminating than true statements. As commented by Barker and Ebel (1981), cited in Ebel (1993, p.149):

> In the absence of firm knowledge, students seem more likely to accept than to reject a declarative statement whose truth or falsity they must judge. If the false statement tend to be higher in discrimination, it would seem advantageous to include higher proportion of them, perhaps as many as 67%. Even if students come to expect a greater number of false items, the technique still seems to work…

ii. Word the item so that superficial logic suggests a wrong answer.

iii. Make the wrong answer consistent with a popular misconception.

iv. Use phrases in false statements that give respondents "the ring of truth".

An item is written based on a single proposition. As each statement was written, it was also identified and marked as either TRUE or FALSE. The items were also checked to avoid any double-barreled items (that is: partly true and partly false). Initially, the written statements were grouped separately in two sections: true statements, and false statements. A total of 25 true statements and 37 false statements were tentatively generated.

Every effort was made to ensure the items were expressed as concisely and as clearly as possible. Editing of the items was done where necessary. These statements were then carefully examined to make sure that each of the statement was indeed clearly true or clearly false.

A final selection of statements to be included in the TCR was done. The number of true statement was trimmed down to 20 while the number of false statement was reduced to 30, giving a total of 50 statements. Subsequently, the 50 statements were randomly placed to form the 50-item TCR. Hence, the draft version of the TCR comprised 20 true statements and 30 false statements.

The draft version of the TCR was given for validation to the same panel of reviewers as the TCC (see Section 4.2.1.2). Their feedback was favourable.

Pilot Study of the TCR

The main objective of pilot-testing the TCR was to estimate the time required for the subjects to complete the 50-item test, to estimate the difficulty level and discrimination index of each item. Other objectives were: (i) to detect the presence of any unintended errors in the instrument, such as inappropriate use of words, phrases or any other ambiguity in the items, and (ii) to serve as a trial run to provide useful information for any unexpected problems that might arise in the actual study.

During the pilot study, none of the subjects took more than 30 minutes to complete the test. The difficulty index (p) and discrimination index (ID) for each item were also computed (*Appendix 12*). For a classroom test, normally items with undesirable p value or ID are either reconstructed or discarded to increase the reliability of the test score. However, the ID is only a useful measure of item quality whenever the purpose of a test is to produce a spread of scores, reflecting differences in students' achievement, so that distinction may be made among the performances of respondents. In this study, the TCR served a dual purpose. The main purpose of administering the TCR was to investigate students' conceptions of chemical representations and to identify their alternative conceptions of chemical representations. A check of ID showed that most of the items with low ID (< 0.20) were items with low (between 0.10 – 0.30) to very low (< 0.10) p values (*Appendix 12*). These appeared to be difficult or tricky items even the good students could not answer. Deleting these items defeated the purpose of administering the test.

Reconstruction of the TCR

Feedbacks from the pilot study, wherever relevant, as well as suggestions from the Vetting Committee of the research proposal, University of Malaya, were used to reconstruct the final version of the TCR for the actual study. In the process of reconstruction, minor changes were made to the content domain of the TCR (*Appendix 9a*). The table of specifications was also modified accordingly (*Appendix 10a*). The TCR used for the actual study comprised two parts. Part A contained 30 true-false items while Part B contains 6 multiple choice items (*Appendix 11a*).

4.2.2.2 Validity of the TCR

As the TCR was a new instrument, appropriate and sufficient evidence were gathered to help establish the validity of the new instrument. During the course of developing the TCR, attempts had been made to provide evidence for face validity and content validity.

To check for face and content-related validity, the draft version of the TCR had been reviewed by the same panel of reviewers as the TCC, comprising two experienced chemistry teachers from two different premier secondary schools in Perak, and a Professor in Chemical Education of a reputable university in Malaysia. The items in the TCR appeared to be relevant for the purpose of this study.

Since the TCR is a test, it is important to establish the content validity of the instrument. This was done by:

i. Constructing a concept map of the three levels of chemical representation of matter (*Appendix 8*),

ii. Determining the content domain of the test (*Appendix 9* & *9a*),

iii. Preparing a table of specification for the test (*Appendix 10* & *10a*)

4.2.2.3 Reliability of the TCR

The item format in the TCR is TRUE/FALSE for the pilot study, and TRUE/FALSE for Part A, MCQ for Part B for the actual study. Hence, each item in the TCR was scored dichotomously: one point for a correct response and no point for an incorrect response. The test score reliability was estimated using the Kuder-Richardson formula (Kuder & Richardson, 1937; cited in Mehren & Lehmann, 1973). Kuder-Richardson formulae are for estimating the reliability of a test based on inter-item consistency and requires only a single administration of the test.

Pilot test of the 50-item TCR with a similar sample of student (n=57) gave a KR-20 of 0.31. Apparently, this figure indicated that the new instrument had low reliability. Descriptive statistics of test scores for the 50-item TCR (mean=27.14; standard deviation=3.73; variance=13.87) showed the low reliability coefficient was probably due to the variance of the test scores being small. However, in this study, the TCR is essentially a diagnostic instrument. The main purpose of administering the TCR was to identify students' alternative conceptions of chemical representations. Hence, in the actual study, test scores reliability of the TCR was not estimated using the usual K-R 20 formula for dichotomous items. Instead, test-retest reliability was estimated. The statistical procedure used to examine test-retest reliability is correlation. Since total test scores for the TCR is continuous, the Pearson correlation coefficient is the statistic used to reflect test-retest reliability. A high correlation coefficient indicates the instrument is stable over time. For the pilot study, test-retest with a smaller sample (n=33) after a 3-week interval gave a correlation coefficient of r=0.82. In the actual study, test-retest with a random sample (n=45) after a 1-month lapse gave a correlation coefficient of 0.64. These figures suggested that the scores on the TCR were relatively stable over time. *Appendices 13 & 13a* showed the correlation coefficients and scatter plots of test-retest scores of the TCR for the pilot study and the actual study, respectively.

