

**Accuracy of chemistry performance evaluation of BSc Four-year
programme students: a case study**

by

KGADI CLARRIE MATHABATHE

**Submitted in partial fulfilment of the requirements for the degree
of Master of Science in Science Education**

**In the Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria**

2011

**Supervisors: Prof. M. Potgieter
Dr. S. Human-Vogel**

ABSTRACT

The ability to make realistic judgements of one's performance is a demonstration of the possession of strong metacognitive skills. Metacognition involves the monitoring of one's progress during learning, and the ability to modify learning strategies for increased effectiveness. Poor-performing students are at risk because they generally exhibit high levels of overconfidence when evaluating their performance, and may fail to adjust their learning strategies in time. This study aims to explore the accuracy with which students in the BSc Four-year programme (BFYP) of the University of Pretoria evaluate their performance in a stoichiometry test, as well as the influence of teaching on test performance and on accuracy of performance evaluation. The factors that students rely on when making performance evaluations as well as shifts in the reliance on these factors after teaching are explored. Finally, the study examines the relationship between bias in performance evaluation and the self-protection, self-enhancement motivational factors and gender.

Data were collected by means of a three-tier stoichiometry test instrument, administered as pre- and posttest, as well as a questionnaire administered simultaneously with the pretests to a sample of 91 students. Each test item comprised a stoichiometry question, a confidence rating and a free-response explanation for the choice of confidence rating. The confidence rating was interpreted as an indication of expected performance. The test instrument allowed for the investigation of bias in performance evaluation in the pre- and posttests, the exploration of factors that students rely on when making performance evaluations and how the reliance on these factors shifted in the posttests. The questionnaires were used to collect data on self-enhancement, self-protection and gender. The study shows that the majority of the students were overconfident in the evaluation of their performance in both the pre- and posttests. Performance improved significantly in the posttest but accuracy of performance evaluation did not.


Students were categorised as overconfident (OC), realistic (R) or under-confident (UC) based on the difference between actual and expected performance. Five subgroups were defined on the basis of accuracy of performance evaluation in the pre- and posttests. The five subgroups, labelled first by their pretest and then their posttest category, were the OC-OC (50 students), OC-R (13 students), R-R (11 students), R-OC (15 students) and the R-UC (2 students) subgroups. The results indicated no significant difference between the pre-knowledge and

ability of the students in the four main subgroups. The students differed significantly in terms of performance in the posttest, their pre- and posttest average confidence scores and in performance gain. A significant difference was not found with regard to performance in the CMY 143 end of semester examination. These findings confirmed that we were dealing with four discrete subgroups with different characteristics. The OC-R subgroup achieved the highest learning gain by a significant margin. Moderate learning gains were demonstrated by the R-R and OC-OC subgroups and the R-OC subgroup did not achieve any learning gain at all. Careful analysis of qualitative data revealed that accuracy in the evaluation of posttest performance was associated with both a reduction in the prevalence of vague subjective judgments and with higher performance gain. Similarly, an increase in the tendency to base metacognitive monitoring on vague global judgments of performance in the posttest was associated with reduced accuracy of self-evaluation and lower learning gain. The tendency by the four performance evaluation subgroups to self-enhance or self-protect was not found to be statistically different. P-values greater than 0.05 in the pre- and posttests indicated that males and females were not significantly different in their accuracy of performance evaluation.

The study suggests that an element of bias in performance evaluation may be beneficial to learning. Inaccuracy in self-evaluation in the pretest did not hamper learning for both the OC-OC and OC-R subgroups. Students who were over-optimistic about their performance in the pretest may have been less intimidated by the challenges of the new content material than those who were better calibrated (R-R and R-OC subgroups). Students who remained overconfident in the posttest, i.e. in the OC-OC subgroup did not gain from the learning experience as much as those who entered overconfident but became better calibrated. Those who entered tentatively as realists and then, with a little exposure, became unrealistic in their performance evaluation were shown to be the most vulnerable based on their lack of learning gain. Furthermore, increasing content knowledge alone may not be enough to raise the metacognitive ability of students. Finally, chemistry educators should be aware that students often make vague subjective judgements of performance even on a topic like stoichiometry, which requires predominantly procedural knowledge and formal reasoning. Our study has shown that this deficiency, when associated with poor accuracy of self-evaluation, may hamper learning gain.

DECLARATION

I, Kgadi Clarrie Mathabathe, declare that the dissertation, which I hereby submit for the degree Master of Science in Science Education at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: 

DATE: 03 May 2011

DEDICATION

This dissertation is dedicated to my son, Neo Tshegofatso Mathabathe and my husband, Neo Lucas Mathabathe, for his support and understanding and for taking care of our son allowing me to concentrate on the completion of the study.

ACKNOWLEDGEMENTS

I am grateful to God the Almighty for His protection, guidance, strength and wisdom.

I would like to thank my supervisors, Professor Marietjie Potgieter and Dr. Salome Human-Vogel. The success and completion of this study would have not been possible without their guidance, constructive criticism and encouragement. I am grateful for their invaluable input and support which led to the final research report.

Special thanks, to the BFYP staff for their support and advice on various aspects of the study. My sincere thanks to the students and educators who willingly participated in the study.

I would also like to acknowledge the financial aid and support made by the Canon Collins Trust and the Graça Machel scholarship for women which resulted in the completion of this study.

TABLE OF CONTENTS	PAGE
ABSTRACT	i
DECLARATION	iii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	vii
LIST OF TABLES	ix
LIST OF APPENDICES	xi
LIST OF ACRONYMS	xii
Chapter 1 Introduction to the study	1
Chapter 2 Literature review	10
Chapter 3 Research design and methodology	39
Chapter 4 Results and discussion	64
Chapter 5 Conclusions and recommendations	129
REFERENCES	151
APPENDICES	158

LIST OF FIGURES	PAGE
Figure 2.1 An overview of monitoring control processes	14
Figure 2.2 An overview of metamemory components during the retrieval stage	17
Figure 2.3 A concept map showing metacognitive monitoring processes we propose are experienced by students during test-taking	21
Figure 2.4 Conceptual framework of the study	28
Figure 2.5 Three levels of chemistry	34
Figure 3.1 Embedded mixed methods design	44
Figure 3.2 Embedded experimental model	44
Figure 3.3 Embedded Experimental model used in the study	45
Figure 3.4 Overview of the study	53
Figure 4.1 Actual versus perceived scores of students in their quartile ranks	79
Figure 4.2 Expected versus actual pretest performance of students in the four performance quartiles	80
Figure 4.3 Expected versus actual posttest performance of students in the four performance quartiles	81
Figure 4.4 Scatterplot of pretest average confidence scores against pretest scores categorised by pre-post performance evaluation subgroups	84
Figure 4.5 Scatterplot of posttest average confidence scores against pretest scores categorised by pre-post performance evaluation subgroups	85
Figure 4.6 Boxplot showing a comparison of the four pre-post performance evaluation subgroups in terms of performance in the first semester chemistry module (CMY 133)	89
Figure 4.7 Boxplot showing a comparison of the four pre-post performance evaluation subgroup in terms of pretest scores	91
Figure 4.8 Boxplot showing a comparison of the four pre-post performance evaluation subgroups in terms of posttest scores	91
Figure 4.9 Boxplot showing a comparison of the four pre-post performance evaluation subgroups' pretest average confidence scores	93

LIST OF FIGURES	PAGE
Figure 4.10 Boxplot showing a comparison of the four pre-post performance evaluation subgroups in terms of posttest average confidence scores	95
Figure 4.11 Boxplot showing a comparison of the four pre-post performance evaluation subgroups in terms of CMY 143 performance	96
Figure 4.12 Boxplot showing a comparison of the four pre-post performance evaluation subgroups' average gain in performance	97
Figure 4.13 Scatterplot showing the relationship between bias in performance evaluation in the pretest and SEa scores	100
Figure 4.14 Scatterplot showing the relationship between bias in performance evaluation in the posttest and SEa scores	100
Figure 4.15 Scatterplot showing the relationship between bias in performance evaluation in the pretest and SEb scores	101
Figure 4.16 Scatterplot showing the relationship between bias in performance evaluation in the posttest and SEb scores	101
Figure 4.17 Scatterplot showing the relationship between bias in performance evaluation in the pretest and SP scores	102
Figure 4.18 Scatterplot showing the relationship between bias in performance evaluation in the posttest and SP scores	102
Figure 4.19 Boxplot showing a comparison of the self-enhancement levels (SEa) of the four pre-post performance evaluation subgroups	104
Figure 4.20 Boxplot showing a comparison of the self-enhancement levels (SEb) of the four pre-post performance evaluation subgroups	105
Figure 4.21 Boxplot showing a comparison of the self-protection levels (SP) of the four pre-post performance evaluation subgroups	106
Figure 5.1 Question 20 of the stoichiometry test instrument shown in Appendix I	147

LIST OF TABLES	PAGE
Table 4.1 Results of content validity conducted by the educators on the stoichiometry test	69
Table 4.2 Inter-correlations of the three motivational factors	71
Table 4.3 Factor loading matrix showing a pattern of how items loaded into discrete factors	72
Table 4.4 Questionnaire items as grouped together by factor analysis	73
Table 4.5 Internal reliability: Cronbach's alpha coefficients of the instrument upon removal of individual items	74
Table 4.6 Descriptive statistics of students' performance in the pre- and posttest	76
Table 4.7 Descriptive statistics of students' average confidence scores in the pre- and posttest	76
Table 4.8 Student categories based on the evaluation of their performance in the pretest	78
Table 4.9 Student categories based on the evaluation of their performance in the posttest	78
Table 4.10 Shifts in student groups after teaching and learning	82
Table 4.11 Pre- and posttest performance data according to performance evaluation subgroups	86
Table 4.12 Multiple comparisons (p values) of CMY 133 performance of students in the four pre-post performance evaluation subgroups	90
Table 4.13 Multiple comparisons (p values) of posttest performance of students in the four pre-post performance evaluation subgroups	92
Table 4.14 Multiple comparisons (p values) of pretest average confidence scores of students in the four pre-post performance evaluation subgroups	93
Table 4.15 Multiple comparisons (p values) of posttest average confidence scores of students in the four pre-post performance evaluation subgroups	95
Table 4.16 Multiple comparisons (p values) of average performance gain of students in the four pre-post performance evaluation subgroups	97
Table 4.17 Categories and super-categories generated from emerging codes	109

LIST OF TABLES	PAGE	
Table 4.18	Cohen's kappa values and the level of agreement observed between the coding systems of the two coders per item	111
Table 4.19	Students' open-ended responses about how they would explain their choice of confidence judgement ratings	114
Table 5.1	Summary of categorization of students in terms of accuracy of performance evaluation in the pre- and posttest	134
Table 5.2	Summary of student performance and average confidence scores in the pre- and posttest	134
Table 5.3	Five pre-post performance evaluation subgroups and the number of students in each subgroup	136
Table 5.4	Pre- and posttest performance data according to performance evaluation subgroups	137

LIST OF APPENDICES	PAGE
I. Stoichiometry test	158
II. Student questionnaire (Pilot Study)	181
III. Educator Instruction Sheet (Pilot Study)	182
IV. Educator Questionnaire: Section B (Pilot Study)	183
V. Educator Questionnaire: Section C (Pilot Study)	185
VI. Questionnaire used in main study	205
VII. Consent form	208
VIII. Letter of Approval from the University of Pretoria's Ethics Committee	209

LIST OF ACRONYMS

BFYP	BSc Four-year programme
EOL	Ease-of-learning
GPA	Grade Point Average
FOK	Feeling of knowing
FOnK	Feeling of not knowing
JOK	Judgement of knowing
JOL	Judgement of learning
JOnK	Judgement of not knowing
LTM	Long-term memory
OC	Overconfident
R	Realist
SC	Super-category
SE	Self-enhancement
SOJ	Second order judgement
SP	Self-protection
TIMSS	Trends in International Mathematics and Science Study
TOT	Tip-of-tongue
UC	Underconfident
UCT	University of Cape Town
UL	University of Limpopo
UNIFY	University of Limpopo Foundation year programme
UNISA	University of South Africa
UP	University of Pretoria
UPFY	University of Pretoria Foundation year programme

CHAPTER 1

INTRODUCTION TO THE STUDY

CONTENTS	PAGE
1.1 Introduction	2
1.2 Background and context of the study	2
1.2.1 BSc Four-year programme	2
1.3 Statement of the problem	3
1.4 The rationale for the study	5
1.5 The aim of the study	8
1.6 Research questions	8
1.7 Sequence of the research report	8

CHAPTER 1

INTRODUCTION TO THE STUDY

1.1 INTRODUCTION

In this chapter the background and context of the study, the rationale, the aim as well as the research questions the study attempts to answer are discussed. The chapter concludes with a description of the sequence followed in the research report.

1.2 BACKGROUND AND CONTEXT OF THE STUDY

An academic development programme is an intervention by a tertiary institution to address under-preparedness of incoming students for the mainstream programmes offered by that institution with the goal of ultimately widening access for under-prepared students to maths and science-related careers (Potgieter, Dawidowitz and Mathabatha, 2007). South Africa has a school system that mainly produces students who are under-prepared for tertiary science studies. This emphasises the need to introduce such programmes in South African institutions.

Previously the University of Pretoria (UP) offered two academic development programmes, namely University of Pretoria foundation year programme (UPFY) and the BSc extended programme. A foundation year programme at UP focused primarily on subject content contained in high school syllabi that was assumed as pre-knowledge for mainstream coursework. Extended programmes on the other hand offered first-year subject matter at a slower pace so that deficiencies could be addressed. With time the South African government decided to change its conditions for funding academic development programmes. Government preferred to fund extended programmes as opposed to foundation year programmes because these were degree programmes rather than bridging programmes. The two programmes namely UPFY and the BSc extended programme were replaced by the BSc Four-year programme (BFYP) in 2008.

1.2.1 BSc Four-year programme (BFYP)

Prior to admission to the BFYP prospective students are subjected to a stringent selection process which entails meeting a minimum score requirement for their Grade 12 mathematics and physical science marks and passing an admission test endorsed by the university. The

programme is an 18 month programme of three semesters. The first six months focus on foundational knowledge at school level. In the next six months the mainstream semester one chemistry workload is covered and in the last six months students are introduced to aspects of theory to be covered in mainstream semester two chemistry. This is done to enable weaker students to learn the work at a slower pace. However in the third semester the pace is increased to acclimatise students to the pace of mainstream teaching. Upon completion students receive credits equivalent to the successful completion of a first-year mainstream course in a particular subject. Candidates with suitable marks at the end of the first calendar year may apply for a transfer to engineering or the health sciences. The majority of students, who pass however, proceed to the third semester.

During a typical week students are exposed to teaching in the form of two large group lectures as well as to two tutorials or small group sessions per module. In the large group sessions students are exposed to an overview teaching of prescribed topics by different lecturers using different teaching styles. During the small group sessions students are assigned to only one tutorial lecturer. It is in the small group sessions where in-depth teaching of the topic, extensive supervised problem-solving and continuous evaluation take place. Students also receive continuous feedback on their performance in the small group sessions. Apart from the evaluation activities carried out in the small group sessions, it is compulsory for all students to complete a computerised quiz per topic which is posted on the university's intranet.

1.3 STATEMENT OF THE PROBLEM

Social and economic development in South Africa is largely dependent on mathematics and science. The Trends in International Mathematics and Science Study (TIMSS) 2003 science and mathematics reports (Martin, Mullis, Gonzalez & Chrostowski, 2004; Mullis, Martin, Gonzalez & Chrostowski, 2004) suggest that South African learners are currently experiencing the greatest challenges with regard to these two subjects. Under-prepared students fail to meet tertiary science entry requirements because of poor mathematics and science results.

Having failed to meet the entry requirements for mainstream courses, students generally find themselves in an academic development programme (Mabila, Malatje, Addo-Bediako, Kazeni & Mathabatha, 2006). Upon admission to the programme they have to pass the

modules offered in the programme if they want to be accepted into a mainstream course. To gain entry into an academic development programme, students have to participate in selection processes and meet stringent requirements (such as an admission test) that may influence the students' academic self-concept unfavourably and give rise to them having high perceptions about their ability to do well in the sciences. Failing may not impact significantly on this perception and they may choose to ignore any cues of failure when this does not fit in with their academic self-concept. It is also important to bear in mind that students admitted to these programmes were often the best in their high schools and many had never failed a grade before. Experiencing failure in a test may prove to students in academic development programmes that there is a lot that they still do not know or misunderstand content-wise. If these students fail to accept this message and do not seek ways to improve on time, failure may be inevitable.

Once admitted to an academic development programme, under-prepared students are faced with the task of making wise choices such as how best to prepare for examinations and which career path to follow. All these decisions require an accurate evaluation of one's strengths and weaknesses, i.e. what one knows or does not know well, what one does well and where one needs improvement. In order to succeed one of the things students are required to do, is to evaluate themselves and their performance accurately throughout the programme. Ehrlinger (2008) states that there is no simple answer to the question of whether accurate performance evaluation is on the whole good or bad. However there are many cases in which accurate performance evaluation is an important goal. It for example is important for students in academic development programmes to be accurate in their performance evaluation in order to efficiently regulate their own learning.

In their study Potgieter *et al.* (2007) found inaccuracy in performance evaluation to occur more among students admitted to UP's and the University of Limpopo's (UL) foundation year programmes. If under-prepared students are to fully benefit from these programmes they would have to be accurate in judging their performance and progress. Literature has shown that inaccuracy in calibration is not conducive to academic success (Nowell & Alston, 2007), for example students end up studying less than if they had had an accurate perception of their ability.

1.4 THE RATIONALE FOR THE STUDY

Prior to consolidation of the two programmes at UP, i.e. UPFY and the BSc Extended programme, a study was conducted by Potgieter *et al.* (2007) which focused on the relationship between confidence and performance of first-year chemistry students at three tertiary institutions namely University of Pretoria (UP), University of Limpopo (UL) and the University of Cape Town (UCT). The sample used for data collection at the beginning of the year 2005 consisted of three groups of first-year chemistry students at UP (Mainstream chemistry, BSc extended programme and UP's foundation year programme abbreviated as UPFY), two groups at UCT (mainstream chemistry and UCT Academic Development programme) and UL (mainstream chemistry and UL Foundation Year programme known as UNIFY). Academic development programmes at UP (BSc extended programme) and UCT consisted mainly of black students coming from disadvantaged backgrounds. In the two universities' academic development programmes a period of two years was utilised to cover the first-year chemistry syllabus as opposed to one year for the mainstream chemistry course. In academic development programmes additional support was given in the form of extensive supervised problem-solving sessions. The foundation year programme on the other hand consisted of black students with good potential but coming from disadvantaged academic backgrounds. The foundation year programme sought to strengthen such students' secondary education in preparation for tertiary studies. So generally foundation year programmes consisted of students who had entered the university under-prepared for tertiary studies. Immediately after answering a question in a test designed to probe for knowledge and understanding of chemistry as well as the participants' level of skills development, participants were required to report on a four-point scale how certain they were that the answer they had provided was correct. A comparison of the students' performance and confidence indices was used to gauge the quality of judgements the students in the different cohorts were capable of making. In their study Potgieter *et al.* (2007) found that students from the UPFY programme were overly optimistic about the correctness of their answers, i.e. despite much poorer performance, in some of the subsets of test items the confidence they had in the accuracy of their answers was similar to or higher than that of UP mainstream students. These students displayed poor calibration in the sense that many of the answers which they expected to be correct were indeed wrong.

Inaccurate judgements about one's competence in specific subject matter can potentially have serious consequences as an accurate performance evaluation is critical in decisions on the

time required to study for the specific course, what study methods to employ as well as what topics to give the most attention to. An inaccurate performance evaluation of how much one knows and understands may lead one to study less than if one had accurate perceptions (Grimes, 2002; Nowell & Alston, 2007).

The role of assessment in the development of an important metacognitive skill such as accurate performance evaluation was described by Carvalho (2007: 2): “Test-taking is a particularly challenging academic requirement and a valuable opportunity for students to learn how to regulate their own learning in a certain domain. In the process of preparing for a test and while taking it, students have the opportunity to make decisions about the efficiency of their learning strategies, to learn how to better monitor their performance in that domain, to make attributions to their failures and successes, and to learn how to behave in future similar situations.” The implication here is that assessment should offer students an opportunity to monitor and evaluate their performance in order to effectively pace and regulate their own learning.

The study conducted by Ochse (2003) showed that overconfidence or being overoptimistic may have a negative effect on subsequent performance. Third-year students were asked to give an estimate of their average exam mark in psychology as a percentage. In addition they had to indicate how sure they were of obtaining their expected mark on a Likert scale ranging from 100% to 0%. Basically, students in this study were required to predict their final exam score and also to report the level of confidence they had in the accuracy of their predictions. Students whose expected mark was nine or more marks higher than the actual mark were labelled as “Overestimators”. The unjustified high confidence observed in the students who overestimated their mark was referred to as overoptimism. Students, who overestimated when they were asked to predict the score they expected to obtain for the final examination of a module, were significantly more confident about the accuracy of their expected scores, they perceived themselves to have higher ability but yet obtained the lowest final scores compared with scores of students who were realistic in their estimation.

Overoptimism or overconfidence according to Nowell and Alston (2007) has two dimensions. The first kind may be defined as a reflection of an inflated view of an ability to accurately predict future performance. The second type reflects self-assessment that is overly optimistic. Based on the analysis presented by Nowell and Alston (2007), when students display an

inflated view of their ability to accurately predict or estimate future performance they can be referred to as “Overestimators”. In our study, similar to a study conducted by Carvalho (2007) students are not required to report their expected scores. They are rather expected to indicate the level of confidence they have in the accuracy of their answers in a test. We therefore focus on overconfidence rather than overestimation. We define overconfidence therefore as an inflated level of confidence one displays with regard to the accuracy of one’s answers in a test.

It has been reported in the literature that many times people have been found to report overly optimistic judgements when they were asked to evaluate their performance or competence. When we are prompted to make judgements on how we perceive our ability or how well we know something or how well we have performed in a particular task, the judgements we report are called metacognitive judgements (Dunlosky, Serra, Matvey and Rawson, 2005; Fernandez-Duque & Black, 2007; Koriat & Bjork, 2005; Rosenthal, 2000). According to Dunlosky *et al.* (2005) metacognitive judgements have been extensively investigated partially due to the fact that mastering the skill of accurately making them, may result in the effective regulation of self-paced study which is necessary in a tertiary environment where an independent approach to studying is required.

The literature findings reported in the previous paragraphs demonstrate the general occurrence of overconfidence and the potentially negative consequences that it may have on academic success. We have therefore decided to conduct a study to investigate the presence, extent and impact of inaccurate metacognitive judgements in students enrolled in an academic development programme. Findings from such a study will serve to inform staff at tertiary institutions who are involved in academic development programmes. It is anticipated that the findings will have a potential influence on the design, monitoring and presentation of a curriculum and assessment strategies unique to such programmes in order to achieve improved pass rates and therefore increased access for students to mathematics and science fields. “Understanding the factors involved in test-taking that affect students’ planning and monitoring, attributional, and regulatory processes would contribute to creating evaluation practices and conditions that promote learning during the evaluation process.” (Carvalho, 2007: 2).

1.5 THE AIM OF THE STUDY

The purpose of this study is to ascertain how students in UP's academic development programme evaluate their competence to solve chemistry problems when asked to do so and to identify possible factors associated with inaccurate performance evaluation. A further aim is to determine whether exposure to good quality tertiary teaching and feedback after evaluation would improve the students' performance in chemistry, the quality of the judgements made about their command of the subject as well the factors students rely on when evaluating their performance. For the purpose of this study the terms metacognitive judgements, performance evaluations and judgements of performance will be used synonymously.

1.6 RESEARCH QUESTIONS

This study is an attempt to answer the following research questions:

1. How accurately do BFYP students evaluate their performance in a stoichiometry test?
2. What is the influence of teaching of stoichiometry in the BSc Four-year programme on performance and accuracy of performance evaluation?
3. What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?
4. What is the relationship between bias in performance evaluation and self-enhancement, self-protection and gender?