4.2.3 The Test on Representational Competence (TRC)

For the purpose of investigating Form four students' representational competence in chemistry, a paper-and-pencil test – the Test of Representational Competence (TRC) was administered. The TRC was used to collect data for research questions (i) (c) and (iv). These were:

Research Question (i) (c)

What are Form four students' overall levels of representational competence in chemistry?

Research Question (iv)

What are the learning difficulties demonstrated by Form four students when interpreting and using chemical representations to express chemical ideas?

The TRC was designed by the researcher in this study. The test was divided into two parts: Part A and Part B. The draft TRC contains 25 multiple choice items in Part A and 7 short response format items in Part B (see *Appendix 15*).

4.2.3.1 Development of the TRC

The section on the development of the TRC provides a detailed description of the representational skills assessed in this study, stages and procedure involved in the construction of the TRC, scoring procedure of the TRC, pilot study of the TRC, and reconstruction of the TRC.

Concepts and abilities tested

Although the TRC was mainly a test of application and skills in interpreting and using chemical representations, knowledge and understanding of basic chemical concepts and chemical representations were needed to answer the test items. These chemical concepts had been determined based on the content area selected for this study (see Section 1.9). A table of specification for the TRC has also been prepared (*Appendix 14).*

Representational skills assessed

Although in practice, representational competence covers a wide range of skills and practices, the sample in this study were novices to chemistry. Hence, it was only appropriate that the representational skills assessed be confined to five representational skills only. These skills were: (i) the ability to interpret meanings of chemical representations; (ii) the ability to translate between different representations at the same level; (iii) the ability to translate between different representations across levels; (iv) the ability to use representations to generate explanations and (v) the ability to make connections between representations and concepts (*Appendix 14*).

The representational skills assessed had been determined after careful examination of related curriculum and instructional materials such as the Form four Chemistry Syllabus and Curriculum Specifications, school textbooks and past examination papers. Therefore, although some of the test items may appear unfamiliar compared to most text book questions, the main framework for the TRC was based on the Malaysian Chemistry Syllabus (Curriculum Specifications, Chemistry Form four, 2006).

Construction of the TRC

A survey of literature such as Journal of Chemical Education, Journal of Research in Science Teaching, School Science Review; Curriculum and instructional materials such as Form four Chemistry syllabus and Curriculum Specifications, school textbooks, SPM past-year examination papers; a well as past-year papers for both the Australian National Chemistry Quiz (ANCQ), National Chemistry Quiz organized by the Malaysian Institute of Chemistry (IKM), online search and many more were carried out in order to prepare the pilot version of the TRC. Questions selected from various sources were modified appropriately where necessary while some items were generated by the researcher.

The pilot version of the TRC comprised a total of 32 questions in two parts. Part A contained 25 multiple-choice items while Part B contained 7 short answer format items. Sources of the items adopted or adapted were shown in brackets (see *Appendix 15*). Items with no sources quoted were designed by the researcher, based on classroom experience such as students' common errors.

The initial draft of the TRC was given for validation to two experienced secondary school chemistry teachers and a university Professor (same panel of reviewers as for the TCC and TCR). Their feedback was used to fine tune the structure of the TRC.

Scoring procedure for the TRC

The TRC was a two-part paper-and-pencil test. Part A comprised 25 multiple-choice questions while Part B comprised 7 short answer items.

For Part A, each correct answer was awarded one point while no point was given for an incorrect answer. Hence, total possible test point awarded for the 25 dichotomous items was 25. For Part B, one point was awarded for each part of the correct answer. Total possible test points awarded was 15 (see *Appendix 14, 15, 15a*).

Test score for the TRC was the sum of total test points for Part A and Part B. Therefore, test score may range from a minimum of 0 to a maximum of 40 points.

Pilot study of the TRC

For the TRC, specifically, the main objectives of pilot-testing the instrument were to find out the time required for the subjects to complete the test, the difficulty level, as well as the discriminating power of each of the item. During the pilot study, the subjects took not more than 60 minutes to complete the test. Item analysis had been done by computing the difficulty index (p) and the index of discrimination (ID) for each item (see *Appendix 16* and Tables 16a, 16b).

The difficulty indices of the items, which ranged from 0.10 to 0.93, provided a wide range of difficulty in the items. 15 of the 25 multiple choice items were moderately difficult, with p values ranging from 0.30 to 0.70 (Table 16a, *Appendix 16*). Besides, the discrimination indices of the items ranged from 0.31 to 0.94 for 20 of the 25 items, with 18 of the items having ID > 0.40 (see Table 16b, *Appendix 16*). Two items with low ID had low p values too. These were: item 19 (p=0.12, ID=0.13), and item 22 (p=0.10, ID=0.00). In fact, these were difficult items even many good students could not answer. These two items would not be deleted as they were considered very good items to test for students' ability to translate from the symbolic level to the sub-microscopic level. Hence, taken together, the p values and ID indices of the items showed that the MCQ items in the TRC were good items for a norm-referenced test.

Reconstruction of the TRC

Based on the data in *Appendix 16 and Tables 16a, 16b*, items with undesirable difficulty index or discrimination index were either reconstructed or discarded in order to increase the reliability of the test score. Such fine-tuning procedures were necessary to increase the reliability of the new instrument. *Appendix 15a* shows the reconstructed TRC used in the actual study.

4.2.3.2 Validity of the TRC

Since the TRC was not an existing instrument, appropriate and sufficient evidence were gathered to help establish the validity of the new instrument. During the course of developing the TRC, attempts had been made to provide evidence for face validity and content-related validity.

To check for face validity and content-related validity, the initial draft of the TRC had been reviewed by two experienced secondary school chemistry teachers and a university professor specializing in chemical education. The panel of reviewers was the same as for the TCC and the TCR. Feedbacks showed the test appeared difficult for Form four students and could be time-consuming too. After a careful review of the TRC, Section B had been restructured while Section C had been removed.

The TRC was considered an achievement test and focused on representational skills. Therefore, it was considered important to establish the content validity of the instrument. This was done by: (i) predetermining the representational skills assessed in the study, (ii) preparing a table of specification for the test (*Appendix 14*).

4.2.3.3 Reliability of the TRC

Since each item in Part A of the TRC was scored dichotomously, the test score reliability was estimated using the Kuder-Richardson formula (Kuder & Richardson, 1937, as cited by Mehrens & Lehmann, 1973). Kuder-Richardson formulas are for estimating the reliability of a test based on inter-item consistency and require only a single administration of the test.

Pilot test of the TRC with a similar sample of students (n=60) in a secondary school in Kinta District gave a KR-20 of 0.96 for the 25 MCQ items in Part A (see *Appendix 17*).

All items in the instrument were marked by the researcher herself. However, to ensure reliability of marking, 25% of the scripts (n=15) were randomly selected from the total scripts and were given to two experienced chemistry teachers for cross validation. As a check on the level of agreement between raters of the TRC scores across the 7 items in Part B of the TRC, Cohen's Kappa, an index of inter-rater reliability that corrects for chance agreement between raters, were computed. For the pilot study, the k values obtained were 0.845 (rater 1*rater 2), 0.769 (rater 1*rater 3) and 0.920 (rater 2*rater 3). These k values indicate a high level of agreement between the raters (see *Appendix 17*). The average k value is 0.845. [Rater 1 was the researcher in this study].

Further discussions with rater 2 and rater 3 were subsequently held to refine the scoring procedures. A final version of the marking scheme was ascertained. It was also agreed that some of the words in Part B item No. 3 might be problematic. Hence, in the actual study, the sentence "You can answer using words or drawings or both" was changed to "Answer using drawings only". The word "labeled" in item No. 7 (a) was also deleted (see *Appendix 15 & 15a*).

In the actual study, values of KR-20 of 0.81 and 0.87 were recorded for the 25 multiple choice items in Part A and the 7 short answer items (15 points) in Part B, respectively. KR-20 for all the 40 items of the TRC was also computed. A value of 0.90 was recorded (*Appendix 17a*). These KR-20 values indicated that the new instrument (TRC) has very high reliability. Additionally, Cohen's Kappa was also computed for the 7 items in Part B of the TRC. To ensure reliability of marking, 25% of the scripts (n=96) were randomly selected from the total scripts and were given to two experienced chemistry teachers for cross validation. To further ensure consistency in the marking, the inter-raters in the pilot study and the actual study were the same. The k values obtained were 0.795 (rater 1*rater 2), 0.807 (rater 1*rater 3) and 0.989 (rater 2*rater 3). [Rater 1 was the researcher in this study]. The average k value is 0.864. These k values indicate a high level of agreement between the raters (*Appendix 17a*).

4.2.4 The Classroom Test of Scientific Reasoning (CTSR)

In this study, the developmental level of the subjects was determined by their CTSR score. The CTSR was a 12-item, 2-tier multiple choice paper-and-pencil test designed to assess students' ability to conserve weight and volume, separate variables, use of proportional logic, combinatorial reasoning and correlations (*Appendix 18*).

Since every correct item was awarded one point while no point was given for a wrong response, the minimum and maximum attainable score was 0 and 24 respectively. The subjects were categorized into three Piagetian developmental levels of concrete operational (scores of 0 to 7), transitional (scores of 8 to 16), and formal operational (scores of 17 to 24), based on the criteria used by Lawson (1978). The test was a modified version of Lawson's Classroom Test of Scientific Reasoning (Lawson, 1978, 2000). The modified test contained 6 of the original 15 items.

The original items were based on the Piagetian tasks and involved conservation of weight and displaced volumes, the identification and control of variables and proportional, probabilistic, correlational and combinatorial reasoning (Lawson, 1978). The original test items were constructed for the classroom test. Each item involved a demonstration using some physical materials and/or apparatus. For each item, the demonstration was used to pose a question or call for a prediction. The students responded in writing in individual test booklets. The booklets contained only the questions followed by a number of possible answers. Students were instructed to respond by checking the box next to the best answer and then explaining why that answer was chosen.

The revised version (Lawson, 2000) used in this study (see *Appendix 18*) was a multiple-choice, paper-and-pencil test that could be administered to a large group of respondents. No demonstration was involved. For each item, the subjects were required to respond by selecting the correct answer from a list of five options given in the question booklet. Table 4.5 gives a summary of the items that were used in this study.

Table 4.5
CTSR item summary

Type of reasoning ability	Item No.
Conservation of mass	1, 2
Conservation of displaced volume	3, 4
Proportional thinking	5, 6
Advanced proportional thinking	7, 8
Identification and control of variables	9, 10
Identification and control of variables, and probabilistic thinking	11, 12, 13, 14
Probabilistic thinking	15, 16, 17, 18
Correlational thinking (includes proportions and probability)	19, 20
Hypothetico-deductive thinking	21, 22, 23, 24

4.2.4.1 Scoring procedure and classification of developmental level

Every item in the CTSR consisted of two parts. It began with a problem statement and was followed by an explanation for the answer to the problem. For example, Item 1 (see *Appendix 18*) was a problem on conservation of weight. It required students to respond to the question and explain how they obtained the answer by choosing the correct answer from the options given. For each item, 1 point was awarded if the correct answer was chosen or if a correct explanation was given. No point was awarded for incorrect answers. Hence, the score for this instrument ranged from a minimum of 0 point to a maximum of 24 points. Following the criteria used by Lawson (1978, 1992), the respondents were categorized into three cognitive or developmental levels based on the CTSR score, as shown in Table 4.6.