1.7 SEQUENCE OF RESEARCH REPORT

In the first chapter the background, the statement of the problem, the rationale for the study, the aim of the study and the research questions were discussed. The second chapter reviews literature relevant to the study. In the third chapter the research methodology is outlined. The theoretical paradigm the study is situated in, research design, sample, instrumentation, validity, reliability, pilot study, main study, procedures used to analyse main study results and ethical considerations are described. Chapter four presents an analysis and discussion of the results. The fifth chapter summarises the findings of the study and draws conclusions from them. In closing, Chapter five discusses the educational implications of the study's findings, limitations of the study as well as areas for further research. Recommendations on how the

developed test instrument and the findings can be used in future research are also made. References and appendices then follow for easy cross-referencing.

CHAPTER 2

LITERATURE REVIEW

CONTENTS	PAGE
2.1 Introduction	11
2.2 Metacognition	11
2.3 Factors associated with bias in performance evaluation	21
2.3.1 Task-related factors	22
2.3.2 Personal factors	24
2.4 Conceptual framework of the study	27
2.5 Inaccuracy of performance evaluation: methodology	29
2.6 Why use stoichiometry to investigate bias in performance evaluation	31
2.6.1 Stoichiometry and the multistep problem	33
2.6.2 Stoichiometry and representational competence	34
2.6.3 Stoichiometry and misconceptions	35
2.7 Conclusion	37

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter reviews literature, first on several constructs such as metacognition, metacognitive monitoring and control, metacognitive judgements, bias in performance evaluation, and overconfidence. To summarise, all the constructs which are explored are built into a conceptual framework of the study. This is followed by a brief description of different ways that have been used to investigate bias in performance evaluation in other studies. The chapter proceeds to a defence of why amongst all the other topics featured in the first-year chemistry curriculum the study focused on the topic of stoichiometry as a means to investigate bias in performance evaluation. This is done by way of a presentation and discussion of literature on stoichiometry and misconceptions students have on the topic. The chapter concludes with a brief summary of the literature reviewed.

2.2 METACOGNITION

Metacognition according to Rosenthal (2000: 203) is “the only access we have to whether, or how likely it is that, we know something”. J. H. Flavell invented the term “Metacognition”, and he defined the concept as follows: “Metacognition refers to one’s knowledge concerning one’s own cognitive processes or anything related to them, e.g. the learning-relevant properties of information or data. For example, I am engaging in metacognition if I notice that I am having more trouble learning A than B; if it strikes me that I should double-check C before accepting it as fact” (Flavell, 1976: 232). Traditionally metacognition is defined as the knowledge and experiences we have about our own cognitive processes (Flavell, 1979).

Metacognition involves monitoring one’s progress as one learns and making changes and adapting one’s strategies when one realises that one is not doing well. Metacognitive skills are therefore some of the skills that differentiate a novice learner from an expert learner. An expert learner knows how to learn and also knows which strategies work best (Halter, 2008). Because of this, metacognitive skills may be critical ingredients if successful learning is going to take place. It is therefore very important for a learner to be aware of the effectiveness of his or her learning strategies and skills as well as the correctness of the knowledge and understanding he or she has of the concepts in a particular subject area.

Metacognition consists of metacognitive knowledge and metacognitive experiences or regulation. Metacognitive experiences entail the use of metacognitive strategies or regulation. Metacognitive strategies are sequential processes (planning, monitoring cognitive activities and checking the outcomes of those activities) an individual follows to control cognitive activities and to ensure that cognitive goals are met (Livingston, 1997). Livingston (1997) gives a good example that demonstrates the use of metacognitive strategies as a sequential process. When a learner reads a paragraph with the cognitive goal of understanding the text, the learner may use self-questioning as a metacognitive comprehension monitoring strategy to determine whether the cognitive goal has been achieved. In the event of the cognitive goal not being achieved i.e. if the learner cannot answer her own questions, the learner must determine what needs to be done in order to achieve the cognitive goal of understanding the text. Metacognitive knowledge on the other hand refers to knowledge of cognitive processes and the knowledge that can be used to control cognitive processes. Flavell (1979) further divides metacognitive knowledge into three categories: knowledge of person, task and strategy variables.

Knowledge of person variables refers to individual knowledge of one's own learning processes. For example you may be aware that for your study to be more effective you need to study in an environment with fewest distractions and minimal noise levels. Knowledge of task variables refers to knowledge about the nature of the task as well as the type of processing demands that it will place upon the individual. For example you may know that it will take more time for you to study for a science test than for an English test. According to Hartman (2001) knowledge of the strategy variables refers to knowing what (factual or declarative knowledge), knowing when and why (conditional or contextual knowledge) and knowing how (procedural or methodological knowledge). All the facets of metacognitive knowledge are necessary for one to self-regulate one's thinking and learning effectively (Hartman, 2001).

A learner who possesses metacognitive skills should then be able to plan and select relevant strategies, monitor the progress of learning, correct errors, evaluate the effectiveness of learning strategies and change strategies and learning behaviours when necessary (Ridley, Schutz, Glanz & Weinstein, 1992). While learning, novice learners do not stop to evaluate their understanding. When faced with solving a problem, novice learners usually do not try to examine a problem in depth; they are satisfied just to scratch the surface. They do not try to

see the relevance of the material they are learning to the contexts outside the classroom (Hartman, 2001).

To demonstrate metacognitive skills however, the learner needs to have a good knowledge and understanding of the content. How else can one know if the strategy is relevant when one does not even know which strategy is required to solve the problem in the first place?

For example when a learner is asked to solve a stoichiometry problem in chemistry to demonstrate metacognitive skills or ability, the learner must be able to:

1. recognise and understand what is being asked or what is expected of him or her;
2. recognise the suitable strategy or approach to solve the problem (e.g. convert grams of reactants to moles, determine the limiting reactant and use the moles of limiting reactant to determine the moles and eventually the grams of the product);
3. perceive and acknowledge when he or she cannot solve the problem;
4. reflect on the reasons why he or she cannot solve the problem and make changes to, and adaptations of his or her strategies in order to improve.

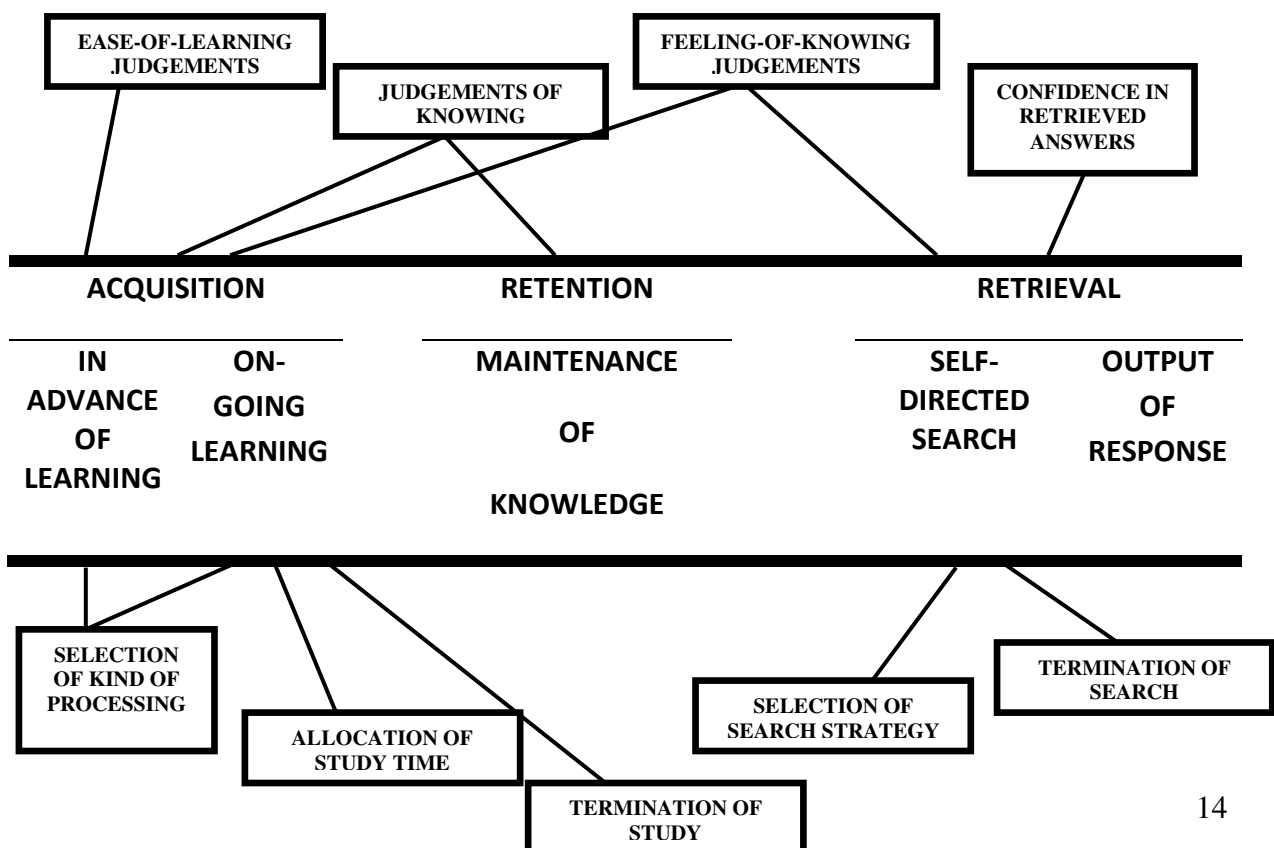
The problem arises when students do not perceive when they are not doing well or when students think they are doing well while the performance is contrary to the perception.

Modern research in metacognition stems from two parallel roots. One emerged from the cognitive psychology of the 1960s e.g. Hart (1965) and the other emerged from the post-Piagetian developmental psychology of the 1970s, an example being the work of Flavell (1979). Hart (1965) was more interested in the accuracy of judgements people made of their memory abilities. Flavell (1979) on the other hand was interested in determining the relationship between a greater understanding of the rules that govern memory and cognition and improvement in children's memory abilities (Schwartz and Perfect, 2002), and although the two paths have remained separate, modern research was introduced to the construct of metacognition through the publication of Nelson and Naren's (1990) theory of monitoring and control. According to Schwartz and Perfect (2002) the theory was able to integrate almost all of the existing research on metacognition. The theory focused on the interaction between metacognitive monitoring and control. Metacognitive monitoring entailed processes that enabled individuals to observe, reflect on, or experience their own cognitive processes (Flavell, 1979) whereas metacognitive control could be observed in the decisions individuals consciously or unconsciously made based on the outcome of their monitoring. Monitoring is

revealed by asking participants to make judgements about their memory, knowledge, learning or comprehension. Control on the other hand is revealed by the actions an individual engages in as a result of the monitoring, for example decisions about which items to study and the amount of time allocated to study (Schwartz and Perfect, 2002). Without the work of Nelson and Narens (1990) on metacognitive monitoring and control, Flavell's (1979) research could not show any strong correlation between metacognitive thinking and improvements in memory (Schwartz and Perfect, 2002).

When one is asked to judge whether one knows something, or how easily one will learn an item, or even whether one has successfully learned an item, these judgements according to Rosenthal (2000) are metacognitive judgements. Nelson and Narens (1990) identified several types of metacognitive judgements namely ease-of-learning judgements (EOL), judgements of knowing or judgements of learning (JOL), feeling-of-knowing judgements (FOK) and confidence judgements. The theoretical framework of Nelson and Narens (1990) which is shown in Figure 2.1 below shows an overview of how different metacognitive judgements guide the monitoring and control processes that occur when a student studies for an upcoming examination and when a student retrieves information during an examination.

Figure 2.1: An overview of monitoring control processes (adapted from Nelson & Narens, 1990)



The framework consists of three stages, namely the acquisition, retention and retrieval stages. The acquisition stage takes place prior to studying for the examination. The retention stage occurs when a student is busy studying for the test and the retrieval stage is when the student is taking the test and information is being retrieved. The metacognitive monitoring part of the acquisition stage entails the student setting goals he wants to achieve. This is guided by a judgement the student makes on the level of mastery that will have to be attained during acquisition. The metacognitive control aspect of the acquisition stage entails the student formulating a plan of how he intends to achieve the set goals. The student formulates the plan guided by several metacognitive judgements forming part of the monitoring part of the acquisition stage. Prior to acquiring the necessary knowledge, the student makes a judgement on what will be easy or difficult to learn in the target content information and which strategies will make learning easier. These are ease-of-learning judgements. During or after acquisition of information, the student makes judgements on expected test performance based on currently recallable items. These are judgements of learning. Also occurring during and after acquisition are feeling-of-knowing judgements. These are judgements a student makes on whether a currently non-recallable item is known or will be remembered in the upcoming test. As part of metacognitive control, based on these metacognitive judgements the student can then make decisions such as the amount of time to be allocated to studying, type of information processing which will ensure retention as well as when to stop studying.

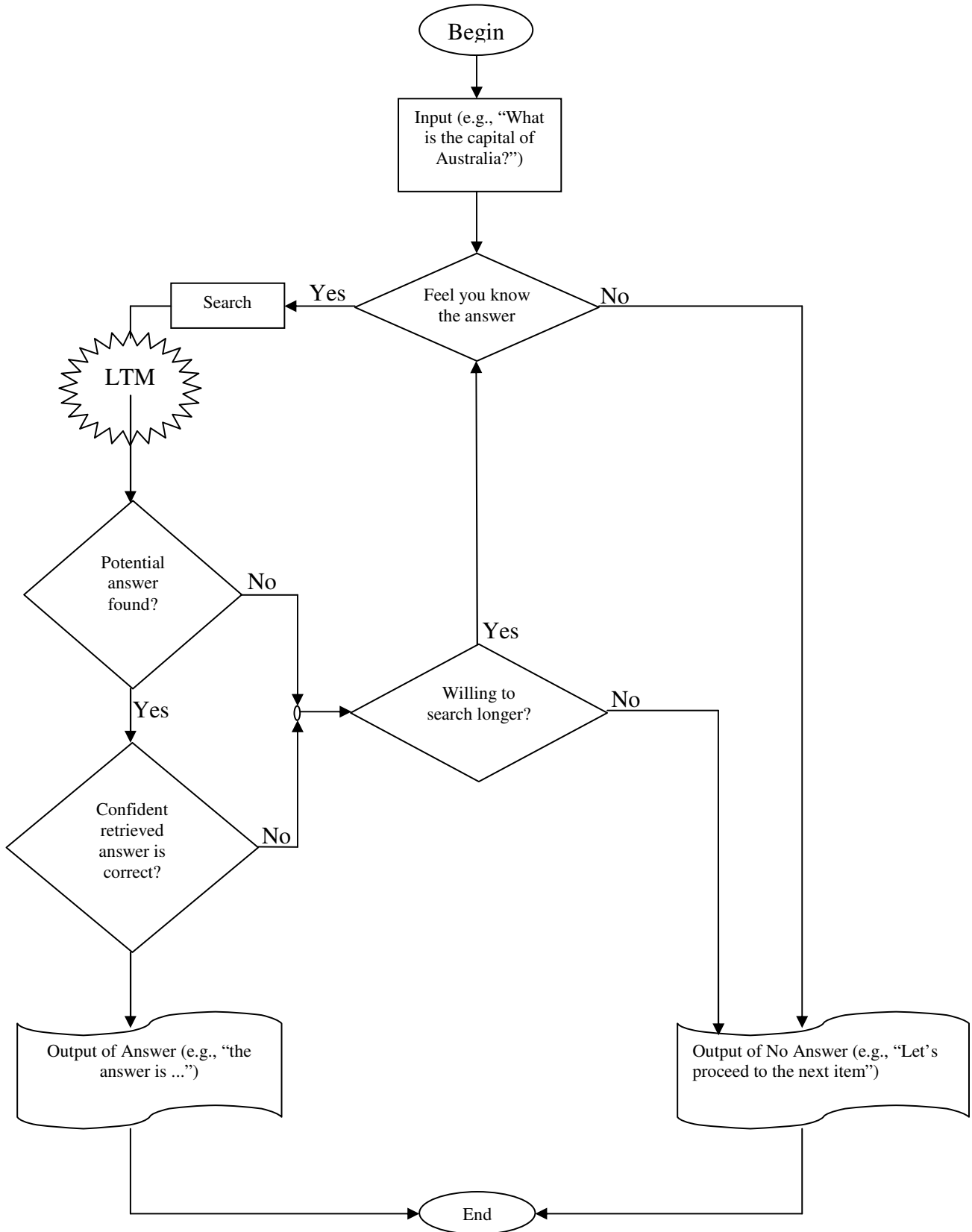
During the retention stage the student will make the decision whether to review the material or not based on a judgement he makes on how much information he has been able to retain in his long-term memory. As part of the student's metacognitive monitoring he will make a judgement of knowing and based on the judgement, he will control his learning by adopting strategies that would ensure better retention for retrieval during test-taking (Nelson & Narens, 1990).

The retrieval stage takes place when the test is being taken. Figure 2.2, taken from Nelson and Narens (1990), shows the process that occurs during information retrieval. The question the student is confronted with in a test is described as an input. Nelson and Narens (1990) posit that the urge to search for the correct answer in metamemory is initiated by a feeling of knowing.

In the absence of a FOK, there should be no answer given as output. Jing, Kazuhisa and Yuejia (2003) define the feeling of not knowing (FOnK) as the accurate negative FOK predictions that accurately anticipate “not-knowing”. The explanation presented by Jing *et al.* (2003) of how an individual reaches a FOnK judgement is consistent with the accessibility hypothesis (Metcalf, 2000). They argue that FOK predictions are as a result of an effortful process of retrieval while FOnK predictions are based on a “null” retrieval process. They posit that due to little or no information retrieved, subjects make a judgement of “I don’t know”. On the other hand, consistent with the cue-familiarity hypothesis proposed by Reder and Ritter (1987), if the information in the question asked does not elicit any familiarity, an individual quickly reaches the conclusion that the information is not available in memory. According to Glucksberg and McCloskey (1981), “do not know” decisions can be divided into two basic types. Firstly, when a person is in possession of concrete content information relevant to the question asked, the person will locate and evaluate the information in order to determine whether what is stored in memory can be used to answer the question. This results in a slow, low confidence decision when the person finds that the information in memory is not sufficient to answer the question. Secondly, a rapid response of not knowing is produced when a person has no knowledge relevant to the question asked. The response produced is fast because when the initial search for information draws a blank, the person stops searching.

In the event that the FOK is positive, the student proceeds to search his long-term memory (LTM) for an answer. When the student is confident that his retrieved answer is correct he can then report the answer as an output. When the process of searching results in no potential answer, the student may decide to spend more time searching or stop searching. The output in this case is described by Nelson and Narens (1990) as an omission error. An incorrect answer retrieved after the process of searching and output as correct however is described as a commission error (Nelson and Narens, 1990). The explanation that Nelson and Narens (1990) give for commission errors is that people’s FOKs are not completely accurate and sometimes mistaken because they refer to the wrong information. Nelson and Narens (1990) also report that commission errors are more prevalent in college students rather than omission errors. In other words rather than not give an answer after failing to retrieve a potential answer students would report an answer based on the wrong information as correct. Several studies reveal why it is not enough to base the judgement that the answer is correct, solely on FOKs.

Figure 2.2: An overview of metamemory components during the retrieval stage (adapted from Nelson and Narens, 1990).



(i) Feelings of Knowing (FOK)

According to Winnie and Nesbit (2010) a feeling of knowing is a belief that information is in memory even though it cannot be retrieved. Koriat (2000) relates the feeling of knowing with the tip-of-tongue (TOT) phenomenon which is experienced by an individual when he/she struggles to retrieve an elusive name from memory. The TOT phenomenon distinguishes between the subjective conviction that the individual knows the name and the actual inability to produce the name (Koriat, 2000). During the TOT experience one can sense the missing name or word and not just acknowledge its existence. Koriat (2000) further states that during a TOT state people can sense the emergence of the target they want to recall into consciousness and are able to judge its closeness or imminence. People might refer to the feeling aroused by the TOT experience as an intuitive feeling, a hunch or “just knowing”. Koriat (2000), states that this is a kind of feeling that is self-evident, requiring no justification. Thus, when making a JOK some people may rely on a feeling of knowing. However, there are instances when feelings of knowing judgements may not be accurate and attributes of metamemory hypotheses are observed in the factors that influence FOK judgements and their accuracy. The cue familiarity hypothesis (Reder & Ritter, 1987) implies that an individual’s metacognitive judgement is based on how familiar he/she is with the information provided in a question. Meaning that if an individual is familiar with the topic or terms, on which a question is based, he/she is likely to judge that he/she knows the answer to the question. The individual will however more likely judge that he/she does not know the answer to a question which presents new or unfamiliar topic or terms. Lastly, the competition hypothesis (Maki, 1999) points out that the danger of relying solely on familiarity is that an individual may mistakenly assume familiarity with an object due to its similarity with the target object. Thus fewer memory traces result in low level of competition and ultimately a more accurate FOK judgement rating.

(ii) Affective feelings

The other type of feelings influencing the making or construction of metacognitive judgements is discussed by Greifeneder, Bless and Pham (2010). They describe affective feelings as subjective experiences that may or may not be directly related to an object such as moods and emotions. Affective feelings can be conceptualised as experiential information people rely on when forming judgements (Greifeneder *et al.*, 2010). This is called the feelings-as-information hypothesis. Feelings are experienced; therefore the information is qualitatively different from activated content information. During the formation of such a

judgement people are thought of asking themselves questions like “How do I feel about it? Therefore such a judgement is more sensitive to moods and attitudes at the time. Affective feelings are characterised as either being incidental or integral to the target. Incidental feelings are elicited by an external source other than the target being judged for example, the negative mood a student may be in while writing an exam due to insufficient time spent studying. The implication for teaching and learning here is that when a student cannot recall nor has no knowledge of required content information to answer the question, the student may randomly choose an answer in a multiple choice test situation and make JOnKs and negative confidence judgements informed by how he/she may be feeling at the time of making the judgements rather than the lack of content information. In the defence of his/her choice of JOnK and negative confidence judgement the student may then motivate his/her choice in terms of external factors such as lack of preparation or memory. Integral feelings are on the other hand elicited by features of the target object whether the features are real, perceived or imagined (Greifeneder *et al.*, 2010). Integral feelings may for example be feelings of difficulty or ease experienced for example, when solving a problem in a chemistry test situation. In summary, integral feelings may be attributed to the target object while incidental feelings are unconnected to the target. People may therefore rely on integral or incidental feelings when making metacognitive judgements such as JOKs, JOnKs and confidence judgements.

(iii) Cognitive feelings

Cognitive feelings are experiences that reflect activated information accompanied by cognitive processes such as the ease with which information can be retrieved from memory (Greifeneder *et al.*, 2010). During the formation of a metacognitive judgement people may use cognitive feelings of ease-of-retrieval as a source of information rather than rely directly and solely on content information. A student may feel that he or she is familiar with the content information required in a test question which would increase the ease with which he or she can retrieve the information.