Table 4.6
Classification of developmental level based on CTSR scores

Score Range	Developmental Level
0 – 7	Concrete operational reasoning (CR)
8 – 16	Transitional reasoning (TR)
17 – 24	Formal operational reasoning (FR)

4.2.4.2 Validity of the CTSR

The validity of the CTSR was established through rigorous steps (Lawson, 1978), in which three types of evidence were sought.

The first type of evidence concerned the face validity whereby a panel of judges responded with 100% agreement that the test items appeared to require concrete and/or formal operational reasoning. Convergent or concurrent validity was obtained by computing the Pearson product-moment correlations between the classroom test total score and level of response on the bending rods and balance beam tasks, which measured formal thought. A coefficient of 0.76 at $p<0.001$ was obtained and this relatively high correlation indicated that the classroom test had convergent or concurrent validity. For the third type of evidence of the classroom test's validity, the classroom test and all 4 interview tasks loaded heavily on the same factor supporting the hypothesis that they measured aspects of the same psychological parameter, that is, formal operational reasoning.

Thus, it was concluded that the classroom test was a valid measure of formal reasoning, concrete reasoning, and reasoning that could be considered intermediate.

4.2.4.3 Reliability of the CTSR

Reliability of the CTSR had been reported in several different studies, with Cronbach alpha reliability coefficient ranging from 0.75 to 0.81. In a study involving a sample of 189 tenth-grade students, Cavallo (1996) reported a CTSR Cronbach alpha coefficient of 0.75. When the classroom test was administered to a group of 663 undergraduates by Lawson, Alkhoury, Benford and Clark (2000) in another study, a Cronbach alpha reliability coefficient of 0.81 was obtained.

A split half reliability of 0.76 was reported when Lawson (1983) field-tested the CTSR with 96 undergraduate students. In another study, Lawson, Clark, Meldrum, Falconer, Sequent, and Kwon (2000) reported a test-retest reliability coefficient of 0.65 by comparing the CTSR scores of 667 undergraduates. When the reliability of the Bahasa Malaysia's version of the 20-item CTSR was estimated by Eng (2002) using the KR-20 formula in a study for 294 sixth form students, a coefficient of 0.51 was obtained. Nagalingam (2004) reported a reliability coefficient of 0.95 in a study involving 381 Form four science students.

In this study, test score reliability of the 24-item CTSR was estimated using the KR-20 formula for the 214 Form four science students who took the test. A reliability coefficient of 0.82 was recorded (*Appendix 18a*).

4.2.5 The Digit Span Backwards Test (DSBT)

Working memory capacity of the subjects was estimated using the DSBT. This test was part of the Wechsler Adult Intelligence Scale (Wechsler, 1955), and involved both storage and processing. The question or problem had to be understood (translated) or represented, held in memory and then manipulated (rearranged). This test not only measures memorization of data, but also combines the number of data and the operations carried out on them.

In the administration of the test, students listened to sequences of digits and had to write them in reverse orders. There were two (2) sequences of digits with two (2) digits each, two (2) sequences with three (3) digits each, and so on, up to two (2) sequences of eight (8) digits each. Students had to write the digits by filling in printed grids, with one (1) digit in each square.

The test had been used by Johnstone (1997, 2000b, 2006) and his group (Al-Naeme & Jonestone (1991); El-Banna & Johnstone (1986) in all their relevant work.

4.2.5.1 Administration and scoring of the DSBT

For this study, the conventional DST, which attempted to measure the STM capacity, was used as a practice for the students. It consisted of a procedure in which the students were asked to listen to a series of numbers and then give them back to the researcher exactly, without any processing.

The test began by reading three digits and students responded by writing in exactly the same order. This was repeated for another three-digit sequence. Then, the students were given four digits and then four others. This increased to two sets of five digits and so on until the students failed both sequences at a given level. If the students failed one and succeed in another of the same complexity, he/she was taken to have succeeded at that level. When he/she failed both at a given level, his STM was taken to be the last level at which he/she succeeds.

The actual test was the DSBT. This involved a similar procedure, but the students were asked to listen to the sequence without writing, inverted the sequence in his/her mind, and then wrote it down. The maximum number of digits that were successfully written for at least two out of three corresponding sequences was taken as the value of working memory capacity.

The scores obtained from the DSBT were not a measure of STM because holding and inversion of the sequence (processing) had taken place. The scoring procedure was exactly the same as in the original test. The DSBT scores were usually less than the STM scores because the capacity had been doing two things: holding and processing.

Ideally, both tests should be administered individually and face to face with the researcher and entirely verbally, without writing. However, the sample size for this study as relatively large and the process would be too tedious. Therefore, a group method was used where the students wrote their responses on a prepared form (see *Appendix 20*).

4.2.5.2 Validity of the DSBT

To avoid the possibility of cheating (writing the digits in reverse order from right to left, and in particular simultaneously with listening), students had to write the digits by filling in printed grids, with one digit in each square (see *Appendix 20*). To ensure that the students kept to the rules, another teacher was assigned the task of supervising during the actual test.

4.2.5.3 Reliability of the DSBT

To estimate the reliability of the DSBT, Nagalingam (2004) did a test-retest using the same DSBT (*Appendix 20*) with a sample of 100 respondents who were the subjects of the study. A reliability coefficient of 0.97 was reported.

In this study, test score reliability of the DSBT was estimated using a test-retest with a smaller sample of 56 respondents. A Pearson correlation coefficient of r=0.86 was recorded (*Appendix 20a*).