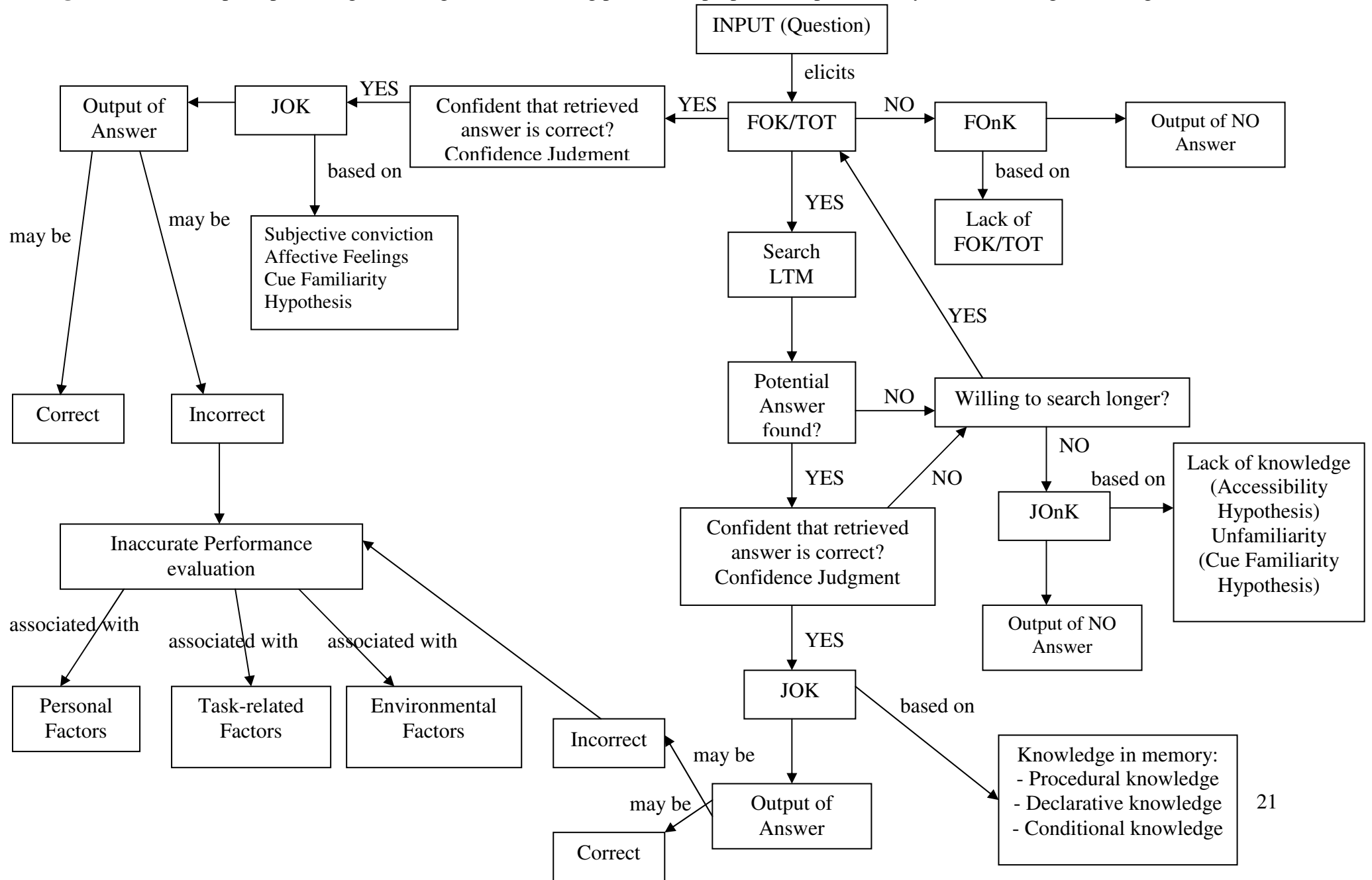
Apart from FOKs several psychological mechanisms may underlie the decision to output a single specific answer as correct (Nelson & Narens, 1990). A recognition stage occurs in which the student retrieves the answer based on a recognition judgement. In the event that only one answer is retrieved the student then gauges that answer against his confidence judgement indicated in Figure 2.2. The student may choose to continue searching but if the

retrieved answer is not associated with enough confidence and no other potential answer is retrieved, the student may output the initial answer even though it was not associated with enough confidence. Another strategy that Nelson and Narens (1990) highlights is the one in which people output an answer even when they themselves are not convinced it is the correct one. People would then output such an answer on the basis that it has the likelihood of being correct.

Based on the framework of Nelson and Narens (1990) and explaining the framework in the context of students taking a test, it seems that after the search for information has occurred, the students would have to make a confidence judgement about the correctness of their retrieved answer before putting it out as the answer. An error or bias in performance evaluation is observed when students indicate high confidence levels regarding the correctness of a retrieved answer and when the answer itself is incorrect.

Figure 2.3 is an overview of the anticipated metacognitive monitoring processes students may experience while taking the test. The concept map shows that a question asked in the test may elicit either a feeling of knowing (FOK) or feeling of not knowing (FOnK). When a FOK is elicited the students may choose to either claim a judgement of knowing (JOK), make a confidence judgement and give an answer based on how they feel or alternatively, search their long-term memory (LTM) prior to making a confidence judgement as well as a JOK. A judgement of not knowing (JOnK) is made when attempts to retrieve a potential answer from memory have failed. A JOnK and a FOnK should lead to an output of no answer based on reasons that a tip-of-tongue experience is not elicited by the question or information given in the question for the latter and lack of knowledge or unfamiliarity with question or information in the question for the former. A JOK made without consulting information in the LTM is based on merely a subjective conviction while a JOK made after searching LTM should be substantiated by objective evidence such as possession or demonstration of declarative, procedural or conditional knowledge. Literature has however shown that JOKs based on feelings or objective evidence in memory may be incorrect due to the competition hypothesis. Incorrect answers reported post JOK and positive confidence judgements may be interpreted as bias in performance evaluation.

Figure 2.3: A concept map showing a metacognitive monitoring process we propose is experienced by students during test-taking



Research on the benefits of inaccurate self-insight has largely focused on one type of inaccuracy – overconfidence (Ehrlinger, 2008). Investigations in behavioural science have shown in different domains that people are most of the time not good at accurately assessing their performance or competence. In fact most people have been found to be overconfident about their performance (Ehrlinger, 2008). People have been found to be overconfident in assessing their likelihood of developing health problems (Strecher, Kreuter & Kobrin, 1995) , Lawyers have been found to be overconfident in assessing their likelihood of winning cases they were about to try (Loftus & Wagenaar, 1988) . Laboratory technicians have shown very little insight into how well they have performed in tests of their skills (Haun, Zeringue, Leach & Foley, 2000). Weak students have been found to be overconfident in the assessment of their performance in tests (Carvalho, 2007; Carvalho & Yuzawa, 2001; Kruger & Dunning, 1999; Kennedy, Lawton & Plumlee, 2002, Potgieter *et al.*, 2007). Nelson and Narens (1990) pointed out that the tendency to report confidence in the correctness of answers which were indeed wrong was found to be prevalent in college students. In addition to the exaggerated confidence judgements made on the basis of JOKs based on feelings or incorrect information, several task-related, personal and environmental related factors may be associated with bias in performance evaluation.

2.3 FACTORS ASSOCIATED WITH BIAS IN PERFORMANCE EVALUATION

Carvalho (2007) identifies three kinds of factors that may influence students' performance and monitoring processes in academic tasks, namely personal factors, task-related factors and environmental factors. This study only focuses on several personal and task-related factors identified in literature. Research has identified several factors associated with bias in performance evaluation or monitoring. In my discussion of how these factors have been found to influence metacognitive monitoring and therefore accuracy in performance evaluation, I have categorised them as either personal or task-related factors. Under personal factors, the following factors are discussed: the tendency to rely on chronic self-views to evaluate performance; the need for self-protection and self-enhancement; theories of intelligence respondents adhere to; personality traits and gender. Task-related factors discussed are lack of knowledge; properties of the task; format selected for evaluation and the quality of feedback received. Task-related factors are discussed followed by a brief discussion of several personal factors identified in literature.

2.3.1 Task-related factors

Several task-related factors may influence the accuracy with which students may judge their performance.

(i) Lack of knowledge

When learners are asked to report metacognitive judgements on their performance in a test in an academic context, they are really asked to evaluate their metacognitive knowledge of task variables and strategy variables. When learners illustrate good metacognitive knowledge of these variables, it is an indication that they know and acknowledge what they know and do not know. However, when they illustrate poor metacognitive knowledge of these variables, this becomes an indication that they do not know what they know and do not know. Research has identified lack of knowledge as the most basic level factor associated with error in performance evaluation (Kruger & Dunning, 1999).

“For a person to know whether he or she has answered each test question correctly, he or she must know which the correct answer is. By definition, those who lack skill do not know the correct answer and, as such, lack the knowledge necessary to realize that they have not performed well. Indeed, those who lack skill have greater difficulty than their more skilled counterparts in distinguishing between correct and incorrect responses, whether those responses are their own or are provided by another individual” (Ehrlinger, 2008: 385). At least a certain level of knowledge is required for students to accurately distinguish between strong and weak performances. Poorly performing students have been found to have a greater tendency to be overconfident than do those who perform well (Beyer, 1999; Nowell *et al.*, 2007; Potgieter *et al.*, 2007). According to Carvalho (2007), students can efficiently monitor their test-taking according to their content knowledge. In their study Kruger and Dunning (1999) observed that the level of skill possessed by participants determined the degree of error in the participants’ performance evaluation. Participants were asked to report estimates of how well they had performed immediately after writing an examination. They were divided into quartiles according to their examination performance. When comparing the participants’ expected and actual examination performance, Kruger and Dunning (1999) observed that participants in the top quartile were actually more modest in making their estimations than the participants who were in the bottom quartile. Participants in the bottom quartile were dramatically overconfident.

The sample of this study is made up of students who are placed in an academic development programme because they have been found to be academically unprepared in terms of their mathematics and science Grade 12 results and performance in the university's admission test. Being under-prepared suggests that their metacognitive knowledge of task and strategy variables such as content knowledge, understanding and application may be poor. This may influence the accuracy with which they evaluate their test performance.

(ii) Properties of the task

Lichtenstein and Fischhoff (1997) have made a distinction between types of knowledge, i.e. explanatory (or conceptual) knowledge which requires a higher level of understanding than the knowing of facts and procedures, and procedural knowledge which requires more factual knowledge. They have found that people are more accurate in evaluating their performance in providing correct factual knowledge than explanatory knowledge i.e. when asked to answer questions that require factual knowledge people tend not to be so overconfident than when asked to answer questions requiring them to explain.

(iii) Format selected for evaluation

Results obtained by Carvalho (2007) showed that the difference between actual and expected performance was significantly larger for multiple-choice than for short answer tests in undergraduate psychology. Students were more confident in multiple-choice than in short-answer tests, however their judgements were more accurate in the short-answer tests than in the multiple-choice tests, i.e. the elevated confidence levels did not match their performance in the multiple-choice tests. A possible explanation for this finding is that, multiple-choice tests require tasks of lower cognitive demand such as recognition, as compared to the higher demand of recall and self-construction of responses and this may tempt students into reduced metacognitive activity. Moreover short answer items require deeper engagement which forces the student to critically and accurately judge his/her performance on them.

(iv) Quality of feedback received

Irrespective of ample feedback, most people remain overconfident. Therefore Carter and Dunning (2008) have pointed to the contribution of missing information and deficits in the quality of feedback received towards error in performance evaluation. In fact Nowell *et al.* (2007) found that the grading practices of an instructor can be associated with overconfidence.

2.3.2 Personal factors

Even skilful students may rely on wrong information when assessing their performance and this may influence the accuracy of their performance evaluation (Ehrlinger, 2008). Literature has shown that sometimes underlying personal and psychological factors may explain respondents' bias in the evaluation of their performance.

(i) A tendency to rely too much on chronic self-views to evaluate performances

This is a tendency to draw on pre-existing perceptions of how skilled one is in a particular domain when one is asked to predict how well one has performed on any specific task. For example, “a doctor might evaluate the accuracy of a suspected diagnosis by drawing on her general perception of how knowledgeable she is about diseases and symptoms of that kind or even how skilled she is as a doctor in general”(Ehrlinger, 2008). Ehrlinger (2008) has also found that these self-views are in part a reflection of cultural beliefs and of wishful thinking.

(ii) A need for self-protection and self-enhancement

Gramzow *et al.* (2003) argued that the exaggeration of the grade point average of the previous semester by 68% of their sample of psychology students was motivated by either the students' need for self-enhancement or self-protection. Their prediction that students with low actual grades would exaggerate their GPAs to a greater extent than students with high actual grades was supported by results that showed a negative correlation between actual GPA and exaggeration. Gramzow *et al.* (2003) argue that exaggerated self-report by students with low actual grades are self-protective. They argue that students with low actual grades tend to self-protect in order to avoid the negative implications associated with acknowledgement of poor performance. On the other hand exaggerated self-reports by students with high actual grades are self-enhancing. They argue that students with high actual grades have a great need for achievement and hence report exaggerated self-reports when asked to recall and indicate their GPA from the previous semester. Results showed a positive correlation between exaggeration and need for achievement or self-enhancement.

The self-worth theory of achievement motivation explains the motivation of some students as attempts to enhance or protect self-worth (Seifert, 2004). The theory postulates that people possess a sense of self-worth which is a critical dimension of human functioning. Self-worth has to do with the judgement a person makes about his/her worth and dignity (Seifert, 2004). The western culture holds the belief that self-worth is related to performance from which the

belief that a person's ability to do something well is connected to a person's worth, emanates. There is no denying that our current system of education has been highly influenced by the western culture and its beliefs. Hence smart students are seen as those who obtain top grades in the school context and therefore deemed more worthy than their counterparts who do not do well in the academics (Seifert, 2004). For many students ability is the source of performance and performance a source of self-worth. Therefore it is the desire of every student to be deemed worthy. Students cannot afford to be seen as stupid or unable to perform academically not only because of the worth associated with academic excellence but because for most, how they perceive themselves is built on academic self-concept (Woolfolk, 1998). In the absence of actual performance, looking like one who possesses the capability to perform becomes a means with which poor performing students would protect their self-worth.

According to the affect mechanism of the self-worth theory great effort which results in failure implies low ability, leading to feelings of shame and humiliation which are feelings students would rather not experience. As a result the students may spend most of the time engaging in failure-avoiding strategies to avoid any implications of failure. These strategies include excuses and defence mechanisms students may use to protect ability perception in the event of failure. Procrastination, maintaining a state of disorganisation, setting goals too high or too low, cheating or asking for help are some of the behaviours associated with these strategies. The self-worth theory of achievement motivation helps us understand that students may resort to a self-protective process by reporting exaggerated assessments of their performance in order to gain favourable judgements of competence and also that students may become biased in the evaluation of their performance because no matter what their actual performance reveals, they do not wish to look incompetent (Seifert, 2004).

(iii) Theories of intelligence which respondents adhere to

Ehrlinger and Dweck (cited in Ehrlinger, 2008) distinguish between incremental theorists and entity theorists. The incremental theorists hold the belief that intelligence can be improved over time while on the other hand entity theorists believe that intelligence is fixed and unchangeable. In their study Ehrlinger and Dweck (cited in Ehrlinger, 2008) found that the incremental theorists were more accurate in their performance evaluation than entity theorists. The incremental theorists were motivated to learn and hopefully improve while the

entity theorists were determined to maintain a positive view of their fixed intelligence by choosing easier tasks to avoid feedback contrary to their self-views.

In his discussion of how motivation theories influence students' behaviours in academic settings, Seifert (2004) describes two types of students. These students differ in terms of two dominant goals they pursue. Students who pursue mastery goals are self-regulating, self-determining and their dispositions foster cognitive development. They believe that intelligence is malleable and a controllable factor just as the effort they put in their academic work ultimately determines success or failure. The mastery goal students are also less likely to deny responsibility for failure. On the other hand students pursuing performance goals are more concerned with ability and to them intelligence is a fixed entity. They are more concerned about how they perform relative to others and how others will perceive them. However, for students pursuing performance goals failure is a result of inability.

(iv) Personality traits

Different individuals possessing different personality traits have different dispositions towards overconfidence with some having a disposition towards overconfidence whereas others show the opposite trend. Pallier, Wilkinson, Danthiir, Kleitman, Knezevic, Stankov and Roberts (2002) in studying the independent metacognitive trait within the domains of personality and intelligence, found a small but significant relationship between the confidence factor and the personality constructs of proactiveness and activity which are the two variables associated with extraversion. In their study Schaefer, Williams, Goodie and Campbell (2004) asked a sample of psychology students to immediately report their level of confidence in the accuracy of their answers in a multiple-choice test. High levels of confidence reported for inaccurate answers were defined as overconfidence. Having categorised students into different personality types based on their scores on Goldberg's Big Five Personality Inventory, they found that among the five most commonly used measures of personality only extraversion and openness to experience/intellectance were positively correlated with overconfidence; however, they found that extraversion predicted confidence but not accuracy implying that extroverts are significantly overconfident. On the other hand openness to experience/intellectance predicted confidence as well as performance which means that the elevated confidence levels associated with this personality type accurately reflected their elevated performance.

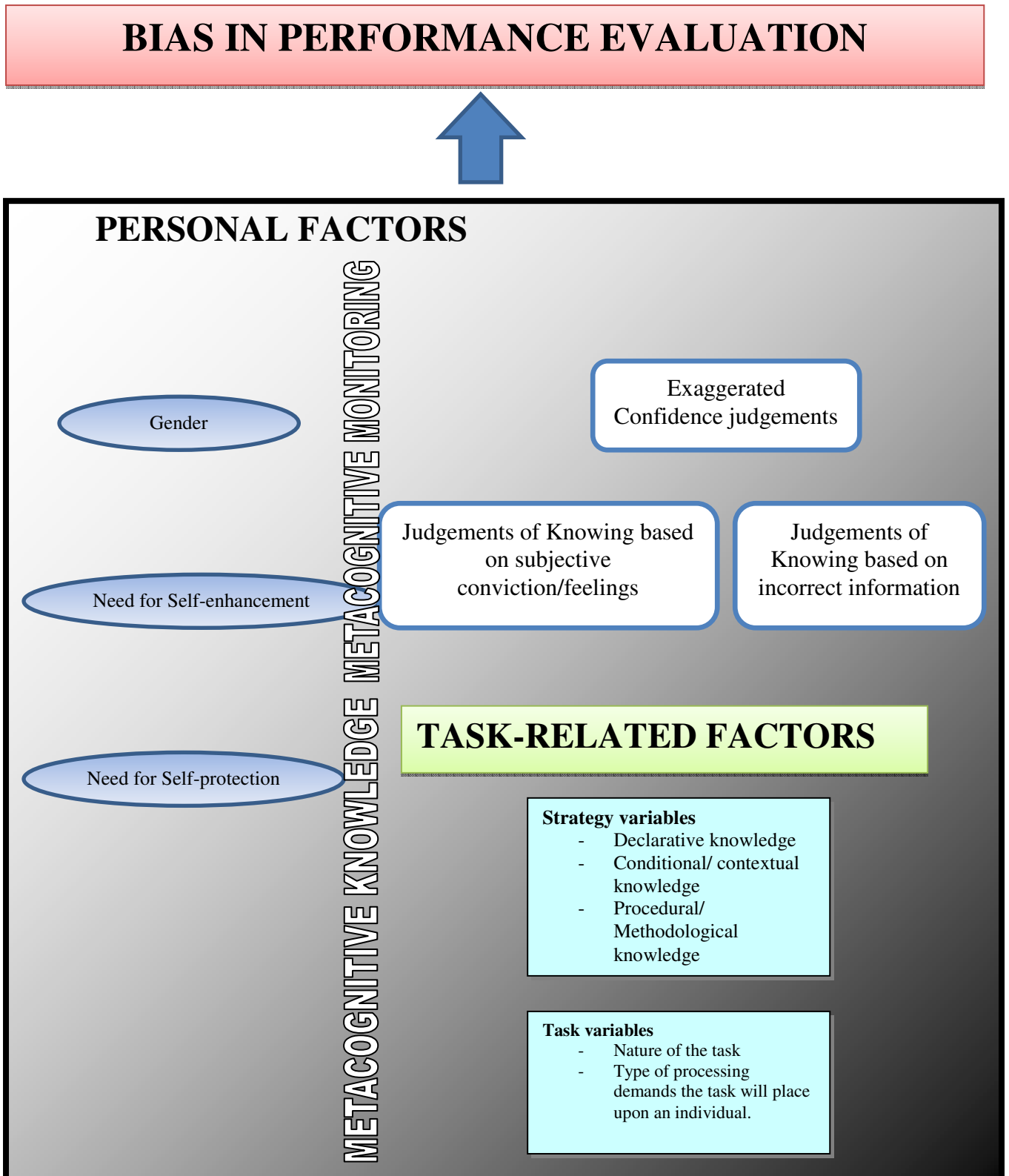
(v) Gender

Gender is another variable that may influence an individual's ability to accurately evaluate his or her performance. In their study Beyer and Bowden (1997) defined performance evaluations as post-task estimates of performance without the benefit of performance feedback. They found that women significantly underestimated their performance in masculine tasks while men accurately evaluated their performance. Hannover (1999) as cited by Beyer (1999) found that women significantly underestimated their performance in mathematics compared with men who reported an accurate evaluation. In a study conducted by Nowell and Alston (2007) overconfidence was defined as the difference between expected and actual grades. Economics male students appeared to exhibit greater overconfidence than female students.

2.4 CONCEPTUAL FRAMEWORK OF THE STUDY

The use of a conceptual framework as a benchmark assisted in guiding the investigation of the factors underlying students' bias in performance evaluation during test-taking. Based on the reviewed literature Figure 2.4 is a depiction of the conceptual framework of this study. The conceptual framework indicates that personal factors such as gender, the need to self-protect or self-enhance may underlie both metacognitive monitoring and knowledge variables. Metacognitive judgements made during the monitoring process may be associated with bias in performance evaluation and so is the knowledge of strategy as well as task variables. Data were collected and analysed to determine the association between bias in performance evaluation, personal factors, task-related factors and judgements made during metacognitive monitoring in our sample of students.

Figure 2.4: Conceptual framework of the study



2.5 INACCURACY OF PERFORMANCE EVALUATION: METHODOLOGY

Unlike investigating bias in self-evaluations with regard to attributes such as talent and being well adjusted, the advantage of investigating bias in self-evaluation with regard to performance is that this bias can be objectively verified (Carvalho, 2001; Carvalho & Yuzawa, 2001). Different procedures have been used to study bias in performance evaluation. In a study conducted by Gramzow, Elliot, Asher and McGregor (2003), subjects were asked to recall and indicate their grade point average, i.e. GPA from the previous semester. The grading scale used by the university where the study was conducted ranges from 0.0 to 4.0. The actual GPAs were obtained through the university's registrar and these were compared with the students' self-reported GPAs. An exaggeration index was calculated by subtracting each student's actual GPA from the self-reported GPA. On average students were found to have over-reported their GPA by over 1/10th of a point. Few students under-reported their GPA while many students over-reported their GPA. In fact three per cent over-reported their GPA by a full point or more.

In his study Carvalho (2007) asked psychology majors ($N = 129$) to make confidence judgements immediately after answering each test item. Students were asked to make ratings of how sure they were that each of their answers were correct on a Likert scale ranging from 0 to 100 per cent. The sum of all correctly answered items on a test was used as an indication of performance. There were 50 items in a test. Each correctly answered item received two points resulting in a total of 100 points when all the items were answered correctly. Accuracy in judgement was determined by subtracting the sum of correctly answered items from the average of confidence judgement ratings.

In their study Dunlosky *et al.* (2005) asked participants to predict their likelihood of correctly recalling studied items on a scale of 0 (0% chance of recall) to 100 (100% chance of recall) in a test. These predictions they called judgements of learning (JOLs). Participants were instructed to make immediate JOLs for half of the items in the test and delayed JOLs for the other half. The delay was for 30 seconds. Immediately after each JOL, participants were instructed to rate the confidence they had in the accuracy of the JOL made for each item, from 0 (definitely not accurate) to 100 (definitely accurate). The second type of judgement was called a second order judgement, i.e. SOJ. After the judgement-making phase of the study, participants were asked to recall studied items. The relative accuracy of a participant's JOLs was computed by correlating his or her JOLs with recall performance across items.

Dunlosky *et al.* (2005) found that the correlation between recall performance and delayed JOLs was greater than the correlation between recall performance and immediate JOLs. The relative accuracy was substantially greater for delayed than for immediate JOLs. Participants were more confident in their delayed JOLs than in their immediate JOLs and the high confidence in delayed JOLs was justified by higher recall performance. Dunlosky *et al.* (2005) further posit that the use of intermediate JOLs percentage values and overall low confidence in immediate JOLs suggest that participants were aware that their predictions were poor.

In the methodology employed by Sinkavich (1995) a sample of educational psychology students was instructed to choose the correct option in a multiple choice examination and for each item rate their confidence in the answer on a five-point Likert scale. The Likert scale had two anchor points noted as “not correct” (-2) to “correct” (+2). Zero which was the midpoint was interpreted as “maybe it is correct” or “maybe it is not correct”. To determine relative accuracy in calibration each individual student’s confidence ratings across all responses were summed up and a correlation was determined between the sum of confidence ratings and the total number of correctly answered items (examination score). Sinkavich (1995) found that students who expressed a higher degree of confidence had higher examination scores compared with students who expressed a lower degree of confidence (positive correlation). The confidence of good students was justified by performance in correctly answered items. Finally good and poor students differed significantly in their ability to predict what they knew and did not know. Good students were better in predicting their test item performance compared with poor students in multiple choice examinations.