4.2.6 The Learning Approach Questionnaire (LAQ)

The LAQ is a Likert-scale instrument designed to assess students' learning orientation, ranging from meaningful to rote (Entwistle & Ramsden, 1983). The LAQ was adapted from the Approaches to Studying Inventory (ASI) devised by Entwistle and his colleagues (Entwistle & Ramsden, 1983, p.35-55). A version of the instrument adapted and employed by previous researchers was used in this study (Cavallo & Schafer, 1994; Cavollo, 1996; BouJaoude & Barakat, 2003; BouJaoude et al., 2004; Sim, 2006).

The LAQ consisted of 23 items in two subscales, with 13 items on the meaningful learning subscale and 10 items on the rote learning subscale, although the items were randomly placed within the LAQ (see *Appendix 19*). Table 4.7 shows the item to subscale key of the LAQ.

A likert scale was adopted as it is a good way of writing closed-ended questionnaire items to measure people's attitudes and opinion with intensity scale (Nardi, 2003). A 4-point Likert scale (A=Always True to D=Never True) was used for responding. The use of a 4-point scale was to overcome the tendency of respondents selecting the neutral option (BouJaoude et al., 2004; Sim, 2006).

Table 4.7
Item to subscale key of the LAQ

LAQ subscale	Item No.	Number of items
LAQ-Meaningful	1, 2, 4, 6, 8, 9, 10, 11, 13, 15, 17, 20, 21	13
LAQ-Rote	3, 5, 7, 12, 14, 16, 18, 19, 22, 23	10

To control for response set, some of the items were negatively worded, and scoring was reversed for these items (BouJaoude et al., 2004). For example: a response of (A=Always True) on Item No. 2 indicates a strong tendency towards meaningful learning, whereas a response of (A=Always True) on Item No. 5 indicates a strong tendency towards rote learning (see *Appendix 19*). Items from the LAQ-Rote subscale were reverse-scored. Hence, a high score in the LAQ represents a more meaningful learning orientation while a low score in the LAQ represents a more rote learning orientation (Cavallo, 1996).

The selected items were modified to measure approaches to learning chemistry by changing the subject matter in the item to “chemistry” (see item No. 2 and item No. 4; *Appendix 19*).

4.2.6.1 Scoring system and categorization scheme

The items in the LAQ were scored by assigning points to the option selected by the respondents for each of the item. The total number of points accumulated for the 23 items gave the LAQ score. Table 4.8 summarizes the scoring system.

Table 4.8
Scoring system for the LAQ

Option selected	Positively worded Items	Negatively worded items
A = Always True	4	1
B = More True than Untrue	3	2
C = More Untrue than True	2	3
D = Never True	1	4

Total number of item	= 23
Maximum score	= 23 x 4 = 92
Minimum score	= 23 x 1 = 23

Since the LAQ consisted of 23 items and a 4-point Likert scale was used for the responses, total possible score or maximum score is 92, while the minimum score is 23. See Figure 4.1.

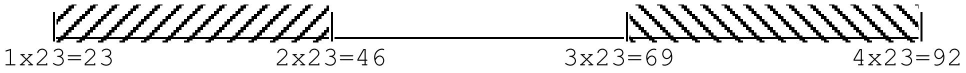

Figure 4.1: Scoring for the LAQ

For the purpose of this study, students' learning orientation had been divided into three categories: rote learning orientation, meaningful learning orientation, and the middle group. Hence, based on Figure 4.1, categorization scheme of the LAQ score was also established, as shown in Table 4.9 below.

Table 4.9
Categorization scheme of the LAQ score

Learning Orientation	LAQ score
Rote Learning Orientation	<46
The Middle Group	46 – 69
Meaningful Learning Orientation	>69

Students' LAQ scores were sorted in descending order, and divided into three categories: meaningful learning orientation, rote learning orientation, and the middle group, based on their LAQ scores as shown in Table 4.9. Students with LAQ scores higher than 69 were categorized as having a `meaningful learning orientation' while those with scores lower than 46 were categorized as having a `rote learning orientation'. Students with LAQ scores of between 46 and 69 were categorized as `the middle group' or `intermediate learners'. This rating, which was based on the LAQ score, was referred to as student self-report or student rating (BouJaode et al., 2004).

4.2.6.2 Validity of the LAQ

The LAQ had been used by previous researchers and had been shown to be a valid instrument for assessing students' learning orientation or approaches to learning. However, to further enhance its validity for this study, several measures had been taken. These included:

(i) Certain terms in the LAQ items had been changed to terms more familiar to secondary school students. For example: the word "lectures", probably more commonly used for higher education, had been replaced by "classes" (Item No. 9, *Appendix 19*).

(ii) General term such as “subject matter” had been replaced by “chemistry” (Items No. 2 and No. 4, *Appendix 19*).

(iii) Items No. 14, 16, 17, and 22 had also been rephrased using simple words as far as possible, to make the statements easier to understand.

4.2.6.3 Reliability of the LAQ

The internal consistencies of the subscales were reported by Entwistle and Ramsden (1983) as Cronbach alphas, and ranged from 0.47 to 0.78. A Cronbach alpha internal consistency coefficient for this instrument was reported as 0.77 for a sample of Grade 11 chemistry students (BouJaoude, 1992). The Cronbach alpha for a 24-item LAQ was reported as 0.54 for a separate sample of Grade 10 biology students (Cavallo & Schafer, 1994). BouJaoude et al. (2004) used the 23-item LAQ on a 4-point Likert-scale for a sample of Grade 11 chemistry students. Alpha coefficient for the LAQ was reported to be 0.60. Sim (2006) used a modified version of the 23-item LAQ on a 4-point Likert-scale for a sample of 168 Form 4 science students and reported an alpha coefficient of 0.58.

In this study, Cronbach alpha coefficient of 0.62 was recorded with the 23-item LAQ for a sample of 211 Form four science students. Alpha coefficient for the meaningful learning subscale (13 items) was 0.77 while that of the rote learning subscale (10 items) was 0.47. See *Appendix 19a*.

4.2.7 The Interviews

Two semi-structured interviews (SSI 1 and SSI 2) were conducted to gain further insights and understanding into selected students’ representation of basic chemical concepts. A small, purposive sample (n=9) constitutes the participants of the interviews. Interview Protocols 1 and 2 (*Appendix 21*) were used as the guidelines to conduct the two interviews. Interview SSI 1 was held for each of the participant. Subsequently, interview SSI 2 was conducted using the same sample.

4.2.7.1 The Interview Sample

A smaller, purposive sample consisting of nine students in three categories were the participants for the two interviews. The three categories were: students with high, average and low test scores for the TRC (see Table 4.10). In order to avoid bias on grounds of their categorization, they were not informed as to whether they were from the high, average or low category.

Table 4.10
Profile of the interview participants

Participant's ID	Scores					
	TCC	TCR	**TRC**	CTSR	LAQ	DSBT
H1	25	26	**40**	22	65	8
H2	22	21	**39**	18	57	8
H3	24	21	**39**	20	80	8
M1	12	19	**17**	8	66	8
M2	15	17	**16**	9	60	8
M3	13	17	**17**	2	63	8
L1	7	15	**6**	10	57	8
L2	6	10	**8**	5	55	5
L3	6	13	**4**	6	54	8

Note: H=High; M=medium; L=Low

TCC=Test of Chemical Concepts, TCR=Test of Chemical Representations, TRC=Test of Representational Competence, CTSR=Classroom Test of Scientific Reasoning, LAQ=Learning Approach Questionnaire, DSBT=Digit Span Backwards Test

4.2.7.2 Choice of Interview Type

A semi-structured interview approach was used in preference to the highly structured or unstructured interviews. This is because in semi-structured interviews, the open-ended questions which are fairly specific in its intent, are phrased to allow for individual responses, provides a high degree of objectivity and uniformity, yet allow for probing and clarification (McMillan & Schumacher, 1993).

In addition, individual interview is preferred compared to focus group interview (FGI). Although FGI is more economical in terms of time and resources as large amount of data can be gathered within a limited time (Patton, 2002), the interview may be dominated by one or two individuals. Besides, in this study, the focus is on cognition: probing on mental process of individual, not social interaction.

4.2.7.3 Purposes of the Interview

Another important source of data in this study is from interviews although interview is a common data collecting technique in qualitative research (Merriam, 1998). While paper-and-pencil tests can only assess performance, interviews allow researcher entering the interviewees' perspective and to find out what is in their mind (Patton, 2002).

Conventionally, interview is a face-to-face conversation with a purpose between two unacquainted individuals, that is: the interviewer, who asks questions, and the interviewee or respondent, who provides the answers (Gubrium & Holstein, 2002, p.57; cited in Chien, 2006). However, in this study, the semi-structured interviews involved more than face-to-face conversation. Participants' drawings, the use of worksheets, online quiz, model-building kit, enabled multiple sources of data to be collected. Triangulation of data increased the validity and reliability of the findings.

Two semi-structured interviews (SSI 1 and SSI 2) were conducted, through which the researcher would be able to probe further into students' conceptions of chemical representations (SSI 1) and representational competence in chemistry (SSI 2). It was hoped that by probing further through in-depth interviews, the researcher could gain a better insight and understanding into students' representations of basic chemical concepts.

4.2.7.4 The Interview Protocols

Interview Protocol 1 focused on aspects such as: symbolic representations and submicroscopic representations while Interview Protocol 2 focused on student-generated representations and multiple levels of representations. See *Appendix 21*.

Questions based on TCR and TRC for the interviews were selected after a careful analysis of students' responses to the items in both the TCR and the TRC. These items were only identified after the administration of both the instruments in the actual study.

4.2.7.5 Pilot Study of the Interview

The interview protocols were pilot-tested with three subjects, one from each category. Pilot-testing was necessary as a check for bias in the procedures, the interviewer, or the questions; provided a means of assessing the length of the interview and gave the researcher some idea of the ease with which the data could be summarized. Any cues suggesting that the participant could not fully understand the question would be noted. Weaknesses identified in the interview protocols were corrected. During the pilot study, it was discovered that: (i) the interview protocols were generally comprehensible, (ii) the average participant took the longest time to complete the interview, with long pauses after each question, and needed the most probing, (iii) the good participant took the shortest time to complete the interview and took the initiative to ask for clarification to some questions, (iv) the poor participant needed guidance to answer both Worksheets 1 and 2.

After the pilot study, slight modifications were made to further improve the interview protocols for the actual study (see *Appendix 21*).

4.3 Data Collection

Data collection involved three main procedures. These were: (i) preliminary procedures, (ii) administration of the TRC, TCC, TCR, CTSR, DSBT, and LAQ, (iii) the interviews.

4.3.1 Preliminary Procedures

Data collection for this study began after permission had been granted by the relevant authorities. These included: (i) Approval from Faculty of Education, University of Malaya, (ii) Permission from the Educational Planning and Research Division (EPRD) of the Ministry of Education, Malaysia, (iii) The State Education Department of Perak, and (iv) Principals of the seven selected schools in this study. Besides, Letters of Information and Consent were also forwarded to school principals and interviewees selected for the semi-structured interviews. See *Appendices 27 (a) to (d).*

Data collection was carried out between September 2008 and July 2009. The duration for data collection was about nine months, excluding the two months year-end school holidays. The entire data collection was conducted by the researcher, with the help of assistant test administrators for the Digit Span Backwards Test.

Data collection was conducted in three stages. Stage (i): administration of the TRC, TCC, and TCR; Stage (ii): administration of the CTSR, DSBT and LAQ; Stage (iii): the interviews.