Another methodology was the one employed by Ochse (2003) of UNISA. Before their final examination, Ochse (2003) asked a group of third-year psychology students to complete a questionnaire which amongst other things asked them to indicate a score they expected to obtain for the final examination of the module and on a Likert scale from 0% to 100%, indicate their confidence in obtaining the mark. After the writing of the examination, actual scores were compared with expected scores and students who had overestimated their actual score by nine or more marks were categorised as overestimators. Students whose expected mark was between nine marks above and nine marks below the actual mark were labelled realists while those who underestimated their actual mark by nine marks or more were categorised as underestimators. Overestimators on the whole, expected higher marks than the

realists and the underestimators were significantly more confident about the accuracy of their expected scores perceived themselves to have higher ability but however had obtained the lowest scores of the three categories. On average the overestimators failed.

In the study conducted by Ochse (2003) confidence judgement ratings were used to indicate confidence levels of the overestimators, realists and underestimators and in the study conducted by Dunlosky *et al.* (2005) confidence judgements were used twofold to indicate the likelihood of recalling studied items (JOLs) and the confidence participants had in the accuracy of their JOLs. In Carvalho (2007)'s and Sinkavich (1995)'s studies confidence judgement ratings were used to give an indication of students' expected performance which in turn indicated the students' perceived competency in correctly answering questions in a test.

As stated in paragraph 1.5 the aim of my study is to investigate the accuracy with which students evaluate their performance in a test. In this study students were required to make confidence judgements immediately after answering a question which was expected to be more sensitive than asking the student to indicate the judgement after completing the test. The methodology adopted for this study is therefore similar to that of Carvalho (2007) and Sinkavich (1995) but different from that of Dunlosky *et al.* (2005) because of the nature of metacognitive judgements involved (see Figure 2.1).

To objectively assess the students' ability to evaluate their performance, a test on a specific topic in chemistry had to be set which would be taken in a real classroom context. A decision to set the test on a difficult topic prior to instruction on the topic was deliberately made to investigate bias in performance evaluation before and after instruction. For the purpose of this study an accurate evaluation or judgement of one's performance is defined as a confidence judgement that can be justified by actual performance. The terms, metacognitive judgement, performance evaluation and judgement of performance will be used synonymously in subsequent chapters.

2.6 WHY USE STOICHIOMETRY TO STUDY BIAS IN PERFORMANCE EVALUATION

Much research on inaccurate metacognitive judgements has been conducted in fields or disciplines such as psychology, economics and computer science (Beyer, 1999; Ehrlinger,

2008; Goodie, 2003; Nowell & Alston, 2007; Schaefer *et al.*, 2003). Our literature review revealed that not much research has been done on the concept of inaccurate performance evaluation in the field of chemistry. Having tested her hypotheses on students' accuracy in performance evaluations, Beyer (1999) pointed out that failure of researchers to test their hypotheses in different courses and different disciplines may explain the current inconsistencies in research on practice effects.

The students in our sample come in weak and under-prepared from high school, i.e. they come in with low entry level knowledge and understanding. It is anticipated that when students are confronted with this difficult topic before receiving instruction in it and they are unable to solve the problems, the perceptions of their performance will be made clear and they will make use of their metacognitive skills to acknowledge and admit when they are unable to solve problems on stoichiometry. This will be revealed in their confidence judgement ratings which are indications of the confidence they have in the accuracy of their answers. Requiring them to evaluate their performance in an unprepared test on a difficult topic before and after instruction will enable us to identify the existence and extent of any bias of performance evaluation as well as the effect of teaching on performance and the accuracy of performance evaluation in a short period of time. After instruction better metacognitive judgments should be made. After all metacognitive skills are some of the skills that differentiate experts from novices.

Huddle and Pillay (1996) mentioned that the topics that cause the most difficulty for first-year chemistry students are chemical equilibrium, the mole, oxidation-reduction and reaction stoichiometry. In addition, electrochemistry was identified by Potgieter *et al.* (2007) as a difficult topic for first-year chemistry students. In the study by Potgieter *et al.* (2007), it became clear that the mole concept which is central to solving stoichiometric problems is still poorly understood by the majority of first-year chemistry students. Kolb (1978) stated that "there is probably no concept in the entire first-year chemistry course more important for students to understand than the mole and one of the main reasons the mole is so essential in the study of chemistry is stoichiometry".

Narrowing the focus of a study ensures that more reliable and clearer results are obtained. Therefore among the five topics namely chemical equilibrium, the mole, oxidation-reduction,

reaction stoichiometry and electrochemistry, which literature reports as the most difficult for first-year students, we have chosen stoichiometry as the main focus of our study.

First-year students find stoichiometry difficult because success in solving stoichiometry problems requires representational competence, formal reasoning and being able to work with multistep mathematical operations. Students enter the academic development programme with misconceptions or alternate conceptions of currently held scientific views and because of these, students may be convinced that they understand and they have done well in an assessment task although the contrary is true. The following properties of stoichiometry problems make stoichiometry an appropriate topic for use in a test instrument intended to assess students' ability to evaluate their test performance.

2.6.1 Stoichiometry and the multistep problem

Stoichiometry is a very mathematical part of chemistry, dealing with calculations on masses (sometimes volumes) of reactants and products involved in a chemical reaction. In stoichiometry students are not only required to demonstrate understanding of chemical reactions, but they must also be able to apply a thorough understanding of the principles involved in ratio and proportion calculations (Ben-Zvi, Eylon & Silberstein, 1988; Huddle & Pillay, 1996). In fact in their study, Potgieter *et al.* (2007) found the inadequacy of mathematical skills to be a factor which must be taken into consideration when teaching topics such as stoichiometry, gas laws, acids and bases and chemical equilibrium. McFate and Olmsted (1999) found that the two features present in some of the best discriminators in university placement tests, were items that required multistep mathematical operations and formal reasoning and stoichiometry does just that. For example in an assessment tool used by Huddle and Pillay (1996) during their quest to investigate the ability of chemistry students to solve problems involving stoichiometric concepts, students' answers to the following problem were analysed:

If the mineral phosphorite ($Ca_3(PO_4)_2$) is heated to $650^\circ C$ with sand (SiO_2) and Coke (C), the products are calcium silicate ($CaSiO_{3(s)}$), carbon monoxide, and phosphorus ($P_{4(g)}$).

Calculate the theoretical mass of P_4 produced if 6.2 kg phosphorite, 4.0 kg sand, and 1.0 kg coke are heated in a furnace to $650^\circ C$.

The problem required students to:

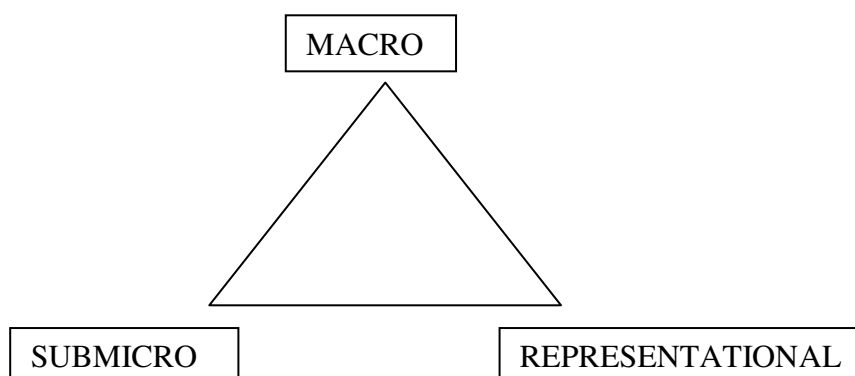
1. write the chemical formula of the substance whose name only was given e.g. carbon monoxide;
2. correctly write and balance the chemical equation;
3. determine the amount of each reactant present in terms of moles or ‘amount of substance’;
4. on the basis of the stoichiometry (stoichiometric coefficients) of the balanced equation, decide on the limiting reagent for the reaction and;
5. calculate the amount P_4 formed in moles and then finally convert it to mass in grams.

Formal reasoning is required in steps 1, 2 and 4 and the use of multistep mathematical operations in steps 3, 4 and 5. Multiple choice tests have been criticised for their use of tasks of lower cognitive demand such as recognition (Carvalho, 2007). However as a result of the necessity to make use of formal reasoning and multistep mathematical operations in order to solve stoichiometry problems, a test item such as the one used by Huddle and Pillay (1996) would discriminate well between students with different entry level knowledge. The process of solving such a stoichiometry problem would require deeper cognitive engagement forcing students into increased metacognitive activity. This may expose the student’s level of understanding and may aid the student in reaching and ultimately reporting an informed metacognitive judgement.

2.6.2 Stoichiometry and representational competence

The model by Johnstone (1991) shown in figure 2.5 suggested that the three thinking levels of chemistry i.e. the macro; submicro and representational levels were integrated and could be thought of as corners of a triangle.

Figure 2.5: Three levels of chemistry (adapted from Johnstone, 1991)



The macro level consists of what can be seen, touched and smelled. The submicro level consists of atoms, molecules, ions and structures. At this level the behaviour of substances is interpreted in terms of the unseen and molecular, and then recorded at the representational level using symbols, formulae and chemical equations. Johnstone (1991) asserts that the difference between an expert and novice chemist is the ease with which an experienced chemist can manipulate all three. Success in solving stoichiometric problems is dependent on students' being able to at least manipulate the last two levels namely; submicro and representational levels. However when the problem statement includes pictorial representations at the atomic or molecular level, students are challenged to reveal the level of their conceptual understanding. This should enable students to make accurate judgements of their understanding and ultimately their performance.

2.6.3 Stoichiometry and misconceptions

Jean Piaget, a Swiss psychologist whose descriptions of children's thinking changed the way we understand cognitive development, concluded that all species inherit two tendencies in thinking namely the tendency towards organisation and towards adaptation. The tendency toward organisation entails the combining, arranging, recombining, and rearranging of behaviours and thoughts in coherent systems whereas the latter involves a species adjusting to its environment (Woolfolk, 1998). Literature usually relates the tendency towards adaptation with misconceptions. The two processes involved in adaptation are assimilation and accommodation. Assimilation is what people do when they use their existing schemes to make sense of events around them. People try and understand something new by making it fit into what they already know and sometimes they may alter the information in order to make it fit. Accommodation on the other hand is when people must develop new structures or adjust their thinking in order to accommodate new information that cannot be altered to fit into their existing schemes of thinking (Woolfolk, 1998). When students are taught they try to make sense of the new information by incorporating it into their existing conceptual schemes by assimilation. However sometimes students have to distort the new information to make it fit and this gives rise to misconceptions. When the information cannot be distorted students are then forced to simply accommodate it by rote learning (Huddle and Pillay, 1996). Misconceptions pose a serious danger to students' performance and cognitive growth because Woolfolk (1998) states that sometimes the nature of people is that when they cannot assimilate or accommodate, they may choose to ignore the new information. When students arrive at tertiary institutions they already have alternate conceptions of currently held

scientific views, and literature has shown that once embedded in students' conceptual schemes, misconceptions are very resistant to remediation (Huddle & Pillay, 1996; Novak, 1988).

Hasan, Bagayoko and Kelley (1999) have defined student misconceptions as “strongly held cognitive structures that are different from the accepted understanding in a field and that are presumed to interfere with the acquisition of new knowledge”. Heller and Finley (1992) call them intuitive conceptions whereas Dykstra, Boyle and Monarch (1992) prefer the term alternative conceptions. However, be it strongly held cognitive structures, intuitive conceptions or alternative conceptions, misconceptions interfere with the acquisition of scientifically accepted conceptions in several domains and may have an effect on both accuracy of calibration and performance in specific domains. The point is that students are very confident about their understanding, but this confidence is misplaced or unjustified because their understanding is flawed. A number of students' misconceptions in stoichiometry have been identified in the literature.

(i) Problems with reactant ratios

Laugier and Dumon (2000) showed that when students feel the need to consider proportions in chemical change they fail to understand that the quantities to be taken into account are amounts of matter which imply using the mole concept. For example in the problem analysed by Huddle and Pillay (1996) students are supposed to recognise that in order to successfully solve the problem the quantities given in grams have to be converted to amounts of matter or ‘moles’ using the mole concept equation ($n = m/M$) where n represents the amount of matter, m represents the amount of substance in grams and M represents the molar mass or molecular weight of the substances. Having converted masses to moles students can now apply their knowledge of ratio and proportion by using the calculated moles and stoichiometric coefficients to determine mole ratios. However because concentration, mass or volume are often used instead of the amount of matter, Frazer and Servant (1986, 1987) found that students fail to establish relationships between different variables like amount of substance/moles(n), mass (m), molar mass (M), concentration (c) and volume (v).

(ii) Stoichiometry and the balancing of chemical reaction equations

From their findings Frazer and Servant (1986, 1987) inferred that successfully writing a balanced equation and interpreting stoichiometric coefficients provides the basis for success in solving stoichiometric problems. BouJaoude and Barakat (2000) found that some students

do not at all understand the significance of coefficients in a balanced chemical equation. For example Huddle and Pillay (1996) noticed that in trying to identify the limiting reagent students used the mole concept to calculate the number of moles of each reactant and then ignoring the stoichiometry of the balanced equation, decided that the reactant with the smallest amount in terms of moles was limiting. One student even wrote, “limiting reagent = least number of moles”. In addition, Furio, Azcona and Guisasola (2002) found that students confuse or do not know the definitions of and relationships between stoichiometric entities in general.

(iii) Limiting reactant and surplus of reactant

According to Gauchon and Méheut (2007) understanding the notion of limiting reactant and surplus of reactant can be considered a fundamental step in understanding stoichiometry. In spite of this Huddle and Pillay (1996) found that students cannot determine the ‘limiting reagent’ in a given problem, when one substance is added in excess. They found that there were students that assume that “limiting reagent” implies “lowest stoichiometry”. Sometimes students choose the limiting reactant randomly, without really justifying their choice. For example, in their study Boujaoude and Barakat (2000) found that students chose the limiting reactant as the reactant whose ‘amount of matter’ had been given in the question or the one whose mass is given or make their choice based on a comparison between the different molar masses.

2.7 CONCLUSION

Metacognition is a broad construct. This study focuses only on the monitoring portion of metacognition. A metacognitive judgement is an evaluation that one makes when he or she is asked to judge his or her performance. Various methods have been used to study the accuracy with which people judge or evaluate their performances. Overall poor performers have been observed to over-exaggerate their performance and this may have negative implications on teaching and learning (Carvalho, 2007; Ehrlinger, 2008). Construction of a JOK and a confidence judgement is often preceded by a FOK as evidence that a question will be successfully answered or a task will be successfully executed. Most studies in our literature review made use of a Likert scale according to which subjects could indicate their level of confidence, this guided the decision to use a Likert scale in our study. Since stoichiometry has been identified as one of the most difficult parts of chemistry for first-year students, it should not be expected that if under-prepared students are asked to make metacognitive

judgements and evaluate their performance in a test on the topic, overconfidence will be observed. However, if it is a good point of departure will be to determine factors associated with this overconfidence and to determine and examine the factors the students may have relied on while making their metacognitive judgements which may have led them to inaccurately assess their performance.

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

CONTENTS	PAGE
3.1 Introduction	41
3.2 Research paradigm	41
3.3 Research methodology	42
3.3.1 Embedded mixed methods design	43
3.4 Research design	46
3.5 Instrumentation	47
3.5.1 Procedure for the development and writing of the test items	49
3.5.1.1 Description of the stoichiometry test	49
3.5.1.2 Description of the three tiers in the stoichiometry test instrument	50
3.5.2 Procedure for the development and writing of questionnaire items	51
3.5.2.1 Description of the questionnaire instrument	51
3.6 Overview of the study	52
3.7 Pilot study	53
3.7.1 Purpose of the pilot study	53
3.7.2 Sample (pilot study)	53
3.7.3 Data collection	53
3.7.4 Validity of the data collection instruments	54
3.7.4.1 Content validity	54
3.7.4.2 Face validity	54
3.7.4.3 Construct validity	55
3.8 Reliability of the instruments	55
3.9 Main study	56
3.9.1 Sample (for main study)	56
3.9.2 Nature of the final data collection instruments	57
3.9.2.1 Chemistry test	57
3.9.2.2 Questionnaire	57
3.9.3 Data collection (main study)	57
3.9.4 Management of data	58

CONTENTS	PAGE
3.9.4.1 Criteria for inclusion of scripts with missing data	58
3.10 Procedures used to analyse the main study results	59
3.10.1 Quantitative data (Stoichiometry test instrument)	59
3.10.2 Quantitative data (Questionnaire instrument)	61
3.10.3 Qualitative data (Part three of the three-tier stoichiometry test instrument)	61
3.11 Ethical considerations	62

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

3.1 INTRODUCTION

The previous chapter focused on the review of relevant literature with regard to metacognition, metacognitive judgements, inaccuracy in performance evaluation, factors associated with bias in performance evaluation, and stoichiometry. This chapter outlines the research paradigm, methodology, design and the data collection methods that were used in the study. I describe the development and piloting of the data collection instruments as well as how the pilot study results helped inform the design and development of the final data collection instruments. Reliability and validity of the data collection instruments are discussed. The chapter concludes with the discussion of the main study, i.e. the nature of the final data collection instruments, the research sample, administration and management of the main study, the procedures used to analyse research results and ethical considerations.

3.2 RESEARCH PARADIGM

Epistemology has to do with the philosophy of knowledge and how we come to know that knowledge. It involves how we come to know reality while methodology which is closely related to epistemology involves the identification of particular practices through which knowledge can be attained (Krauss, 2005). Literature (Ivankova, Cresswell & Clark, 2007) recognises three approaches or methodologies to social science research, namely quantitative, qualitative and mixed methods. Each method or approach is characterised by its own purposes, methods of inquiry, data collection strategies, analysis and criteria for judging quality. The different methods also differ in terms of their epistemologies and theoretical paradigms concerning the nature of reality. A research paradigm can be defined as the world view or basic belief system that guides the investigation or research method (Guba & Lincoln, 1994). A research paradigm helps us identify the underlying basis used to construct a scientific investigation (Bogdan & Biklan, 1982). Qualitative research is based on a relativistic, constructivist philosophy of reality which assumes that there is no objective reality. Instead human beings who experience the phenomenon of interest construct multiple realities; therefore, measurement is not approached with the idea of construction of a fixed instrument with a fixed set of questions. Questions are allowed to emerge and constantly change as the researcher becomes familiar with the study (Krauss, 2005). Quantitative research on the other hand is based on a positivistic philosophy of reality. Independent facts

about a single reality are quantitatively measured. Data are being observed and therefore cannot change. Science is seen as the way to understand the world with the goal of being able to ultimately predict and control the world. The mixed methods approach is based on the realism philosophical paradigm. Realism or critical realism boasts the elements of both positivism and constructivism (Healy & Perry, 2000). Healy and Perry (2000) suggest that the difference between the three paradigms is that positivism is concerned with a single, concrete reality, Constructivism considers multiple realities while realism is concerned with multiple perceptions of a single reality. Krauss (2005) posits that realism recognises the differences that exist between reality and people's perceptions of reality. Researchers who choose to work from a realist perspective make use of a mixture of theoretical reasoning and experimentation to observe the phenomenon of interest in order to discover knowledge of the real world. In the critical realism paradigm both qualitative and quantitative methodologies are considered appropriate for researching underlying mechanisms that drive actions and events (Healy & Perry, 2000). Case studies, structured and semi structured in-depth interviews as well as statistical analyses are deemed acceptable and appropriate within the critical realism paradigm (Krauss, 2005). Cavaye (1996) posits that the methodology chosen should depend on what the research is attempting to do rather than be a commitment to a particular paradigm. The particular phenomenon of interest must therefore drive the employed methodology. Focusing on the phenomenon rather than the methodology enables the selection of an appropriate methodology (Falconer & Mackay, 1999). To allow for a thorough investigation of bias in performance evaluation, the factors underlying and associated with bias in performance evaluation as well as the effect of teaching on bias in performance evaluation, qualitative and quantitative research methods had to be employed in a mixed methods approach, placing our study in the critical realism paradigm.

3.3 RESEARCH METHODOLOGY

The combination of both the quantitative and qualitative research methods constitutes a mixed methods approach. Quantitative and qualitative methods complement each other. Even though they differ in terms of their methods of conducting inquiry, quantitative and qualitative methods can be used together in a mixed methods approach to enable the researcher to obtain an in-depth and a more complete analysis of the problem under study. The mixed methods approach allows the researcher to collect both numeric and text data concurrently or in sequence. This provides the researcher with time to choose variables and

units of analysis which are appropriate for the study's purpose and for finding answers to the research questions (Tashakkori & Teddlie, 1998).

According to Ivankova *et al.* (2007:278), "Mixed methods research is a procedure for collecting, analysing and 'mixing' both quantitative and qualitative data at some stage of the research process within a single study to understand a research problem more completely".

The quantitative and qualitative methods can be combined for four reasons:

- To use qualitative data to elaborate more on quantitative data obtained
- To collect qualitative data and use it to inform the development or design of a new measurement instrument which after being tested can yield quantitative data.
- To collect both quantitative and qualitative data simultaneously and compare it in order to arrive at a well-validated conclusion.
- To enhance a study by supplementing it with either qualitative or quantitative methods.

The four reasons result in four basic mixed methods designs, namely the explanatory, the exploratory, the triangulation and the embedded designs respectively. The difference in the designs is the sequence with which they collect quantitative and qualitative data and how they ultimately mix the two types of data (Ivankova *et al.*, 2007).

This study followed a mixed methods approach by implementing both qualitative and quantitative research methods to investigate and explore accuracy of performance evaluation of BFYP students. To best understand the research problem the embedded mixed methods design was followed to investigate accuracy of performance evaluation of BFYP students, the factors associated with bias in performance evaluation of it, as well as the influence of teaching on performance evaluation, the factors underlying students' bias in performance evaluation and performance.

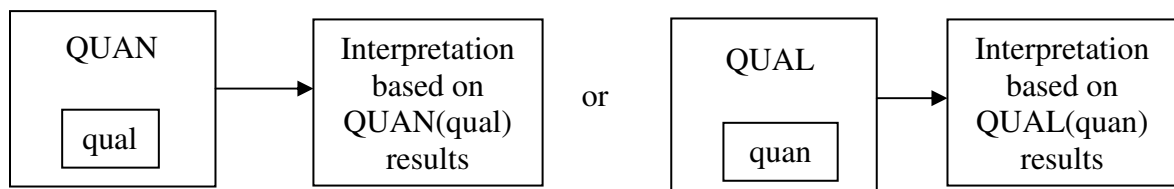
3.3.1 Embedded mixed methods design

Both qualitative and quantitative data are collected in the embedded design. However one of the data types takes on a secondary, supplemental role within the overall design. The embedded design is used when researchers recognise the need to include qualitative or quantitative data to answer a research question within a study largely based on quantitative or

qualitative approach (Creswell & Plano, 2007). The researcher may choose to embed a qualitative component within a quantitative design. For example a researcher may for particular reasons choose to embed qualitative data in an experimental design. Reasons for inclusion of qualitative data in a quantitative design may include development of relevant treatment, examination of the process of intervention or to follow up the results of an experiment (Creswell & Plano, 2007). The different data sets are mixed at the design level, with one type of data being embedded within a methodology guided by the other data type (Caracelli & Greene, 1993).