4.3.2 Administration of the Tests and Questionnaire

The three main paper-and-pencil tests, the Test on Chemical Concepts (TCC), the Test on Chemical Representations (TCR), and the Test on Representational Competence (TRC) were administered to all the 411 subjects of the study. Subsequently, all the other paper-and-pencil tests (CTSR, DSBT) and the questionnaire (LAQ) were administered. Administration of all tests and questionnaire was done under a standardized whole class setting. Table 4.11 shows the maximum time allowed for each of the test or questionnaire.

It was believed that long testing period could cause fatigue, affecting performance. Hence, the tests or questionnaire were administered in three separate sessions on different days to avoid fatigue among the participants (see Table 4.11).

Slight modifications were made to the maximum time allowed for the TCR and TCC in the actual study as the changes were deemed necessary for the reconstructed, final version of these two tests.

It was discovered during the pilot study that the subjects were most serious when the first test was administered. Hence, instead of administering the two tests (TCC and TCR) in the first session and TRC in the second session as were done during the pilot study, in the actual study, the sequence was reversed. See Table 4.11. This was done as TRC was considered the most important instrument for this study.

Table 4.11
Administration of tests and/questionnaire (actual study)

Test/Questionnaire	Time (minutes)	Session No.
1. Test on Representational Competence (TRC)	60	1
2. Test on Chemical Concepts (TCC)	30	2
3. Test on Chemical Representations (TCR)	30	2
4. Classroom Test of Scientific Reasoning (CTSR)	30	3
5. Learning Approach Questionnaire (LAQ)	15	3
6. Digit Span Backwards Test (DSBT)	20	3

During the pilot study, it was also discovered that some of the subjects wrote down the digits as they were being read and not after the series of digits was read. Besides, for the DSBT, some of them wrote the digits from right to left instead of inverting them in their heads. Hence, in the actual study, the help of an assistant test administrator was sought to ensure the participants followed proper test procedure.

The participants were told that the survey questionnaire and the tests were used for research purposes only and that the results would be kept confidential. The question papers, together with the answers were collected by the researcher at the end of the sessions. Any request by any participant to bring back the test paper or the questionnaire was not entertained.

4.3.3 The interviews

Following the administration and scoring of all the paper-and-pencil tests and questionnaire, interviews were subsequently conducted to gain better insights and understanding into selected students’ representations of basic chemical concepts.

Two semi-structured interviews (SSI 1 and SSI 2) were conducted, through which the researcher would be able to probe further into students' conceptions of chemical representations (SSI 1) and representational competence (SSI 2). See *Appendix 21* for Interview Protocol (1) and (2).

Purposive sampling procedure was used to select a representative sample of participants for the interviews (n=9). Only those students who were willing to be interviewed were included in the interviews. See *Appendix 27d* for the letter of information and consent.

Each student was interviewed individually for about 30 to 45 minutes for the semi-structured interview 1, using Interview Protocol 1 as a guideline (*Appendix 21).* The semi-structured interview 2 also took about 30 to 45 minutes, and Interview Protocol 2 was used as a guideline (*Appendix 21*). The interviews were conducted by the researcher in this study in a quiet room provided by the school authorities. The participants were informed of the purposes of the interviews and given the assurance that their responses would be kept confidential.

The participant was asked one question at a time, and each time, the interview protocol was used as a guideline, with modifications where necessary:

Apart from worksheets, focus cards, molecular modeling set, and a laptop with internet access, plain paper, pencil, and other relevant materials were also provided. The participants were told to read out aloud whatever they had written or scribbled on their paper. This was done because it was believed that a good way to get learners to think about chemistry as to get them to talk about it (Schmidt, 1984).

In order to get good data, it is important to ask good question in a language which is familiar to the interviewees and clearly understood by them (Merriam, 1998). Hence, participants were allowed to choose to converse in the language they are most comfortable with (Bahasa Melayu, English, or Mandarin), and free to ask for clarification and translations. Besides, to get meaningful data and keep the interview going, the researcher needs to establish an extended, open-ended exchange relationship with the interviewees (Gubrium & Holstein, 2002; cited in Chien, 2006). Further more, according to Merriam (1998), an interviewer needs to behave as a sensitive instrument, flexible, make adjustments if necessary during the interview, having good probing skills to explore more details, establish a good rapport with the interviewees and equally important, avoid pushing too hard and going too fast with the interviewees.

In this study, the interviewer and the interviewees were well acquainted as the researcher herself was the interviewer, the test administrator for all the paper-and pencil tests (TCC, TCR, TRC, CTSR, DSBT), as well as the questionnaire (LAQ).

In order to understand how students think or to capture their thinking, the interviewer needs to use a language which the interviewees are familiar with. In this regard, the interviewer's 24 years experience as a chemistry teacher in four different states in Malaysia would be helpful. Familiarity with the chemistry curriculum and content also facilitate better understanding on the interviewees as Patton (2002) believed that researcher's personal experiences and insights are important part of the inquiry and critical to understanding interviewees' thinking.

At the beginning of the interview, interviewees were told that the interviews would be audio-recorded and that it was important for them to speak out loud whatever came to their minds. All interview sessions were audio-recorded with the permission of the interviewees (Bogden & Biklen, 2003; cited in chien, 2006). A hand phone with voice recording function was used as an audio recorder.

Recording is an essential part of the interview because it increases the accuracy of the data collected and allows the interviewer to focus on the conversation instead of busy writing or recording manually (Patton, 2002, p.380).

A digital audio recorder has advantage over the traditional audio tape recorder for recording a long interview without changing the tape manually. Recording can then be transferred directly into a computer via blue tooth service and played back using programs such as real player. To avoid accidental loss or deletion of the interview records, copies of the recording were sent to the researcher's Inbox for safe storage and convenient retrieval. In addition, while transcribing the interview verbatim, the audio recording can be paused, replayed, and the volume can be controlled.