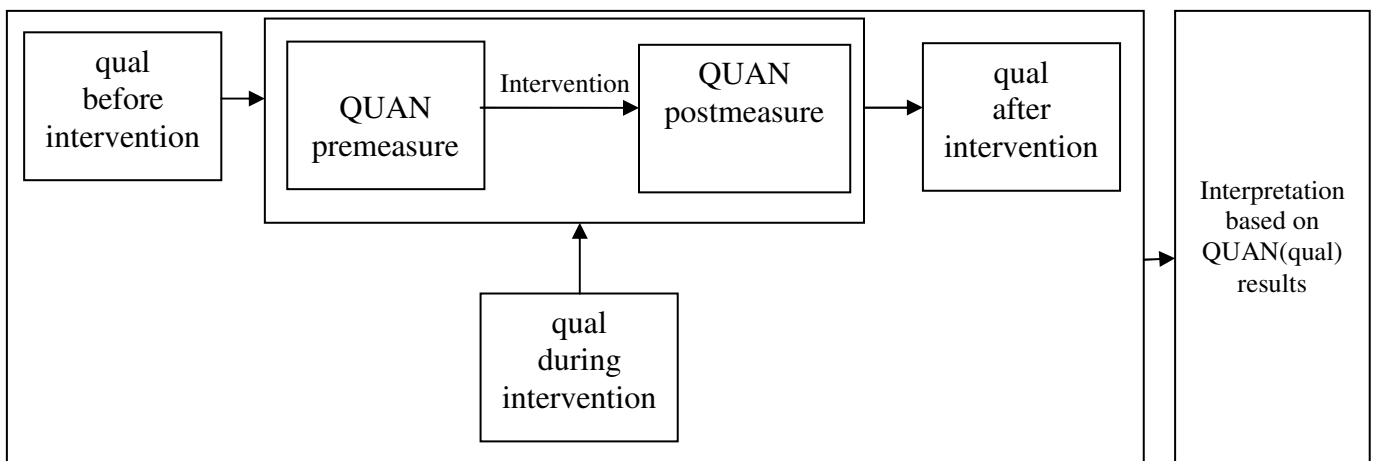
Creswell and Plano (2007) use the following diagram to represent the embedded mixed methods design:

Figure 3.1: Embedded mixed methods design (adapted from Creswell and Plano, 2007)



There are many variations of the embedded design. A typical use of the embedded design can be observed when a primarily quantitative experimental study with a group receiving treatment and the control group receiving none, has a qualitative portion which includes an in-depth interview of the participants before, during or after the treatment, embedded into the study. Creswell and Plano (2007) identify such a study as the embedded experimental model.

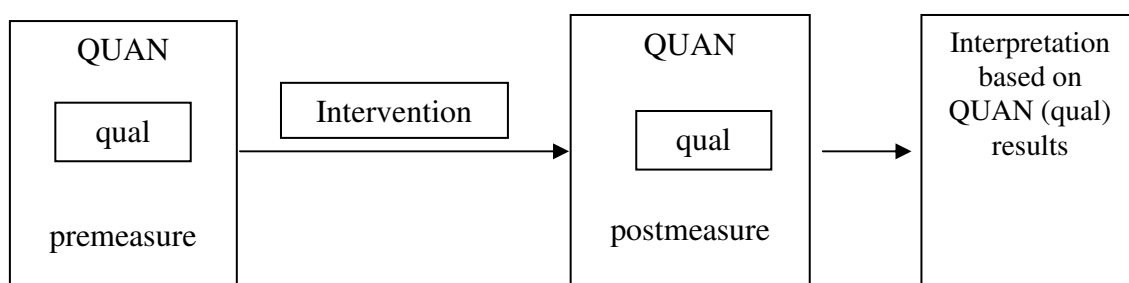
Figure 3.2: Embedded experimental model (adapted from Creswell and Plano, 2007)



This model is achieved by having qualitative data embedded into an experimental design. The primary focus is on the quantitative portion with the qualitative data in a secondary, supplemental role. The model may be used in either a one-phase or two-phase approach. A typical one-phase approach would include embedding the qualitative data in the intervention phase of the study, enabling the researcher to qualitatively examine the intervention process. In the two-phase approach qualitative data may be embedded before or after the intervention phase. Creswell and Plano (2007) suggest that the two approaches can be instrumental in enabling the researcher to acquire qualitative information prior to the administration of the intervention, shape the intervention, develop an instrument, select participants, explain the results of the intervention or follow up the experiences of participants with particular types of outcomes. The design may be used as either a one-phase or two-phase approach depending on the purpose for inclusion of qualitative data.

In this study quantitative data includes raw test scores indicating students' mastery of chemistry content knowledge and associated process skills, students' level of confidence in the accuracy of their answers in the test as well as data on the motivational factors associated with inaccuracy of performance evaluation. Qualitative data constitutes factors underlying students' level of confidence in the accuracy of their answers. Qualitative data in the form of students' free responses to open-ended questions are embedded into a quantitative chemistry test instrument within a quantitative experimental design. This study uses the embedded experimental model which may be depicted by a diagram which is slightly different from the one indicated in Creswell and Plano's model (2007).

Figure 3.3: Embedded Experimental model used in the study



Quantitative and qualitative data are collected simultaneously before and after the intervention phase. Interpretations are then made based on the quantitative and qualitative results. However like any other design the embedded design has advantages and limitations.

The advantage of the embedded design is that it can be used when a researcher does not have enough time or resources to commit extensively to quantitative and qualitative data collection because one data type is given less priority than the other. There are limitations because two types of data sets are collected, and the researcher must make sure to specify the purpose of collecting qualitative or quantitative data within a larger quantitative or qualitative study. To address this limitation the purpose of collecting qualitative data has been provided in paragraph 3.4. The other limitation cited by Creswell and Plano (2007) is that it may be difficult to integrate the results when the two methods are used to answer the different research questions. To address the second limitation, the two sets of results were analysed and reported separately and only at the end were the two results integrated to achieve an understanding of how the qualitative results explained the quantitative results in the pre- and posttest.

3.4 RESEARCH DESIGN

Nieuwenhuis (2007: 70) defines research design as “a plan or strategy which moves from the underlying philosophical assumptions to specifying the selection of respondents, the data-gathering techniques to be used and the data analysis to be done.” Because this study followed a mixed methodological approach a quantitative and a qualitative design had to be combined. Quantitative data were collected through a questionnaire and a chemistry test and qualitative data by means of open-ended sections in the chemistry test.

In Cohen, Manion and Morrison (2007) a case study is defined as a study that focuses on an instance in action. The instance can be a child, clique, class, school or a community. These are referred to as bounded systems. The strength of a case study is that it observes effects in real contexts, i.e. a case is studied in its natural real-life context or setting. Because observations are done in the natural setting, detailed descriptions can be obtained. The disadvantage of a case study is that data obtained are both subjective and objective. The researcher has to be part of the research process and consequently the results are biased.

A case study design allows the use of multiple sources and techniques in the data collection process. Data collection is not only limited to qualitative approaches but can also be quantitative. Surveys, interviews, documentation reviews and observations can be used to collect data. The limitation of a case study design however is that data obtained cannot be used to arrive at a generalising conclusion, i.e. findings based on data from the sample cannot

be generalised to the entire population. Nieuwenhuis (2007) asserts that the aim of a case study is that greater insight and understanding of the specific situation and sample are gained. The advantage however, of a case study, is that anyone reading the report can identify with the findings.

To prevent challenges posed by the chosen design's limitations, biased views and interpretations should be continuously checked by triangulation. The study does not intend to generalise but its intention is to provide rich descriptions of the sample in a real-life context with the objective of informing rather than generalising.

3.5 INSTRUMENTATION

In the following paragraphs an explanation of how the research questions were used to guide the design and development of data collection instruments is presented.

Research question 1: How accurately do BFYP students evaluate their performance in a stoichiometry test?

The objective of research question 1 is to determine how accurately BFYP students evaluate their performance in a stoichiometry test. To investigate the accuracy with which students evaluate their performance in a test on stoichiometry, a chemistry test on stoichiometry had to be designed and students had to be given the opportunity to evaluate their performance on each test item before and after instruction as shown in Figure 3.3. Similar to the scale used by Osche (2003), a Likert scale ranging from 0% to 100% was provided per item for students to evaluate their performance by choosing a rating which indicated the level of confidence they had in the accuracy of their answers.

In the light of what literature had alluded to; to design and construct a data collection instrument that would not make it easy for students to be overly biased towards over-confidence, careful attention needed to be given to the choice of test format, i.e. multiple choice versus short-answer test, the inclusion of different types of questions that assess declarative knowledge, procedural knowledge and conceptual understanding as well as the students' misconceptions. Consideration had to be given to the fact that students had been found to report accurate evaluations of their performance in the short-answer tests because

these answers required deep and higher order thinking skills rather than recall (Carvalho, 2007).

Research question 2: What is the influence of teaching of stoichiometry in the BSc Four-year programme on performance, accuracy of performance evaluation?

Kruger and Dunning (1999) found that when metacognitive skills of poor performers were improved through careful training their accuracy of calibration improved significantly. Therefore the students' lack of knowledge, misconceptions and quality of feedback would need to be addressed during teaching and the students reassessed in order to determine the extent of the influence of teaching on performance and accuracy in performance evaluation. Having the students take the chemistry test before and after instruction would allow us to investigate the effect of teaching on performance and bias in performance evaluation.

Research question 3: What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?

Requiring students to explain their choice of confidence rating would provide us with detailed descriptions which could be analysed to identify the factors the students rely on when they make their confidence judgements. The explanations of the students provided in the posttests would be compared with the explanations provided in the pretest to determine whether shifts could be observed in the reliance on these factors after teaching. An additional tier would have to be included to each item in the chemistry test for this purpose.

Research question 4: What is the relationship between bias in performance evaluation and self-enhancement, self-protection and gender?

From literature it became clear that lack of knowledge or skill might not be the only factor influencing bias in performance evaluation. Personal factors might also be associated with bias in performance assessment. Gramzow *et al.* (2003) argued that exaggerated self-reports might be a reflection of self-protective or self-enhancement motivations. They argued that overly-positive self-reports by students with low actual grades were motivated by the need to self-protect while exaggerated self-reports by students with high actual grades were self-enhancement motivated. Students with low actual grades tended to self-protect in order

to avoid the negative implications associated with acknowledgement of poor performance and students with high actual grades had a great need for achievement and hence reported exaggerated self-reports. In light of Gramzow *et al.* (2003)'s findings we therefore set out to examine whether self-protection and self-enhancement predicted bias in performance evaluation. The study also investigated whether gender differences in the accuracy of performance evaluations existed. Underlying personal and psychological factors associated with bias in performance evaluation warranted the design of an instrument which would enable the collection of rich data through which these factors could be seen emerging or manifesting themselves. This called for a data collection instrument which could assist in investigating the students' tendency to self-protect or self-enhance.

To investigate the tendency to self-protect and self-enhance as well as how these two constructs are related to bias in performance evaluation a data collection instrument comprising items for which reliability and validity had been established in previous studies, would have to be used.

3.5.1 Procedure for the development and writing of the test items

First-year chemistry textbooks used in the BFYP, research literature as well as the internet were used to identify suitable test items on stoichiometry. Research literature on stoichiometry assisted in identifying misconceptions students are reported to have in the topic and these were used to inform the choice of items included in the instrument. Some of the items were modified before inclusion in the test instrument. They were converted to particulate drawings in order to reveal problems with visualisation skills and conceptual understanding. The test comprised a total of twenty stoichiometry and stoichiometry-related multiple choice questions.

3.5.1.1 Description of the stoichiometry test

The test displayed in Appendix 1 comprised items of moderate level of difficulty requiring conceptual understanding, procedural and declarative knowledge, and the use of formal reasoning as well as multistep mathematical operations to solve the problems. Eight out of twenty items measured procedural knowledge, formal reasoning as well as numeric problem-solving skills using multistep mathematical operations. Only two items assessed declarative knowledge. To reveal conceptual understanding, eight items incorporating submicroscopic representations or particulate drawings of atoms and molecules assessed representational

competence. One item assessed students' ability to use knowledge of symbolic representation of atoms and molecules in a balanced chemical equation to interpret graphical representation of a chemical reaction.

3.5.1.2 Description of the three tiers in the stoichiometry test instrument

Each item in the test instrument was divided into three tiers. The first tier comprised a stoichiometry multiple choice question.

Part 1: Multiple choice question

Given the equation $3A + B \rightarrow C + D$, if 4 moles of A reacted with 2 moles of B, which of the following is true?

- The limiting reactant is the one with the higher molar mass.
- A is the limiting reactant because you need 6 moles of A and have 4 moles.
- B is the limiting reactant because three A molecules react with every one B molecule.
- B is the limiting reactant because there are only 2 moles of B available.
- Neither reactant is limiting.

In order to investigate students' accuracy of performance evaluation students were requested to indicate how sure or confident they were that their chosen multiple choice option was correct, and this they had to indicate on a Likert scale from zero to hundred per cent.

Part 2: Indication of Confidence judgement rating

How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

Lastly, to provide us with rich data which could help us to understand the motivation behind students' choice of confidence indicators, in the third tier students were given an opportunity to explain their choice of confidence judgement ratings.

Part 3: Explanation for choice of confidence indicator

Why did you choose that specific confidence indicator?

3.5.2 Procedure for the development and writing of questionnaire items

Six items selected from the self-enhancement motives instrument developed by Yun and Takeuchi (2007) were modified for the African context and culture and used for the measurement of self-enhancement. An additional four items selected from Gramzow's Personality Research Form were modified and added to the six to make a total of 10 items measuring the same construct. Items selected from Gramzow's instrument were deemed appropriate because using this instrument Gramzow (2003) found that bias in performance evaluation demonstrated by top-performing students correlated with self-enhancement.

With the help of an article by Martin, Marsh and Debus (2003), self-protection items were generated. Nine items measuring the construct were used, resulting in a 19-item questionnaire. An additional 11 items on self-regulated learning were added as spacers resulting in a 30-item questionnaire, appearing in Appendix VI. The 30 items were randomly interspersed throughout the questionnaire. Subjects were also requested to report their gender, age and student numbers.

3.5.2.1 Description of the questionnaire instrument

Scales are useful in measuring how respondents feel or think about something. In a scale response options are set up in a way that the variables being measured can be expressed as numerical scores (Maree & Pietersen, 2007). Most commonly used scales are Likert scales and semantic differential scales. The difference between the two scales is that the Likert scale provides an ordinal measure of a respondent's attitude and the semantic differential scale uses adjectives to measure how a respondent feels about a certain concept. Since most adjectives have polar opposites, the semantic differential scale uses these opposites such as "bad" or "good" to create a numerical measure of a particular concept (Maree & Pietersen, 2007). The objective of this study was to investigate students' tendency to self-protect or self-enhance and the association of these constructs with inaccuracy of performance evaluation. Students were therefore presented with statements with which they had to agree or disagree. The purpose of the statements was to gauge students' strength of feeling. Ideally students would just state whether they agreed or not with no possibility of being neutral, but to accommodate a variety of responses representing various strengths of feelings, provision had to be made for students who somewhat agreed or disagreed or who preferred to take on a neutral or undecided stance as far as a specific statement was concerned. The Likert scale was therefore deemed suitable for the purpose of this study.

The Likert scale used in our questionnaire instrument comprised the following seven categories:

1. Strongly disagree
2. Disagree
3. Disagree somewhat
4. Undecided
5. Agree somewhat
6. Agree
7. Strongly agree

According to Maree and Pietersen (2007), the most convenient way to measure a construct can be achieved by using a Likert scale. The objective was to measure two constructs, namely self-protection and self-enhancement. Ten self-enhancement as well as nine self-protection items stated were used. We were careful to make sure that all the items in each construct were stated in the same direction so that agreeing on all the self-enhancement or self-protection items produced the same meaning, i.e. the respondent exhibits the tendency to self-enhance or self-protect. Items 1, 6, 8, 11, 15, 18, 20, 23, 26, 29 measured the construct of self-enhancement and items 2, 5, 7, 12, 13, 16, 21, 24, 27 measure the self-protection construct. In addition two biographical questions, namely, age and gender, were asked to determine the profile of the sample and also to explore possible relationships between gender and other variables in the study such as inaccuracy of performance evaluation and the two constructs measured through the questionnaire. The questionnaire appears in Appendix VI.

3.6 OVERVIEW OF THE STUDY

This study was conducted over a period of three years. Figure 3.4 provides an overview of the whole study. Activities and events are depicted in chronological order.

Figure 3.4: Overview of the study

Designing of data collection instruments	Pilot Study	Refining of data collection instruments	Pretest	Instruction on stoichiometry	Posttest	Data Processing	Data Interpretation	Reporting of Results
2008 - June 2009	01 July – 20 August 2009		28 August '09	September '09	12 October '09	October 2009 – June 2010		July 2010 – April 2011

3.7 PILOT STUDY

3.7.1 Purpose of the pilot study

The pilot study was used to collect as much data as possible to inform the development of the stoichiometry test, to determine the validity of the instrument as well as to refine it for final data collection. The pilot study was conducted to also gauge whether the test could be completed within the allocated time period.

3.7.2 Sample (pilot study)

Four ex-engineering students admitted to the BFYP programme in the second semester took the test. This sample resembled the sample for which the test was intended in that they were identified as weak students since they had failed to cope in mainstream and hence had to be admitted into the BFYP during the second semester of their first-year. Airtime vouchers to the value of R30.00 were used as incentive. The test was also taken by high school teachers teaching physical science at Grade 12 level in good schools in Pretoria. One university lecturer teaching students in a similar programme in the education faculty also took the test.

3.7.3 Data collection

Four students were given the test to complete within one hour. The students were informed that their responses were important for research purposes and that their responses would help to inform the design of the final data collection instruments which would be used in the main study of the research. The aim and purpose of the research was clearly communicated to the students and their consent was obtained for the process. In addition three high school teachers and one university lecturer were also given the test to complete in their own time. However in

order to control the time spent on the test, the educators were requested to report the time it took them to complete the test.

3.7.4 Validity of the data collection instruments

The extent to which an instrument measures what it intends to, is referred to as validity (Pietersen & Maree, 2007). In this study two instruments used for different purposes are used namely the stoichiometry test instrument and the questionnaire instrument. The first instrument takes on the form of a multiple choice test used to measure students' performance in a stoichiometry test. The second instrument is a questionnaire used to measure specific constructs relating to the students' tendency to self-enhance and self-protect. The stoichiometry test was checked for content validity while the questionnaire was checked for construct validity. Face validity was assessed for both instruments.

3.7.4.1 Content validity

Content validity refers to the extent to which the instrument covers the complete content of the specific construct that it intends to measure (Pietersen & Maree, 2007). This validity assesses whether in developing the instrument, all items that cover the different aspects of the measured construct were included in the instrument. For example if the test measures competence in solving stoichiometry problems, all the different aspects of stoichiometry should be included and assessed in the test. Content validity of an instrument is usually assessed by experts in the field before the instrument can be finalised (Pietersen & Maree, 2007). In our study four experts, i.e. three high school teachers and one university lecturer, were invited to assess the content validity of our instrument.

3.7.4.2 Face validity

Face validity on the other hand refers to the degree to which an instrument appears to measure what it is supposed to measure. This was crucial to our test instrument because in addition to textual information, students were given pictorial information in the form of submicroscopic diagrams to assess and expose their conceptual understanding. Therefore the face validity of such diagrams had to also be assessed. Thus experts in the field were invited to simultaneously assess both content and face validity of the stoichiometry test instrument. Moreover after taking the test, the students to whom the test instrument was piloted were instructed to complete a brief questionnaire on the test (available in Appendix II). The questionnaire required students to report on any ambiguities or potential language barriers

that would cause a second-language respondent to misunderstand and comment on the clarity of instructions, vocabulary as well as terminology used in the test. Students had to say whether the level of the test was appropriate for the students for which the test was intended. The face validity of the questionnaire was assessed by my two supervisors.

3.7.4.3 Construct validity

Construct validity has to do with how well groups of related items measure the constructs covered by an instrument. Statistical techniques used to measure construct validity are factor and item analysis. In our study factor analysis was employed to ascertain whether the response patterns indicate that only two constructs were measured, namely, self-enhancement and self-protection. Factor analysis is used to determine which items “belong together” in a sense that when answered in a test or questionnaire, they are answered similarly indicating that they measure the same factor (Pietersen & Maree, 2007). Pietersen and Maree (2007) further posit that this type of analysis is well suited for items measured on a 5- or 7-point Likert scale. For these reasons, factor analysis was deemed necessary for our study, since our items were measured over a 7-point Likert scale. We also wanted to verify that all items designated to measure self-protection and items meant to measure self-enhancement would all do so.

When a factor analysis is performed on a set of items, it produces a factor loading matrix as its primary output. This matrix contains, for each item, a loading on each factor in a form of correlation coefficients between items and factors. Correlation coefficients greater than 0.25 or big values are indicative of which items belong to which factor (Pietersen & Maree, 2007). A correlation matrix is also produced and it aids in identifying intercorrelations between items and factors. This matrix is useful in revealing whether or not factors measure the same construct.

3.8 RELIABILITY OF THE INSTRUMENTS

Reliability of a test or a questionnaire is an indication of the extent to which the instrument is likely to result in consistent scores. Another way of looking at reliability is that if the same person were to complete a questionnaire or test, that person should obtain the same score on that test or questionnaire as if they completed it at two different times (Field, 2009). A test or questionnaire with more items is likely to have a higher reliability and a test on more diverse subject matter is likely to have a lower reliability. Reliability is usually represented in terms

of Cronbach's alpha or Kuder-Richardson values. As a guideline for classroom examination, a Cronbach's alpha between 0.70 and 0.80 is good and a Cronbach's alpha between 0.60 and 0.70 is somewhat low and an indication that the test needs to be supplemented or some items need to be improved (Nunnally, 1967). However classroom tests need not be stringent in terms of reliability as they are combined with other scores to determine grades. High reliability should be demanded in cases when a single test score is used to make major decisions. According to Kline (1999), 0.8 value is generally accepted for intelligence tests; however a cut-off point of 0.7 is suitable for ability tests. In terms of questionnaires, the wording of items needs to be formulated in such a way that the meaning of related items will be "closer" to each other and to the construct they intend to measure. This will result in higher correlations between the items and eventually a better reliability coefficient (Pietersen & Maree, 2007). High reliability of data collection instruments is also required in research studies to ensure that quality data are obtained for interpretations. To establish reliability of the test and questionnaire instruments Cronbach's alpha coefficients were computed using data collected during the main study.

3.9 MAIN STUDY

Data collection of the main study was conducted in August and October in 2009. Various data collection instruments were used to enable the collection of relevant and sufficient data to answer the research questions.

3.9.1 Sample (for main study)

The sample of the study comprised students in the BFYP. The students completed their first semester chemistry module, CMY 133 and were now enrolled for their second semester chemistry, CMY 143. Some students who were registered for engineering in the first semester also joined this group in the second semester. However, only the data of 91 students (35 males and 55 females) with a median age of 19 years were analysed based on the reasons provided in paragraph 4.1. These students share commonalities in that they were all taught stoichiometry by the same lecturer in the large group lectures and they had to complete compulsory computerised quizzes on the topic. In addition they attended small group lectures or tutorial sessions where they had plenty of problem-solving and constant feedback opportunities (see paragraph 1.2.1 for details of the model of teaching followed in the programme). Even though they had different lecturers with different teaching styles in the

small group lectures, the lecturers worked collaboratively in terms of the material used in the lectures as well the quantity, format and content of tests and tasks given to the students.

The final test instrument was administered twice, i.e. both as pre- and posttest (in the main study). The questionnaire was only used once to collect data during the pretest. One hundred and seventy students wrote both the pre- and posttest. The data and results obtained through these data collection instruments will be presented in chapter four.

3.9.2 Nature of the final data collection instruments

In order to answer the research questions posed in this study the following data collection instruments were used.

3.9.2.1 Chemistry test

After making the necessary modifications drawn from the pilot study a 20-item, three-tier paper and pencil test was developed. See Appendix I. Each item consisted of a stoichiometry or stoichiometry-related multiple choice question with only one option correct and the other options serving as distractors. Despite the weakness associated with the multiple choice format, the format was chosen because a test in this format is easy to administer, mark and statistically analyse. This is especially important for large groups similar to the one used in this study. A multiple choice test can be taken in a short period of time. In addition the scoring of multiple choice questions is accurate and objective as opposed to open-ended questions (Higgins & Tatham, 2003). The analysis of chosen distractors can aid in the identification of misconceptions. In addition, boxes for coding were added per item for easy data coding and data capturing.

3.9.2.2 Questionnaire

Appendix VI shows the 30-item questionnaire which the participants had to complete in the main study. In addition to responding to items based on self-enhancement and self-protection, participants were requested to report their age and gender.

3.9.3 Data collection (main study)

The final data collection instruments were administered in the second semester of the BFYP. The pretest and questionnaire were administered prior to students receiving instruction on stoichiometry. Students were given two hours to complete both the pretest and the

questionnaire instruments and the time allowed was adequate for every participant. A fellow colleague was asked to help with the administration of the data collection instruments for both groups. After six weeks of exposure to lecturing and teaching on stoichiometry and tutorial sessions in which students were given ample opportunity for guided and unguided problem-solving, the same test instrument was administered as a posttest. Since only the test instrument was administered, students were given one hour to complete the test.

3.9.4 Management of data

Raw data from the pretest, posttest and the questionnaire were electronically captured by a data capturer provided by the university. Before data analysis, electronic data was returned to the researcher to check for any mistakes that might have occurred during the data-capturing process. Any mistakes made during data capturing were identified and corrected. The nature of the test instrument was such that each item had to have three responses and as a result scripts with missing data were submitted. For some the missing data could be inferred from the available data. Therefore objective criteria for inclusion of scripts with missing data had to be drawn up.

3.9.4.1 Criteria for inclusion of scripts with missing data

The following criteria were drawn up and used to include scripts with missing data:

1. Multiple choice answer not selected

1.1 *Criterion #1*: Script **included** if:

A very low or 0% confidence judgement rating was indicated and an explanation given for choice of confidence rating was one of the following:

- a. I forgot the topic;
- b. I cannot calculate;
- c. I don't know;
- d. I don't know or remember.

Answer inferred from explanation would be coded as the letter Z

(Z = would have been an incorrect multiple choice option if it had been chosen by respondent)

1.2 *Criterion # 2*: Script **excluded** if:

Confidence was indicated but the explanation for choice of confidence indicator was omitted; or confidence rating indicated was 100% and an explanation for choice of confidence indicator was given. This could be attributed to students' negligence.

2. Both confidence and answer missing

2.1 Criterion #1: Script included if:

- Explanation given for the choice of confidence indicator was for example “I don’t know”

Answer inferred from the given explanation: Z

Confidence inferred from the given explanation: 0%

2.2 Criterion # 2: Script excluded if:

Confidence was not indicated, answer was not selected and explanation was not provided or items on the last page were unanswered possibly because of negligence on the part of the student or items in the middle were unanswered possibly because these items had been skipped on purpose or by accident.

3. Confidence indicator missing

3.1 Criterion #1: Script included if:

Answer was selected and an explanation given for choice of confidence indicator was one of the following:

- a. “I forgot the topic”
- b. “I don’t know”

Confidence inferred from the given explanation: 0%

3.2 Criterion #2: Script excluded if:

All other scripts with missing confidence indicator excluded even if the answer and the explanation had been given. There was no way of knowing what confidence indicator could have been chosen even if the chosen option was correct and the explanation showed that the student believed the option was correct.

3.10 PROCEDURES USED TO ANALYSE THE MAIN STUDY RESULTS

3.10.1 Quantitative data (Stoichiometry test instrument)

The multiple choice test data for both the pre- and the posttest were scored. Correct answers were awarded a score of 1 and the incorrect answers 0. A total score out of a maximum of 19 was then calculated and converted into a percentage value. An explanation for scoring the tests out of 19 instead of 20 is provided in chapter four, paragraph 4.2.1.2. The confidence judgement ratings were reported per item in the test on a scale of 0 to 100%. The confidence judgement ratings per item were calculated to determine the average of confidence judgement ratings for each individual student’s test. The difference between the score obtained in the

multiple choice test (as a percentage) and the average confidence score was used to categorise subjects as realistic, overconfident or under-confident before and after instruction. Test scores out of 19 were converted to a percentage, then subtracted from the percentage average confidence to determine accuracy of judgement. The following equation was used to determine accuracy of performance evaluation:

$$\Delta = \text{Average confidence score (\%)} - \text{Test score (\%)}$$

The sample in our study comprised weak, under-prepared students and poor students have been known to exhibit high levels of overconfidence when evaluating their performance. The students were expected to evaluate their performance on a difficult topic like stoichiometry which lends itself to misconceptions. Misconceptions may contribute towards students' tendency of being biased towards overconfidence when asked to evaluate their performance in a test, because in their study Hasan *et al.* (1999) interpreted highly exaggerated confidence levels as an indication of the presence of strong alternative conceptions making the students feel confident about their choices even when these were incorrect. A test on specifically, stoichiometry was utilised to afford the students an opportunity to use their performance in the test as a benchmark in order to generate a clear, objective perception of their ability to solve stoichiometry problems and as a result make accurate evaluations of their performance. Taking all these factors into consideration we were more than generous to allow the students leeway to misjudge their performance on a maximum of three out of 19 items. Even though the number three was chosen arbitrarily, allowing the students to misjudge their performance on more than three items in a 19-item test could have been too generous. Even the slightest errors should not be tolerated in exams. In test situations 1% can be the difference between a pass and a failure. Learners are also expected to demonstrate the ability to monitor their own learning and this they can demonstrate in the accuracy with which they evaluate their performance. In addition the adoption of a wide margin of error held the prospect of enabling the identification of real outliers in terms of performance judgement. Converted into a percentage value this resulted in a 15.8% margin of error in performance perception. In a hypothetical test if a student obtained a test score of 42% and an average confidence score of 59%, the difference between the two scores would be 17% which is more than the allowed error margin of 15.8%. This would then be interpreted as an indication of overconfidence.

In summary subjects whose average confidence score exceeded the test score by more than 15.8% were labelled as overconfident. The realistic group were subjects whose average

confidence scores were between 15.8% and -15.8% (-15.8% and 15.8% included). Subjects whose test scores exceeded their average confidence scores by more than 15.8% were labelled as under-confident. The number of subjects in the realistic, overconfident and under-confident groups in the pre- and the posttest were compared.

To aid in the analysis of the obtained data, descriptive and inferential statistics were carried out on the quantitative data. According to Pietersen & Maree (2007:183), “descriptive statistics is a collective name for a number of statistical methods that are used to organise and summarise data in a meaningful way. This serves to enhance the understanding of the properties of the data”. The mean, standard deviation, minimum test score and maximum test score were determined for both the pretest and the posttest data. For inferential statistics p-values were calculated to determine whether the performance and accuracy in performance evaluation were significantly different after instruction. p-values were also calculated to determine statistically significant difference in the performance and accuracy of judgement amongst the groups in the pre- and posttest.

3.10.2 Quantitative data (Questionnaire instrument)

Students’ responses on the Likert scales as well their biographical information were electronically captured. For each respondent and per construct (the instrument was used to measure more than one construct), values from 1 to 7 (seven categories were used, see paragraph 3.4.2.1) were assigned based on each respondent’s responses and then added to obtain a total score. Correlation coefficients and p values were computed to establish the relationship between mean scores of motivational factors, i.e. self-enhancement and self-protection and bias in performance evaluation and to also ascertain whether the relationship was significant.

3.10.3 Qualitative data (Part three of the three tier stoichiometry test instrument)

In the third tier of the stoichiometry test instrument students were asked to explain their choice of confidence judgement rating in a free-response format (see paragraph 3.4.1.2); this constituted qualitative data. The students’ responses were retyped. ATLAS.ti version 4.2 software package was used to systematically organise the students’ free responses for coding the qualitative data and for categorising codes into themes. Thematic analysis was used to analyse the students’ free responses. Although qualitative data are analysed differently from quantitative data, in using thematic analysis, a more deliberate and rigorous way of analysing

qualitative data is applied. According to Braun and Clarke (2006), during thematic analysis patterns or themes within data are identified, analysed and reported. A theme represents some level of patterned response or meaning within the data set (Braun & Clarke, 2006). Themes can be identified inductively or deductively. In the inductive approach identified themes are strongly linked to the data and data are coded without fitting them into a pre-existing coding frame. In the deductive approach identified themes are determined by theory or the researcher's analytical interest. The deductive approach however tends to provide a less rich description of the data compared to the inductive approach. According to Nieuwenhuis (2007), inductive analysis allows research findings to emerge from the frequent, dominant or significant themes inherent in raw data. The deductive approach wherein themes are formulated in advance, was not found suitable for the purpose of this study, as it tends to obscure or render key themes invisible (Nieuwenhuis, 2007).

Conducting thematic analysis entails following the following steps: familiarising oneself with the data; generating initial codes; searching for themes; reviewing themes; defining and naming themes; and producing the report. In analysing the data all these steps were carefully followed. Although thematic analysis is applauded for being a flexible method and easy to conduct on qualitative data especially for novice researchers, the following pitfalls need to be avoided when applying it: failure to analyse the data at all by collating together extracts with little or no analytical narrative; using the data collection questions as themes that are reported; generating an unconvincing analysis where there is too much overlapping between themes and the themes are not consistent; claims that cannot be supported by the data and lastly a mismatch between theory and analytic claims (Braun & Clarke, 2006). In this project all necessary precautions were taken to avoid all the pitfalls as highlighted in the literature.

3.11 ETHICAL CONSIDERATIONS

Before commencement of the study, ethical clearance was obtained from the University of Pretoria's Ethics Committee. Appendix VII is a letter granting ethical clearance by the University of Pretoria's Ethics Committee. None of the participants were minors and therefore the use of consent forms by parents were unnecessary. Participants were asked to sign a consent form signalling their willingness to participate in the study. Participants were promised complete anonymity and that the results of the study would not affect their grades in any way. Participants were duly informed of the objectives of the study before the administration of data collection instruments. The scripts, i.e. questionnaires and tests, were

handled by the researcher, her assistant, supervisors, statisticians and data capturer only. After marking and data capturing, the scripts were stored in a safe place and will be destroyed three years after the study. The findings of the study will be used to compile a report which will be submitted to my supervisors. The findings will possibly also be published in a science education journal and presented at a science, mathematics and technology education conference.

CHAPTER 4

RESULTS AND DISCUSSION

CONTENTS	PAGE
4.1 Introduction	66
4.2 Validity and reliability of data collection instruments	66
4.2.1 Stoichiometry test instrument	67
4.2.1.1 Content and Face validity	67
4.2.1.2 Reliability	70
4.2.2 Questionnaire instrument	71
4.2.2.1 Construct validity	71
4.2.2.2 Reliability	74
4.3 Data analysis	75
4.3.1 Performance in the pre- and posttest	75
4.3.2 Performance evaluation in the pre- and posttest	76
4.3.3 Interpretation of quantitative data	77
4.3.3.1 A detailed view and interpretation of quantitative data	82
4.3.3.1.1 Relationship between the pre- and posttest scores and average confidence scores	84
4.3.3.1.2 Comparison of the four performance evaluation subgroups in terms of average pre- and posttest scores, pre- and posttest pass rates and performance gain	85
4.3.3.1.3 Comparison of the performance evaluation subgroups in terms of prior chemistry knowledge	88
4.3.3.1.4 Comparison of the performance evaluation subgroups in terms of pre- and posttest performance	90
4.3.3.1.5 Comparison of the performance evaluation subgroups in terms of pre- and posttest average confidence scores	92

CONTENTS	PAGE
4.3.3.1.6 Comparison of the performance evaluation subgroups in terms of CMY 143 performance	95
4.3.3.1.7 Comparison of the performance evaluation subgroups in terms of performance gain	96
4.3.3.1.8 Summary	97
4.3.4 Questionnaire responses	98
4.3.4.1 The relationship between bias in performance evaluation and the self-enhancement motivational factor	99
4.3.4.2 The relationship between bias in performance evaluation and the self-protection motivational factor	101
4.3.4.3 The relationship between bias in performance evaluation and gender	103
4.3.4.4 Comparison of the four performance evaluation subgroups in terms of the three motivational factors (SEa, SEb, SP)	104
4.3.5 Qualitative data analysis	106
4.3.5.1 Qualitative data section of the three-tier instrument: Interrater reliability	110
4.3.5.2 Presentation and interpretation of qualitative data	113
4.3.5.3 Discussion of qualitative results	121
4.3.5.4 Conclusion	126

CHAPTER 4

RESULTS AND DISCUSSION

4.1 INTRODUCTION

The previous chapter focused on the research design and methodology. In this chapter results obtained through data collection instruments are presented and discussed. Firstly, validity and reliability of data collection instruments are discussed and next, raw test scores are reported; then quantitative analysis of stoichiometry test data and analysis of questionnaire responses are presented. This is followed by a presentation and discussion of the qualitative data on the subjects' free response explanations as obtained from the open-ended section of the chemistry pre- and posttest.

Of the 170 students who took the pre- and posttest, only fifty per cent, i.e. 85 scripts had complete chemistry and confidence entries. For some incomplete item responses, inferences could be made from the available data about the missing responses. Using criteria drawn up for inclusion of scripts with missing data explained in paragraph 3.9.4.1, nine of the 85 scripts with missing data could be included resulting in a sample of 94 students. However, of the 94 students, the records of three students had to be excluded. One of the students missed the second semester final chemistry examination and two only joined the programme in the second semester. Only those students who had joined the programme at the beginning of the year and had written the final second semester examination were included in the sample for two reasons: It was important that all subjects received the same instruction in the first semester; secondly, I also wanted to investigate the relationship between inaccuracy in performance evaluation and performance in the final second semester examination. Ultimately the sample consisted of 91 participants whose data were analysed and are discussed in this chapter.

4.2 VALIDITY AND RELIABILITY OF DATA COLLECTION INSTRUMENTS

Before using data collected by means of our data collection instruments we subjected these instruments to several measures to ensure content validity for the stoichiometry test, construct validity for the questionnaire as well as reliability for both instruments. Results obtained with regard to how the instruments performed in terms of validity and reliability, will now be presented and discussed, first for the stoichiometry test followed by the questionnaire instrument.

4.2.1 Stoichiometry test instrument

To enable the simultaneous collection of quantitative and qualitative data through a single instrument, the stoichiometry test instrument (Appendix I) was designed to have three tiers. In the first tier, students were given a multiple choice question on the topic of stoichiometry to solve. Immediately after choosing an answer, students were prompted in the second tier to indicate their level of confidence in their chosen response on a Likert scale from 0 to 100 per cent. Following that, in the third tier, students had to provide an explanation for their choice of confidence rating. Data collected by means of the first and second tiers represent quantitative data. The students' free responses indicated in the third tier constitute qualitative data. In the following paragraphs, I will discuss the results obtained after subjecting the test portion of the instrument to several measures to evaluate the validity as well as reliability. The discussion of the reliability and validity issues as relevant to the qualitative data will follow later in the chapter.

Several steps were followed to assess both the content and face validity of the stoichiometry test. The reliability of the test was determined by calculating a Cronbach's alpha coefficient. Detailed explanation of how these measures were interpreted follows in the paragraphs below.

4.2.1.1 Content and Face validity

Experts were approached and requested to assess the content validity of our stoichiometry test instrument. Three high school teachers and one university lecturer were each given a package consisting of a copy of the Stoichiometry test displayed in Appendix I, a questionnaire with four open-ended questions on the test instrument as shown in Appendix IV as well as a list of all the multiple choice questions together with what each question seeks to assess or measure shown in Appendix V. Literature was used to identify what some questions seek to measure as some of the questions were taken from research literature. The educators were requested in an instruction sheet, provided in Appendix III, to answer the questions in the test and record the time it took them to complete the test before proceeding to other documents in their packages. In the questionnaire educators were requested to identify any ambiguities or potential language barriers that may cause a second-language speaker to misunderstand. Educators had to say whether it was reasonable to expect a Grade 11 or 12 learner to answer the questions in the test. Lastly, educators had to comment on the overall presentation of the test, clarity of instructions, soundness of chemistry content, and suitability

of vocabulary and terminology used in the test. In the last document, the educators had to rate each question in terms of whether or not it measured what it intended to and make suggestions on how corrections could be made to the particular question (See Appendices III, IV and V).

Only one of the three high school teachers did not complete the validity check part of the educator questionnaire displayed in Appendix IV. The outcome of the validity check on the stoichiometry test performed by high school teachers and a university lecturer is reported in Table 4.1.

Table 4.1: Results of content validity conducted by the educators on the stoichiometry test

	Teacher 1 (T1)	Teacher 2 (T2)	Teacher 3 (T3)	Lecturer (L)	Total
Question 1	1	1		1	3
Question 2	0	0		0	0
Question 3	1	1		1	3
Question 4	0	1		1	2
Question 5	1	1		0	2
Question 6	1	1		1	3
Question 7	1	1		1	3
Question 8	0	1		1	2
Question 9	0	1		1	2
Question 10	1	1		1	3
Question 11	1	1		0	2
Question 12	1	1		1	3
Question 13	1	1		1	3
Question 14	1	1		0	2
Question 15	1	1		1	3
Question 16	1	1		1	3
Question 17	1	1		1	3
Question 18	1	1		1	3
Question 19	1	0		0	1
Question 20	0	0		0	0
TOTAL	15	17		14	
PERCENTAGE	75	85		70	

Key (Does question measure what it intends to?):

YES	1
NO	0
No comment	

All the educators unanimously agreed that items 2 and 20 were ambiguous but did not provide any recommendations for improvement. The following recommendations regarding the test items were made by the educators:

- i. change the sequence of questions 5 and 4 (T1);
- ii. remove decimal points after each value given in the distractors of question 8 (T2);
- iii. question 19 does not seek to measure symbolic representation but rather it intends to measure the student's understanding of how a reaction occurs and the role of the limiting reactant in a chemical reaction (L).

The sequence of questions was changed from questions four, five and six to six, four and five based on the educators' recommendation. Question five requires a student to state how balancing a chemical equation is achieved; question six requires a student to balance a chemical reaction equation, whereas question four requires the use of a balanced equation to determine moles of reactant from moles of product. To reduce the number of variables resulting in the ambiguity of item 20, molar mass of calcium carbonate was given; option E in question 8 was changed from 1350g to 1354g and decimal points appearing after each distractor value were removed. Decimal points were also removed from all distractor values in question 8. All the above mentioned changes were applied and this resulted in the stoichiometry test instrument as it appears in Appendix I. It is in this format that the test was used to collect data in the main study. Item 2 was retained in order to gather statistical evidence before taking a decision about its possible removal.

In addition the four students who participated in the pilot study (see paragraph 3.7.3) unanimously agreed that there were no ambiguities and potential language barriers for second-language speakers, instructions are clear, and vocabulary and terminology used in the test are appropriate for the level of students for which the test is intended. Thus, information obtained from these students confirmed the accessibility of the instrument in terms of the appropriateness of language used, clarity of instructions as well as the level of difficulty of the test items, bearing in mind the sample for which the test was intended.

4.2.1.2 Reliability

After administration of the test instrument as both pre- and posttest in the main study, Cronbach's coefficient alphas of 0.63 and 0.67 were found, respectively. Item-total correlations were calculated for each item. Item 2 was found to have a very weak item-total correlation in the pretest and a negative item-total correlation in the posttest which

corroborated with the educators' feedback in the previous paragraph. With item 2 excluded, the Cronbach's alpha of the pretest improved marginally to 0.64 and that of the posttest to 0.69. In the light of what the experts and the statistical analysis alluded to, it was decided to omit item 2 from the analysis of test data. Thus, even though all 20 items were answered in the test, only 19 items were analysed.

The reliability of the instrument was therefore found to be somewhat low with marginal improvement in the posttest. According to Nunnally (1967) a low Cronbach's alpha may be an indication that some items still need improvement. However it is also possible that the small sample, limited number of items and assessment of a diverse subject could contribute to an instrument having a low Cronbach's alpha. All these factors could be revisited and addressed in future studies.

4.2.2 Questionnaire instrument

I will subsequently discuss the outcomes with regard to construct validity and reliability of the questionnaire instrument.

4.2.2.1 Construct validity

Factor analysis indicated that items in the questionnaire were measuring three constructs rather than two, i.e. two forms of self-enhancement and one of self-protection. A correlation of 0.25 or greater between two factors is an indication that the two factors are measuring the same construct. Table 4.2 shows that the correlations between factors one and two, factors one and three, and factors two and three were 0.21, 0.19 and 0.17, respectively, and these are smaller than 0.25. Thus the poor correlation between factors was further confirmation that we were indeed dealing with three discrete factors. In this discussion and others factor one will be labelled Self-Enhancement a (SEa), factor two Self-Enhancement b (SEb), and factor three Self-Protection (SP).

Table 4.2 Inter-correlations of the three motivational factors

VARIABLE	FACTOR1	FACTOR2	FACTOR3
FACTOR1	1.000		
FACTOR2	0.206	1.000	
FACTOR3	0.192	0.165	1.000

A table of factor loadings (Table 4.3) shows the strength with which items correlated with one another, resulting in three separate factors measuring different constructs. Bearing in mind that a correlation coefficient greater than 0.25 between two items indicates that the two items measure the same construct, Table 4.3 shows that items labelled SE11, SE15, SE20, SE23, SE26 and SE29 had a correlation coefficient greater than 0.25 loading onto factor one, i.e. SEa factor. Items labelled SE1, SE6, SE8, SE11 and SE18 also had a correlation coefficient greater than 0.25, loading onto factor two which is the SEb factor and items labelled SP5, SP7, SP12, SP13, SP16, SP21, SP24, and SP27 loaded strongly onto factor three, i.e. SP factor. SE11 showed a correlation greater than 0.25 with both the SEa and SEb factors (SEa: 0.359 and SEb: 0.403), but comparing the two showed that in fact SE11 was more strongly correlated with items in SEb than SEa.

Table 4.3: Factor loading matrix showing a pattern of how items loaded onto discrete factors

ITEMS		FACTOR1	FACTOR2	FACTOR3
SE1	1	-0.087	0.689	-0.022
SP2	2	-0.044	0.087	0.195
SP5	3	-0.101	0.132	0.254
SE6	4	0.011	0.808	0.029
SP7	5	0.027	0.013	0.418
SE8	6	0.248	0.631	-0.006
SE11	7	0.359	0.403	-0.118
SP12	8	0.224	0.036	0.387
SP13	9	0.196	-0.092	0.417
SE15	10	0.557	0.027	0.051
SP16	11	-0.109	0.084	0.433
SE18	12	0.244	0.355	0.048
SE20	13	0.370	0.086	0.023
SP21	14	0.089	-0.055	0.489
SE23	15	0.799	-0.084	-0.147
SP24	16	0.083	-0.103	0.364
SE26	17	0.597	0.047	0.205
SP27	18	-0.016	-0.094	0.390
SE29	19	0.334	0.131	0.142

In general items in SEa express the desire to portray oneself to others as a hardworking, clever student. Items in SEb are concerned with the need to give a good impression of oneself to others while items in SP are concerned with the need to protect one's academic self-worth by attributing failure to external factors. The nineteen statements included in the questionnaire are shown in Table 4.4, grouped according to their loading onto the three factors as indicated by factor analysis.

Table 4.4: Questionnaire items as grouped together by factor analysis

SEa Factor One	<p>Factors concerned with the desire to portray oneself to others as a hardworking, clever student.</p> <p>(15) I like to present myself to others as being a clever person.</p> <p>(20) I set difficult goals for myself so people can see I am serious about my work.</p> <p>(23) It is important that others see me as being the best in my class.</p> <p>(26) It is important to me that others think I work hard.</p> <p>(29) I work harder than I normally do when I know someone is watching me.</p>
SEb Factor Two	<p>Factors concerned with the need to give a good impression about oneself to others.</p> <p>(1) I intend to change my behaviours to create a good impression to others.</p> <p>(6) I try to modify my behaviours to give good images to others.</p> <p>(8) It is important to me to give a good impression to others.</p> <p>(11) I try to create the impression that I am a “good” student.</p> <p>(18) I am sensitive to the impression about me that others have.</p>
SP Factor Three	<p>Factors concerned with the need to protect one’s academic self-worth by attributing failure to external factors.</p> <p>(2) When I perform poorly in a test there are usually external circumstances that are to blame.</p> <p>(5) I do not set goals that are hard to reach, because failure is painful.</p> <p>(13) Often I get poor results in courses because the teacher has failed to make them interesting.</p> <p>(16) It is unrealistic to expect good grades for maths and science because these are hard subjects.</p> <p>(21) Sometimes my success on examinations depends on luck.</p> <p>(24) My performance does not reflect my ability: I was just unlucky not to be taught by a better teacher.</p> <p>(27) It is better to expect poor results and to be surprised than to be disappointed when your expectations are not met.</p> <p>(12) I do not want my friends to know about it when I have failed a test.</p> <p>(7) When I get poor results in a test I just want to get rid of the script and not look at it again.</p>

NB: Items are numbered as they appear in the final questionnaire instrument shown in Appendix VI.