To enhance the quality of the interviews, it is necessary to review the interview transcripts, replay the audio recording, examine and analyze the drawings of the participants immediately after the interview sessions for ambiguity or uncertainty and make necessary clarification with the interviewees (Patton, 2002). Hence, it is crucial to get the contact number of participants, especially in cases where getting back to the participants in person is problematic. Researcher should also jot down ideas, interpretations, and other relevant findings from earlier analysis for review purposes and for future improvement in the coming interviews.

After the interviews, the audio-recordings were sent to a laptop via blue-tooth. Play back was done using the program "real player". Each interview was transcribed and interpreted. The worksheets, drawings and rough papers used by the participants as well as all their test papers were kept as these form part of their output and were essential documents in data analysis.

4.4 Data Analysis

All quantitative data were processed and analyzed using the Statistical Packages for the Social Sciences, SPSS. Both descriptive and inferential statistics were employed. An alpha level of 0.05 was used for all the statistical tests.

Research Question (i): The TCC and TCR test scores were measures of students' overall levels of understanding of chemical concepts and chemical representations respectively while the TRC test score measured their overall levels of representational competence in chemistry. Hence, the mean, standard deviation, minimum and maximum of the TCC, TCR and TRC scores were computed and tabulated.

Research Question (ii): To test whether there were significant differences between students with high, medium, and low overall levels of understanding of (a) chemical concepts, and (b) chemical representations, in their overall levels of representational competence in chemistry, one-way analysis of variance (ANOVA) was used to compare the mean TRC scores in order to determine if any significant differences existed among the three groups of students with different levels of (a) TCC scores, and (b) TCR scores, in their representational competence in chemistry. Separate cumulative frequency curves for the two test scores (TCCt scores and TCRt scores) were plotted and ranges for the low, medium and high groups for each test were determined based on quartiles of the respective test scores (see *Appendix 28*). The ranges that denoted the lower and higher 25% of the students became the range for the low and high groups, respectively. The middle 50% became the range for the medium group (Heitzman & Krajcik, 2005). Parametric assumptions for ANOVA such as normality, homogeneous variance and independence were made in this study. If the *F* ratio for ANOVA was significant, multiple comparison tests were subsequently conducted to determine which pair of means difference was statistically significant. The post hoc Scheffe test, which can be used for any combination, was used in this study.

Research Question (iii): For both the TCC and TCR, the frequency and percent correct responses for each of the items were computed and tabulated. Alternative conceptions and percent respondents were then identified, critically examined, and analyzed.

Research Question (iv): Responses for each of the 25 MCQ items in Part A of the TRC were analyzed and the response pattern was tabulated and interpreted. Answers for the 7 short answer format items in Part B were also analyzed and students' difficulties interpreting and using representations were then identified and documented.

Research Question (v): The nine interviews, which were audio-taped, were also transcribed verbatim and interpreted. Responses to the items in the TCR and TRC selected by the respondents, as well as their responses to questions in the Interview Protocols, were used as guidelines in the analysis of the interview data. New findings during the interviews were also noted. Excerpts from the interview transcripts were used to gain further insights into students' conceptions of chemical representations and their representational competence in chemistry.

Research Question (vi): Bivariate correlation coefficients were explored to examine the relationship between students' prior knowledge, developmental level, working memory capacity, learning orientations, and their representational competence. The Pearson Product-moment correlation coefficient was used as a measure of correlation between TCC, TCR, CTSR, DSBT, LAQ scores and TRC scores. It is a statistic descriptive of the magnitude and direction of the relationship between two metric variables. The assumption of linearity of regression was made. The statistical significance of the correlation coefficient obtained was tested at the 0.05 level to determine if this value was a chance difference from a zero correlation in the population from which the scores were sampled was made.

Research Question (vii): Multiple regression analysis was subsequently employed to examine the influence of each cognitive variable on the overall levels of representational competence and subsequently, to determine which cognitive variable was the best predictor variable for representational competence. Finally, the regression model with representational competence as the criterion variable was generated.

4.5 Chapter summary

Table 4.12 provides a summary of the methodology of the study.

In Chapter 5, findings of the study will be presented and discussed.

Table 4.12
Summary of Methodology

Research Question	*Data Collection (Instruments/Sources of data)*	*Data Analysis*
1. Overall levels of: (a) understanding of basic chemical concepts (b) understanding of chemical representations (c) representational competence in chemistry	Test score for the: (a) Test on Chemical Concepts (TCC) (b) Test on Chemical Representations (TCR) (c) Test on Representational Competence (TRC)	Mean, median, SD, variance, minimum, maximum.
2. Comparing students of different overall levels of understanding of: (a) chemical concepts, and (b) chemical representations, in their overall levels of representational competence	Test score for the: (a) Test on Chemical Concepts (TCC) (b) Test on Chemical Representations (TCR) (c) Test on Representational Competence (TRC)	Subgroup comparison 1-way ANOVA *Ranges for the Low (L), Medium (M) or High (H) groups based on quartiles of test scores.*
3. Alternative conceptions of: (a) basic chemical concepts (b) chemical representations	Students' responses to items in (a) the TCC (b) the TCR	Analysis of items & responses (frequency, %)
4. Difficulties in interpreting and using representations	Students' responses to items in the TRC	Analysis of items & response patterns (frequency, %)
5. To gain further insights into selected students' conceptions of chemical representations and their representational competence in chemistry.	Semi-structured Interviews (SSI) Interview Protocols 1 & 2 Worksheets, Focus cards, Model building kit, Online quiz, video clips.	Analysis of interview data
6. Possible cognitive variables influencing representational competence (prior knowledge, developmental level, working memory capacity, learning orientations)	Test on Chemical Concepts (TCC) Test on Chemical Representations (TCC) Classroom Test of Scientific Reasoning (CTSR) Digit Span Backwards Test (DSBT) Learning Approach Questionnaire (LAQ)	Pearson correlation
7. Best predictor of representational competence	Same as above (6)	Multiple regression analysis