4.2.2.2 Reliability

According to Nunnally (1967) a Cronbach's alpha coefficient between 0.70 and 0.80 is an indication of good reliability. A Cronbach's alpha of 0.74 was obtained for the questionnaire instrument which indicates that it was a reliable instrument for collecting data in this study with this particular sample. The results obtained from statistical analysis were further analysed to verify whether all items, including some poor-performing items, should be retained in the final version of the questionnaire instrument. Internal reliability was calculated within each subset of items representing the three different factors. In particular three items correlated poorly with items within their subgroups, i.e. item 29($\alpha = 0.29$), item 18($\alpha = 0.37$) and item 2($\alpha = 0.19$). A Cronbach's alpha of 0.66 was calculated for all five SEa items. However when item 29 was excluded from the subset of items the Cronbach's alpha increased to 0.67. A Cronbach's alpha of 0.74 was calculated for all five SEb items. Upon exclusion of item 18 the Cronbach's alpha for this subset of items remained the same. A Cronbach's alpha of 0.60 was computed for all nine SP items. When Item 2 was excluded from the subset of items, the Cronbach's alpha also remained the same. Table 4.5 shows that exclusion of an item with the weakest correlation with other items within the questionnaire did not result in any meaningful improvement in terms of reliability of the instrument.

Table 4.5: Internal reliability: Cronbach's alpha coefficients of the instrument upon the removal of individual items

Items	Cronbach's ALPHA
1. SE1	0.7328
2. SP2	0.7439
3. SP5	0.7431
4. SE6	0.7216
5. SP7	0.7344
6. SE8	0.7175
7. SE11	0.7254
8. SP12	0.7227
9. SP13	0.7293
10. SE15	0.7239
11. SP16	0.7396
12. SE18	0.7262
13. SE20	0.7320
14. SP21	0.7319
15. SE23	0.7306
16. SP24	0.7394
17. SE26	0.7136
18. SP27	0.7432
19. SE29	0.7259

ALPHA FOR ALL VARIABLES = 0.7411

To summarise, piloting, statistical measures applied to and refining of both data collection instruments provided sufficient evidence that the conclusions based on the results obtained through the instruments would likely be valid and of an acceptable reliability. I will subsequently discuss how data collected using these instruments were used to answer the research questions of the study.

4.3 DATA ANALYSIS

Quantitative and qualitative data were collected and analysed in an attempt to answer the following research questions:

Research question 1: How accurately do BFYP students evaluate their performance in a stoichiometry test?

Research question 2: What is the influence of teaching of stoichiometry in the BSc Four-year programme on performance and accuracy of performance evaluation?

Research question 3: What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?

Research question 4: What is the relationship between bias in performance evaluation and self-enhancement, self-protection and gender?

Quantitative data were used to answer research questions 1, 2 and 4. Qualitative data were used to answer research question 3. In the paragraphs that follow raw score data collected by means of the stoichiometry test instrument will be presented first. This will then be followed by a discussion of how the quantitative data were analysed and interpreted to assist in answering research questions 1 and 2. Quantitative data obtained by means of the questionnaire instrument will be presented and followed by a discussion of how the data were used to answer research question 4. The chapter will conclude with the presentation, interpretation and discussion of how qualitative data helped us answer research question 3.

4.3.1 Performance in the pre- and posttest

The sum of all correctly answered items in a test out of 19 was used as an indication of performance. A correct answer was scored 1 and an incorrect answer 0 (zero). Scores obtained in the pretest and posttest were used to generate the descriptive statistics presented in Table 4.6.

Table 4.6: Descriptive statistics of students' performance in the pre- and posttests

	Pretest	Posttest
Sample size	91	91
Mean	7.0	9.6
STD deviation	2.9	3.4
Minimum	2 (Max = 19)	3 (Max = 19)
Maximum	15 (Max = 19)	18 (Max = 19)

The mean values in Table 4.6 represent the mean of test scores out of a maximum score of 19 obtained by the students. On raw scores it is clear that students improved in performance. Out of a total score of nineteen, a maximum score of eighteen was obtained in the posttest whereas the highest score obtained in the pretest was fifteen. The minimum score obtained increased from two in the pretest to three in the posttest. A matched t-test run on the data yielded a p-value less than 0.05 ($p = 0.00$) indicating a statistically significant difference between the pre- and posttest performance.

4.3.2 Performance evaluation in the pre- and posttest

To evaluate their performance students were requested to choose, on a scale of 0% to 100%, a rating that best described the confidence they had in the accuracy of each of their chosen responses. An average value of confidence scores was calculated per individual student. The average confidence scores obtained in the pretest and posttest were used to generate the descriptive statistics presented in Table 4.7.

Table 4.7: Descriptive statistics of average confidence scores in the pre- and posttests

	Pretest	Posttest
Sample size	91	91
Mean	63.0	75.7
STD deviation	17.4	13.5
Minimum	16.3	40.5
Maximum	94.7	99.5

The mean score results presented in Table 4.7 are percentages. These results show that average confidence scores increased from 63.0% in the pretest to 75.7% in the posttest. A matched t-test run on the data yielded a p-value less than 0.05 ($p = 0.00$) indicating a statistically significant difference between the pre- and posttest average confidence scores.

To summarise, the results presented in Tables 4.6 and 4.7 indicate that both performance and confidence scores increased by a significant margin in the posttest. The mean performance increased by 14% and the mean confidence by a similar margin (13%).

4.3.3 Interpretation of quantitative data

The information in paragraphs 4.3.1 and 4.3.2 gives an indication of the average performance and confidence scores of the students during the pre- and posttest. It is clear from the discussion above that there was a statistically significant difference between the pre- and posttest performance as well as between the pre- and posttest average confidence levels. However these results mask the finer details of whether individual students were able to accurately evaluate their performance. The results also do not provide us with an indication of whether individual students' ability to evaluate their performance had improved or deteriorated after instruction. An explanation of how accuracy of performance evaluation was defined for this study was provided in paragraph 3.10.1. In short, the total score out of a maximum of 19 obtained in the test was converted to a percentage value and then subtracted from the average confidence score also in the form of a percentage value. Any difference higher than 15.8% was interpreted as an indication of overconfidence and any difference between 15.8% and -15.8% (15.8% and -15.8% included) was taken as an indication of accurate performance evaluation. Finally, any difference more than -15.8% was an indication of underconfidence. This chosen margin of error assisted us in the manipulation of data and in the identification of students who were accurate in the evaluation of their performance and those who were not. Students were categorised based on their accuracy of performance evaluation using this margin of error. More information on how we arrived at the 15.8% value as our allowed margin of error was presented in paragraph 3.10.1. The abovementioned margin of error enabled us to determine the level of accuracy students showed during the evaluation of their performance, assisting our attempt to answer research questions 1 and 2, i.e. how accurately the students evaluated their performance in the test and the effect that teaching had on the students' performance as well the accuracy with which the students evaluated their performance.

Ideally a student should know when he or she has not mastered the content required for a specific question or whether he or she is in command of the procedures necessary to solve a problem. A student should be able to recognise when he/she does not understand or know the answer to a question. Therefore when he/she knows that he/she is unlikely to provide the correct answer to a question, he/she should choose a low confidence rating and when he/she is likely to provide the correct answer, he/she should choose a high confidence rating. Tables 4.8 and 4.9 depict how students were categorised as either overconfident, realistic or underconfident based on the accuracy with which they evaluated their performance in the pre- and posttest.

Table 4.8: Student categories based on the evaluation of their performance in the pretest

Category	No. of students	Males	Females
Overconfident	63 (69%)	25	37
Realistic	28 (31%)	10	18
Underconfident	0	0	0
Total	91	35*	55*

*One record without gender information omitted.

Table 4.9: Student categories based on the evaluation of their performance in the posttest

Category	No. of students	Males	Females
Overconfident	65 (71%)	25	39
Realistic	24 (26%)	9	15
Underconfident	2 (2%)	1	1
Total	91	35*	55*

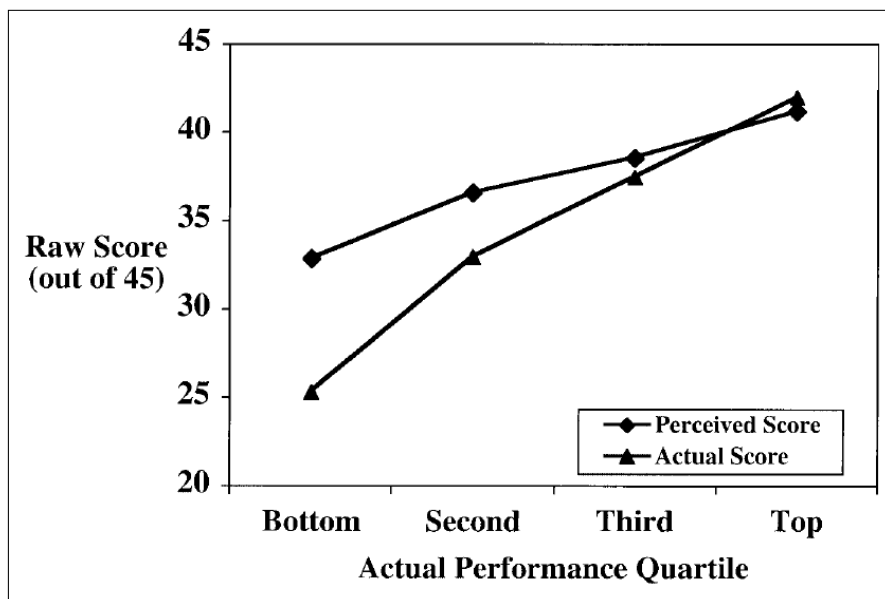
*One record without gender information omitted.

The number of students who were overconfident in their judgement increased marginally, i.e. from 69% in the pretest to 71% in the posttest (Table 4.9). Students who were realistic in their judgement decreased from 31% in the pretest to 26% in the posttest. The data in Tables 4.8 and 4.9 help us to answer the first research question (*How accurately do the students evaluate their performance?*). The answer to research question 1 is that only a minority of

students were able to evaluate their performance accurately, i.e. only 31% in the pretest and 26% in the posttest.

To further explore research questions 1 and 2, we adopted the method of Dunning, Johnson, Ehrlinger and Kruger (2003) of presenting results per performance quartile. According to their method students are divided into four groups based on their actual performance in a test from the bottom 25% of performers to the top 25% resulting in four quartile ranks. Dunning *et al.* (2003) asked participants to estimate their score after writing a psychology examination. The data obtained were used to plot estimated performance against the participants' actual performance. They used this method to study how the estimated scores of participants in the quartiles compared with actual scores. Figure 4.1 is a graphical representation of their findings.

Figure 4.1: Actual versus perceived scores of students in their quartile ranks (adapted from Dunning *et al.*, 2003)



Students in the bottom quartile greatly exaggerated their performance. Whereas their estimated score placed them at about 33 out of 45, they actually obtained an average score of 25. The extent of overestimation decreased from the second to the third quartile. On the other hand, the top quartile which consists of top performers tended to marginally underestimate their performance.

In our case average confidence scores were interpreted as students' perceived scores. For example a student that was 80% certain that the answer was correct, assumed that there was an 80% likelihood of the answer actually being correct. The average of all of the confidence ratings that were chosen by a student would therefore provide a good indication of expected or perceived performance in the test. Students were divided into four quartile ranks based on their actual performance. Mean values for the confidence ratings and test scores of students in each quartile were calculated and a graph of expected against actual performance plotted. Figure 4.2 represents the graph of pretest results and Figure 4.3 of posttest results. If we compare the graphs of our results with that of Dunning *et al.* (2003), similarities and differences are observed. Figures 4.2 and 4.3 show that even for students in the top quartile, a mismatch existed between the expected and actual performance. This was observed in both the pre- and posttest.

Figure 4.2: Expected versus actual pretest performance of students in the four performance quartiles

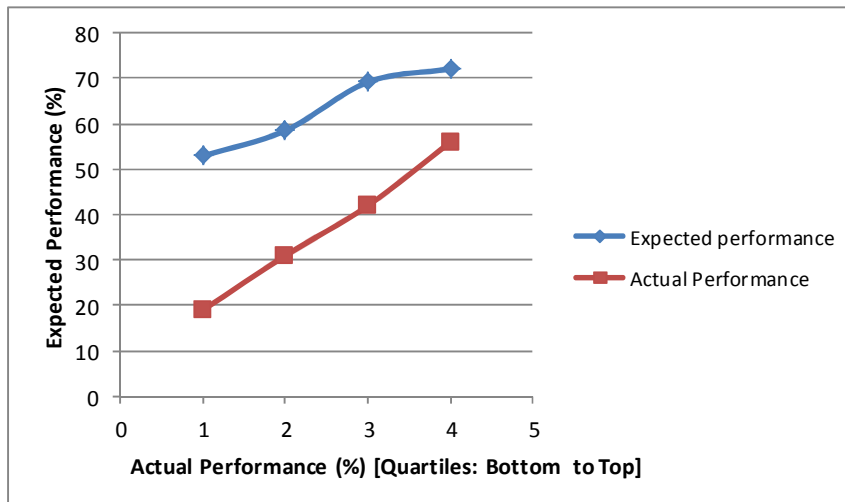
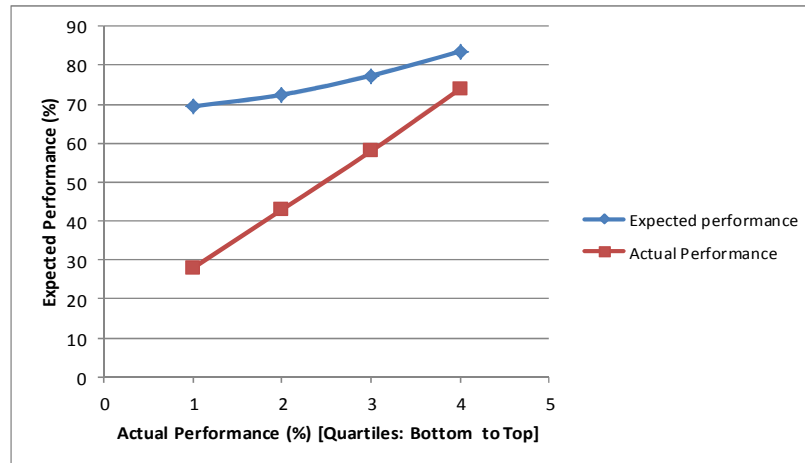


Figure 4.3: Expected versus actual posttest performance of students in the four performance quartiles



Our results corroborate with those of Dunning *et al.* (2003) in that the students in the bottom quartile tended to miscalibrate themselves by the biggest margin. One would have expected the students to calibrate their performance better after instruction, but even the top performers were only marginally better in the evaluation of their performance (Figure 4.3). The only noticeable improvement is that after instruction the gap between the actual and expected performance has been narrowed for students in the third and top quartiles. However, after instruction there was a bigger mismatch between the expected and actual average scores of students in the bottom quartile. In the results obtained by Dunning *et al.* (2003) for psychology students shown in Figure 4.1, the overestimation of performance turned into an underestimation of performance for the students in the top quartile. However, in our case this did not happen as shown in Figures 4.2 and 4.3. The fact that students in the top performing quartile of the posttest still underestimated their performance by an average margin of 9% substantiates the necessity for applying a generous margin of error (15.8%) for this cohort of students for the purpose of classification. Without a wide margin of error all the students in our sample would have been labelled overconfident and discrimination in terms of accuracy of performance evaluation would not have been observed. Stoichiometry is clearly a difficult topic and accuracy of calibration is a difficult skill to master for under-prepared students in the BSc Four-Year Programme.

To answer the question of whether or not teaching had an influence on performance and accuracy of performance evaluation (research question 2), based on the population of the different performance evaluation groups in Tables 4.6, 4.8 and 4.9, it seems that there was

overall improvement in performance and no overall improvement in accuracy of performance evaluation. However a global view of the results as presented in Tables 4.8 and 4.9 makes it difficult to uncover any improvement in accuracy of performance evaluation that may exist for subgroups of students. The global view of results as shown in Tables 4.8 and 4.9 does not show whether it is the same students who remain inaccurate in their evaluation, or what proportion of students showed either improvement or deterioration of accuracy in performance evaluation after teaching. Seeing that inferences could not be made from just a general comparison of the performance evaluation in the pre- and posttest, data had to be analysed and looked at more closely.

4.3.3.1 A detailed view and interpretation of quantitative data

The discussion and interpretation from now on will be an attempt to unmask the results and explore fine details. In order to investigate the effect of teaching on accuracy in performance evaluation students were observed individually, in terms of how they shifted in their performance evaluation groups after teaching and learning had taken place. Table 4.10 is a two-way frequency table which was constructed to depict how the accuracy with which students evaluated their performance, changed or stayed the same after instruction.

Table 4.10: Shifts in student groups after teaching and learning

	Post OC	Post R	Post UC	TOTAL
Pre OC	50	13	0	63
Pre R	15	11	2	28
Pre UC	0	0	0	0
TOTAL	65	24	2	91

Key	
OC	Overconfident
R	Realist
UC	Underconfident

In the first column the pretest groups are listed and the posttest groups are listed in the top row based on categories of accuracy of performance evaluation. Of the 63 students who were overconfident in the pretest, a large percentage showed no improvement in the accuracy of their performance evaluation, i.e. 50 remained overconfident and only 13 were able to make

accurate judgements in the posttest. In the group that was realistic in the pretest (28 students) less than half remained realistic in their judgment in the posttest (11 students). More than half miscalibrated and became overconfident (15 students). The number of students that showed an improvement in the accuracy of their judgements by moving from the overconfident group in the pretest to the realistic group in the posttest (13) matches that of students who slipped from the realistic group to become overconfident (15). Only two of the 28 who were initially realistic became underconfident in the posttest. There were no underconfident students in the pretest.

Five subgroups emerge from the results presented in Table 4.10. The five subgroups are labelled with a set of codes consisting of the category of the pretest followed by that of the posttest, abbreviated as follows: OC-OC, OC-R, R-R, R-OC and R-UC. The subgroup OC-OC consists of all students who were overconfident in the pretest and remained overconfident in their judgement after instruction, i.e. the Overconfident-Overconfident group. The OC-R subgroup is made up of students who were overconfident in their judgement in the pretest, but improved and reported realistic judgements in the posttest. The R-R subgroup is for those students who remained realistic in their judgement and the R-OC subgroup comprises students who were initially realistic in their judgement but became overconfident in the posttest. The last subgroup, R-UC is for two students, one male and one female, who were realistic in their evaluation but became underconfident in the posttest. The female student obtained a score of 10 out of 19 (53%) and 14 out of 19 (74%) in the pre- and posttest, respectively. The male student also showed improved performance by obtaining a score of 6 out of 19 (32%) in the pretest and 15 out of 19 (79%) in the posttest. However since reliable inferences cannot be made from a small sample, the data records of students from this subgroup are subsequently omitted in my discussion.

As an attempt to answer the research question 2 which is about the influence of teaching on the ability of students to make accurate evaluations of their test performance, from this stage onwards, the data will be studied further and analysed in these independent four subgroups ($N = 89$) dubbed *pre-post performance evaluation subgroups*. In my discussion I will compare students in the four performance evaluation subgroups in terms of the following characteristics: the relationship between test scores and average confidence scores for the pre- and posttests, the average scores in the pre- and posttests, the percentage of students who

passed the pre- and posttests, and finally the average performance gain achieved in the posttest as compared with the pretest.

4.3.3.1.1 Relationship between the pre- and posttest scores and average confidence scores

Figures 4.4 and 4.5 are scatterplots of the average confidence score versus test score per student in the pre- and posttests respectively. The students in each performance evaluation subgroup are indicated by a different symbol. Ideally there should be a good match between the average confidence scores and the actual performance. However the plots in Figures 4.1 and 4.2 show that for this particular cohort of students this was not the case.

Figure 4.4: Scatterplot of pretest average confidence scores against pretest scores: categorized by pre-post performance evaluation subgroups

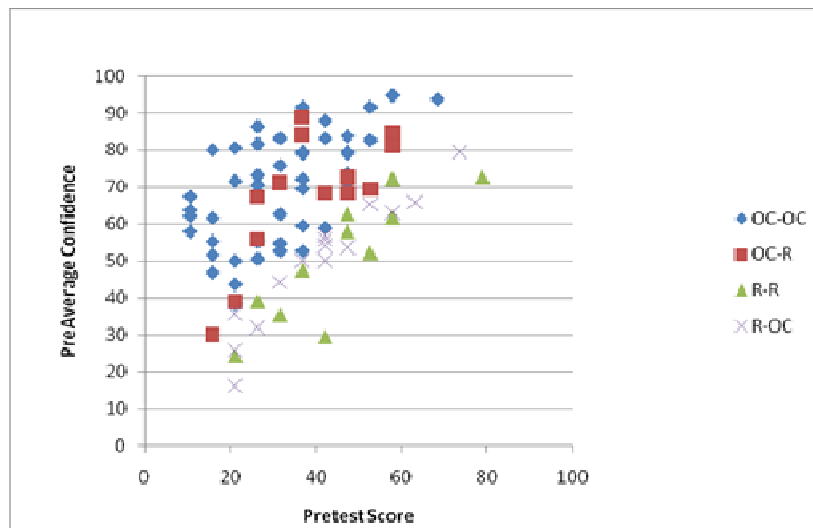
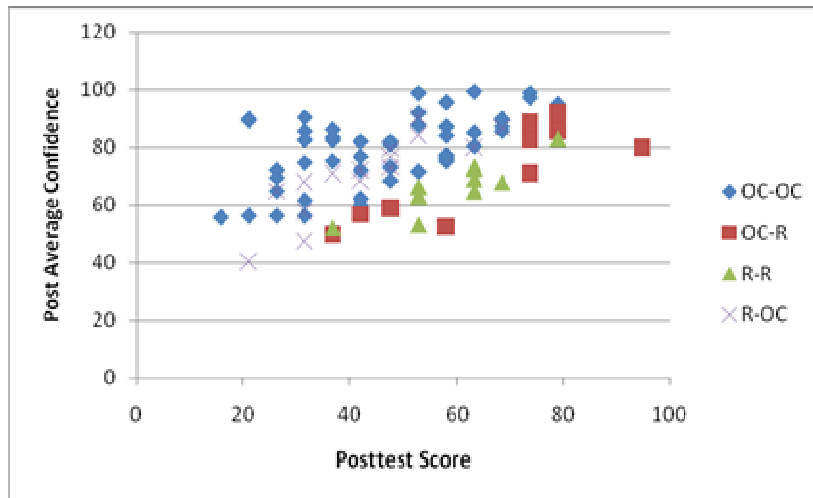


Figure 4.5: Scatterplot of posttest average confidence scores against posttest scores: categorized by pre-post performance evaluation subgroups



Figures 4.4 and 4.5 show that for the majority of students regardless of which category they belong to, the high confidence levels are not justified by high test scores in the pretest. In fact the relationship between the average of confidence judgement ratings and test performance was found to be positive but weak for both the pre- and posttests. The average of confidence judgement ratings was found to be significantly related to the students' pretest scores, $r = .44$, $p = 0.00$ and average of confidence ratings reported during the posttest was also found to be significantly related to posttest scores, $r = .42$, $p = 0.00$.

4.3.3.1.2 Comparison of the four performance evaluation subgroups in terms of average pre- and posttest scores, pre- and posttest pass rates and performance gain

Table 4.11 is a summary of how the four subgroups compare in terms of pre- and posttest mean scores, the number of students who passed the pre- and posttests, respectively, and the average performance gain achieved. Table 4.11 is divided into four quadrants. Each quadrant represents a performance evaluation subgroup. The top left quadrant represents the data of students in the OC-OC subgroup, with the top right representing the OC-R subgroup, the bottom left representing the R-OC group and the bottom right, representing the R-R subgroup. The first row in each quadrant states the number of students in that subgroup. In the second and third row of each quadrant the average pre- and posttest performance of each subgroup is shown respectively. In the fourth and fifth rows of each quadrant the percentage of students who passed the pre- and posttests is shown, and the performance gain of each subgroup is shown in the last rows.

Table 4.11: Pre- and posttest performance data according to performance evaluation subgroups

	POST OC		POST R	
PRE OC	Number	50	Number	13
	Av. Pretest performance (%)	33	Av. Pretest performance (%)	38
	Av. Posttest performance (%)	45	Av. Posttest performance (%)	68
	% Pass Pretest	10	% Pass Pretest	23
	% Pass Posttest	40	% Pass Posttest	77
	Av. Performance Gain (%)	19	Av. Performance Gain (%)	49
PRE R	Number	15	Number	11
	Av. Pretest performance (%)	41	Av. Pretest performance (%)	45
	Av. Posttest performance (%)	43	Av. Posttest performance (%)	61
	% Pass Pretest	27	% Pass Pretest	36
	% Pass Posttest	27	% Pass Posttest	91
	Av. Performance Gain (%)	-1	Av. Performance Gain (%)	25

Pretest performance ranges from 33% to 45%. The performance in the posttest was significantly higher, ranging from 43% to 68%. The OC-R subgroup obtained the highest average score in terms of their performance in the posttest. Less than 50% of the students in each subgroup managed to pass the pretest with the OC-OC subgroup obtaining the lowest pass rate. Performance averages are rather low (less than 50%) for all the subgroups in the pretest. In the posttest the R-R group managed to achieve an almost 100% pass rate. The pass rates achieved by the OC-R and R-R subgroups respectively increased by more than 50% from pre- to posttest. The percentage of students who passed the pretest is equal to the percentage that passed the posttest in the R-OC subgroup which correlates with an insignificant increase in average performance from the pretest (41%) to the posttest (43%). According to Table 4.11 the best performance and the most meaningful improvement in terms of pass rate were demonstrated by the OC-R and R-R subgroups. These two subgroups obtained more than 50% performance averages in the posttest and their pass rates increased by more than 50% as compared with the pretest.

The performance gain is a variable which provides a measure of the extent of improvement in performance in the four subgroups. However, in principle the information that can be

obtained from this variable is limited by the fact that different subgroups may differ in terms of their preknowledge and therefore their room for improvement. A student who scores 16 out of 19 in the pretest can only improve by three points to obtain a full score as opposed to a student who initially scored 6 in the pretest and can gain another 13 points. As a result, rather than comparing subgroups of students in terms of actual performance gain, performance gain results were normalised against scope for improvement. In his comparison of pre- and posttest performance of physics students, Hake (1998) defined normalised gain as the ratio of the actual average gain ($\% \langle \text{post} \rangle - \% \langle \text{pre} \rangle$) to the maximum possible average gain ($100 - \% \langle \text{pre} \rangle$).

$$\begin{aligned} \langle g \rangle &= \% \langle G \rangle / \% \langle G \rangle_{\max} \\ &= (\% \langle S_f \rangle - \% \langle S_i \rangle) / (100 - \% \langle S_i \rangle) \end{aligned}$$

where $\langle S_f \rangle$ and $\langle S_i \rangle$ are the final (post) and initial (pre) class averages

In our study this ratio was used to calculate performance gain for each of the four subgroups. Normalising performance gain against room for improvement in our case yields the following equation:

$$\text{performance gain (\%)} = [(\text{postscore} - \text{prescore}) / (19 - \text{prescore})] * 100$$

In the equation above the difference between the pre- and posttest scores represents the actual gain while 19 minus the prescore represents the maximum possible gain or room for improvement. The performance gain was calculated for each student and the average shown in Table 4.11 subsequently calculated for each subgroup. The OC-OC group showed almost as much learning gain as the R-R subgroup but they started from such a low base that the improvement was not enough to ensure that the majority would pass as was the case for the R-R subgroup. The learning gain of the R-R subgroup was sufficient to ensure that nearly 100% passed the posttest because their pretest performance was higher (45%). Even though they started off low in terms of their pretest performance (38%), the OC-R group was able to achieve the highest learning gain (49%), which was also enough to ensure an increase in the pass rates from 23% in the pretest to 77% in the posttest percentage of students who passed the posttest. The inability of the R-OC to achieve any learning gain resulted in no improvement as far as the pass rate was concerned.

From the pretest performance results it seems that students in all of the subgroups were relatively weak. The pretest performance was between 33% and 45%. However whether the difference between the subgroups in terms of pretest performance is statistically significant needs to be determined. There is a much wider range in the posttest performance of the subgroups, i.e. from 45% to 68%. The percentage pass in the pretest ranged from 10% to 36% and the percentage pass in the posttest from 27% to 91%. All these results required further analysis to determine whether the differences evident in Table 4.11 are statistically significant. The results of statistical analysis will be presented in the next paragraphs.

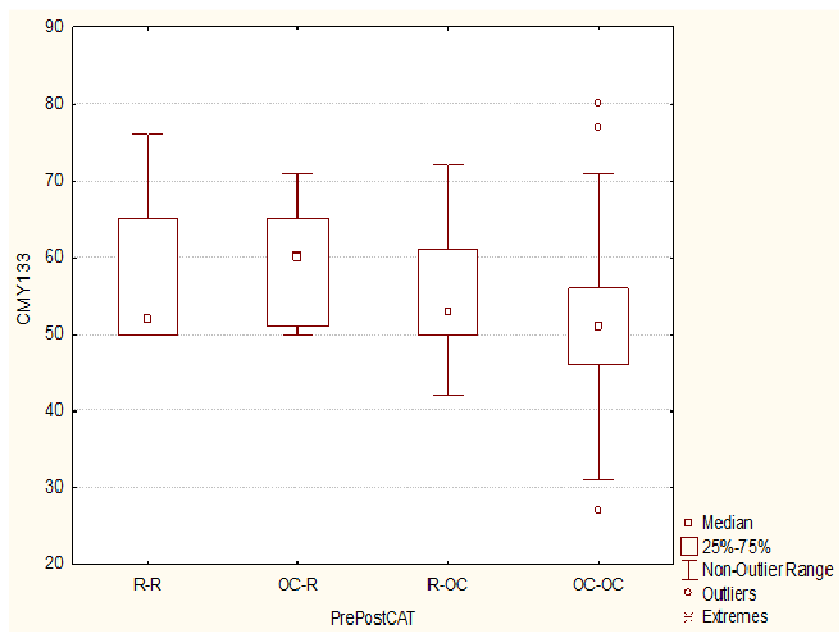
To investigate whether there was a significant difference in the four pre-post performance evaluation subgroups in terms of their pre- and posttest performance, average confidence, first semester chemistry performance as well as end-of-year chemistry performance, Kruskal-Wallis tests were run on the data of the four subgroups and boxplots were drawn. A Kruskal-Wallis test was used because the data set is skewed and the size of the subgroups is small. The Kruskal-Wallis test is a statistical measure used to compare medians of subgroups in a sample. Analysis results are presented in the form of a table with p-values which are generated after comparison. Results are also presented as a box plot showing the data distributions and the medians of the individual subgroups. An overall p-value for the entire data set smaller than 0.05 is an indication of some significant difference hidden in the data set without specifying where the difference lies. However p-values generated after comparing individual groups within a data set reveal where the difference actually lies, and the results of such an analysis are presented in table form. The results of such analyses for the four subgroups in our study will be presented and discussed in the following paragraphs.

4.3.3.1.3 Comparison of the performance evaluation subgroups in terms of prior chemistry knowledge

This study was conducted during the second semester when the students were taking their second semester chemistry module, namely CMY 143. We used the performance results of the students' first semester chemistry module, CMY 133, as an indication of the prior chemistry knowledge that the students possessed upon entry to the second semester. Statistical analysis of CMY 133 results was conducted and the results are presented as a boxplot in Figure 4.6 and in Table 4.12. With regard to how the different groups compare in their first semester chemistry module performance, the box plot in Figure 4.6 shows that there

is a large extent of overlap between the performance distributions of the subgroups, with the OC-OC subgroup being marginally weaker.

Figure 4.6 Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of their performance in the first semester chemistry module (CMY 133)



The overall comparability of the subgroups as shown in Figure 4.6 was confirmed by the Kruskal-Wallis test results in Table 4.12, which also shows a significant but marginal overall difference between the subgroups, $p = .0435$. The table reports no significant difference between any pairs of subgroups except between the OC-R and OC-OC subgroups where a p -value of 0.055 was calculated. The difference in the CMY 133 performance of these two groups is therefore only marginally significant with the OC-R group being better prepared than the OC-OC group upon entry to CMY 143. The medians of all the subgroups range from 50% to 60%. Students seem to have entered the second semester with a similar level of competence with OC-OC being slightly weaker.

Table 4.12 Multiple comparisons (p-values) of CMY 133 performance of students in the four pre-post performance evaluation subgroups

Kruskal-Wallis test: $H(3, N=89) = 8.124621$ $p = .0435$				
	R-R	OC-R	R-OC	OC-OC
R-R		1.000000	1.000000	0.741126
OC-R	1.000000		1.000000	0.055093
R-OC	1.000000	1.000000		1.000000
OC-OC	0.741126	0.055093	1.000000	

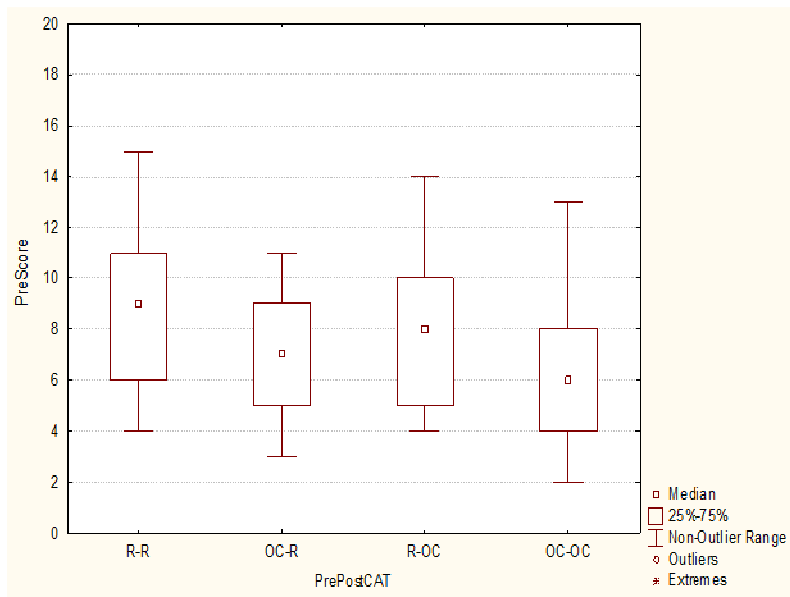
4.3.3.1.4 Comparison of the performance evaluation subgroups in terms of pre- and posttest performance

Table 4.11 showed that average pretest scores of students in the four subgroups ranged from 33% to 45% in the pretest and from 45% to 68% in the posttest. A Kruskal-Wallis test was used to determine whether the differences in the average performance of the subgroups in both the pre- and the posttests were significant.

i. Comparison of pretest scores

Statistical analysis of the pretest scores of the subgroups was conducted and the results are presented as a boxplot in Figure 4.7. This box plot shows extensive overlap between the performance distributions of the four subgroups with the possibility that the OC-OC group may be weaker. However the Kruskal-Wallis test indicated that there was no significant difference in the pretest performance of the four evaluation subgroups, $p = .0543$. The table of multiple comparisons did not provide any additional information and is therefore not included.

Figure 4.7 Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of pretest scores

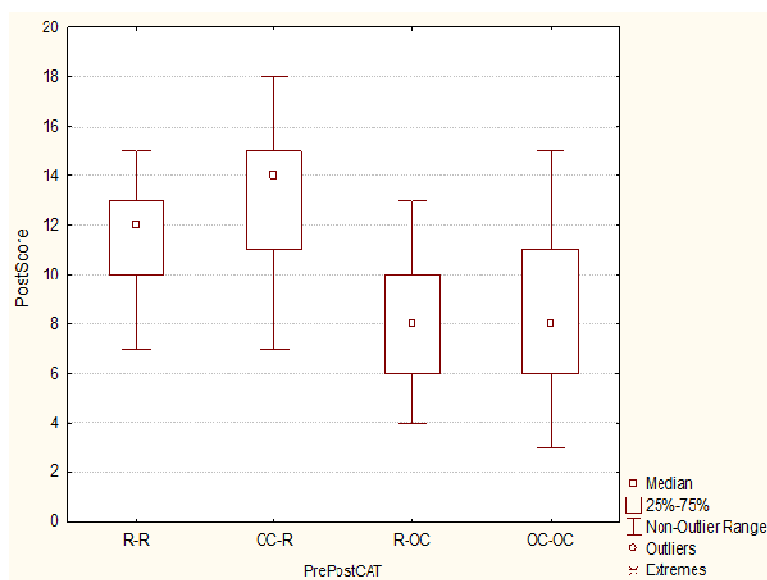


It is evident from Figure 4.7 that extensive overlap existed between the pretest scores of the subgroups. This can be due to the fact that the students were unprepared since they had not yet been taught stoichiometry. There was therefore likely to be a lot of guessing of answers which could not be justified by any knowledge base.

ii. Comparison of posttest scores

Statistical analysis of the posttest results of the four performance evaluation subgroups was conducted and the results are presented as a boxplot in Figure 4.8 and in Table 4.13.

Figure 4.8: Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of posttest scores



The boxplot (Figure 4.8) shows minimal overlap between two sets of the 25 - 75% boxes for the individual subgroups. The medians of the posttest scores of students who were realistic in their judgement during the posttest (R-R and OC-R) are higher than those of the overconfident students (OC-OC and R-OC). It seems that the students who were realistic in their performance evaluation during the posttest performed better than the students who were overconfident. This is confirmed by the results obtained in the Kruskal-Wallis test. Table 4.13 shows that there is an overall significant difference in how the students performed in the posttest, $p = .0001$. Significant differences in terms of posttest scores were observed between the R-R and R-OC ($p = .0443$), R-R and OC-OC ($p = .0347$), OC-R and R-OC ($p = 0.0023$) as well as between the OC-R and OC-OC ($p = 0.0007$) subgroups. Significant differences were observed for the students who were realistic in their judgement during the posttest (OC-R, R-R) and the students who were overconfident in the posttest (R-OC, OC-OC), but not between the two subgroups that were realistic in the posttest (R-R and OC-R) or the two subgroups that were overconfident in the posttest (OC-OC and R-OC).

Table 4.13: Multiple comparisons (p-values) of posttest performance of students in the four pre-post performance evaluation subgroups

Kruskal-Wallis test: $H(3, N=89) = 22.23791$ $p = .0001$				
	R-R	OC-R	R-OC	OC-OC
R-R		1.000000	0.044300	0.034785
OC-R	1.000000		0.002342	0.000703
R-OC	0.044300	0.002342		1.000000
OC-OC	0.034785	0.000703	1.000000	

4.3.3.1.5 Comparison of the performance evaluation subgroups in terms of pre- and posttest average confidence scores

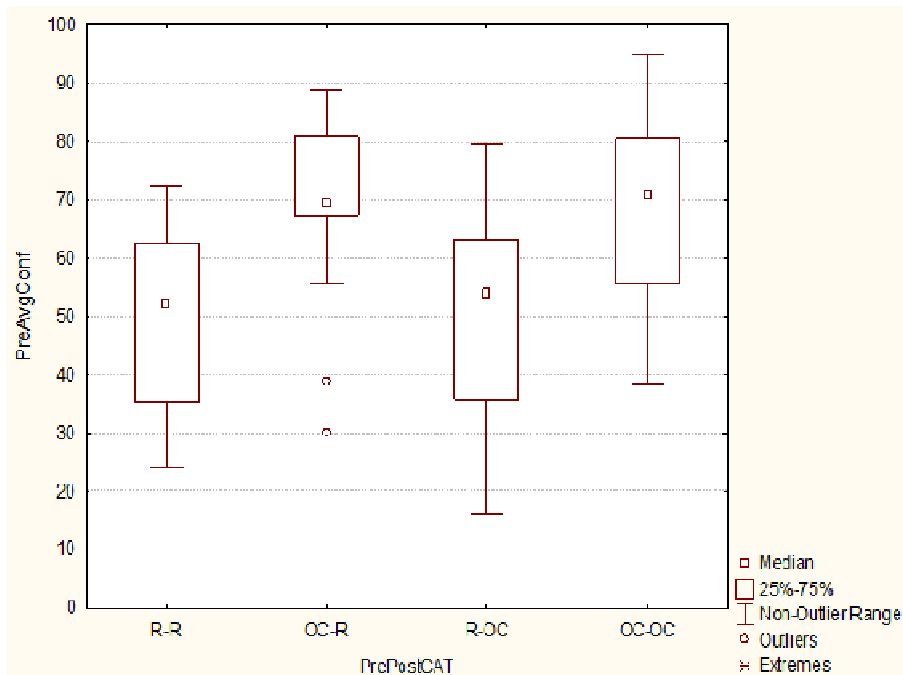
On average the confidence levels of the students were high in both the pre- and posttest (Table 4.7). However, the global view of the data could not show whether there were subgroups whose levels of confidence scores were higher in comparison with the others, whether the subgroups differed in their average confidence scores and whether the difference was significant. The results of comparisons of the four subgroups in terms of their average confidence levels in the pretest are presented in Figure 4.9 and Table 4.14.

i. Comparison of pretest average confidence scores

Table 4.14: Multiple comparisons (p-values) of pretest average confidence scores of students in the four pre-post performance evaluation subgroups

Kruskal-Wallis test: $H(3, N=89) = 19.41114$ $p = .0002$				
	R-R	OC-R	R-OC	OC-OC
R-R		0.081667	1.000000	0.017232
OC-R	0.081667		0.030310	1.000000
R-OC	1.000000	0.030310		0.002332
OC-OC	0.017232	1.000000	0.002332	

Figure 4.9: Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of pretest average confidence scores



The medians for confidence scores of subjects who were realistic in their judgement in the pretest, i.e. subjects in the R-R and R-OC subgroups, are similar, and so are the medians for subjects who were overconfident, i.e. OC-OC and OC-R. The medians of the confidence scores of realistic subjects are also clearly lower than those of the overconfident groups. In the pretest, students in the R-R and R-OC subgroups reported less confidence in the accuracy of their chosen responses than those in the OC-OC and OC-R subgroups.

The significance of the difference in the average confidence scores of the subgroups as shown in Figure 4.9 are confirmed by the Kruskal-Wallis test results in Table 4.14. The Kruskal-Wallis test shows that there is an overall significant difference in the average confidence of the four groups, $p = .0002$. Furthermore the p -values yielded by the individual comparisons of the subgroups in Table 4.14 indicate significant differences between the OC-OC and R-OC subgroups, OC-OC and R-R subgroups as well as between the OC-R and R-OC subgroups. This result should be interpreted together with that of the comparison of pretest performances (Fig. 4.7) where no meaningful difference in the performance of the four subgroups was found ($p = .0543$). The only meaningful difference was found in the average confidence of the subgroups ($p = .0002$). The overconfident students, i.e. students in the OC-OC and OC-R subgroups, were found to report significantly higher average confidence ratings than the students in the R-R and R-OC subgroups. The higher confidence was not justified by higher performance as shown in Figure 4.7.

ii. Comparison of posttest average confidence scores

A statistical analysis of average confidence scores of the four subgroups was conducted and the results are presented as a boxplot in Figure 4.10 and in Table 4.15. In terms of average confidence scores in the posttest, the boxplot in Figure 4.10 shows that medians of groups OC-R and OC-OC are higher than the medians of groups R-R and R-OC. Looking at the finer details, the results of the Kruskal-Wallis test in Table 4.15 below show that a significant difference only exists between the average confidence scores of the R-R and OC-OC subgroups ($p = .0341$). When these results are interpreted together with the statistical analysis of posttest performance (Fig 4.8, Table 4.13) it is clear that whereas the average confidence scores reported by the OC-OC subgroup were higher than the R-R subgroup, their performance was significantly lower. This subgroup clearly lacked the ability of self-evaluation in stoichiometry.

Figure 4.10: Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of posttest average confidence scores

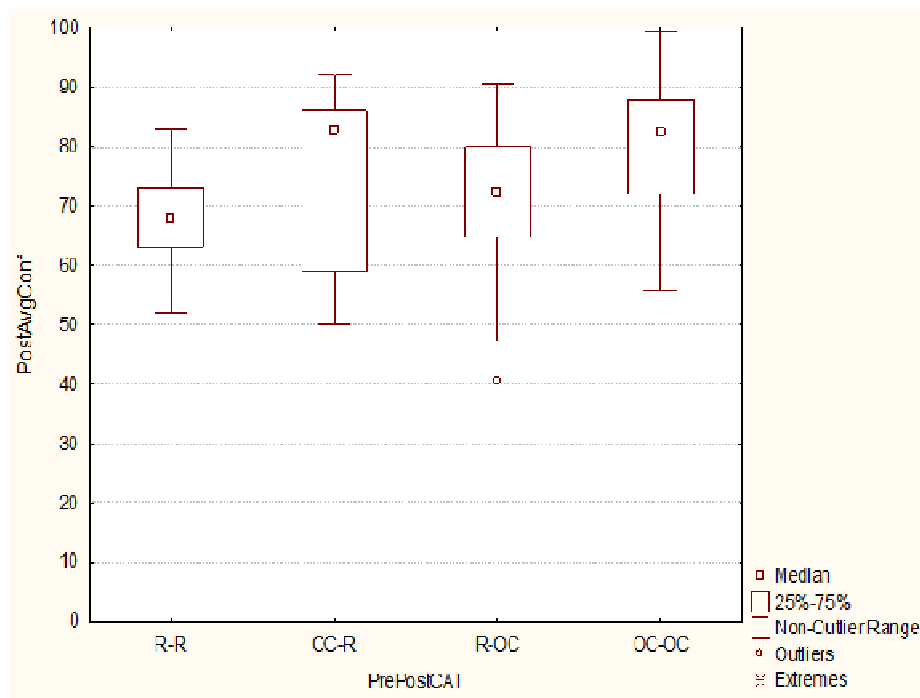


Table 4.15: Multiple comparisons (p-values) of posttest average confidence scores of students in the four pre-post performance evaluation subgroups

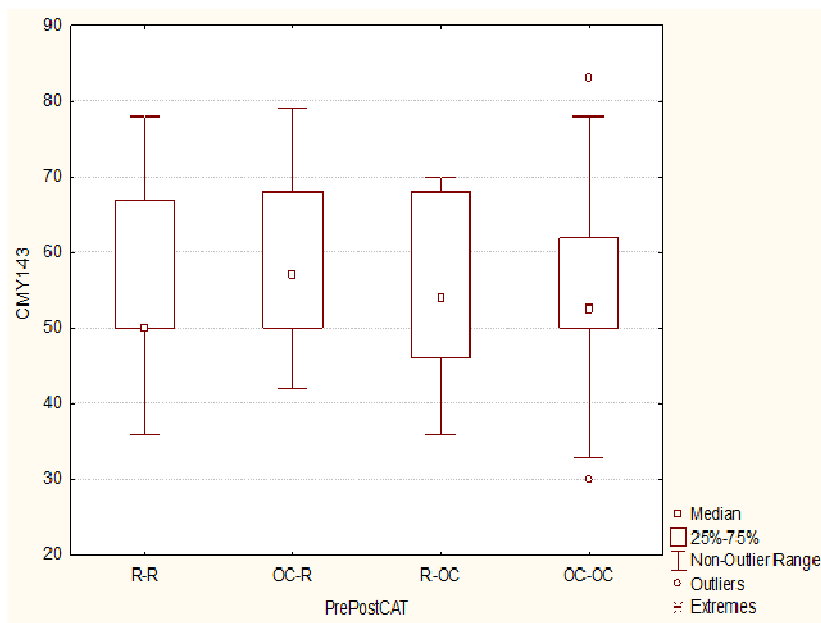
Kruskal-Wallis test: $H(3, N=89) = 10.29770$ $p = .0162$				
	R-R	OC-R	R-OC	OC-OC
R-R		0.485675	1.000000	0.034182
OC-R	0.485675		1.000000	1.000000
R-OC	1.000000	1.000000		0.190336
OC-OC	0.034182	1.000000	0.190336	

4.3.3.1.6 Comparison of the performance evaluation subgroups in terms of CMY 143 performance

A statistical analysis of the CMY 143 end of semester performance of the four performance evaluation subgroups was conducted in order to determine whether similar differences in performance were found as were documented for stoichiometry. The results are shown as a boxplot in Figure 4.11. The difference in the performance of the four subgroups in the final CMY 143 examination was found to be insignificant, $p = .7497$. This is also evident in the extensive overlap between the performance distributions as shown in Figure 4.11.

Stoichiometry is an important, but minor component of the CMY 143 syllabus. It requires analytical reasoning as well as mathematical skills. Students who struggled to master stoichiometry, which is a difficult topic, could perform well in other topics. These results indicate that accuracy of performance evaluation in stoichiometry was not a predictor for end of semester performance.

Figure 4.11: Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of CMY 143 performance



4.3.3.1.7 Comparison of the performance evaluation subgroups in terms of performance gain

The students in the different pre-post performance evaluation subgroups showed different performance gains as reported in Table 4.11, but we wanted to explore whether the difference was significant or not. Statistical analysis was conducted and the results are reported in Figure 4.12 and in Table 4.16.

Figure 4.12: Box plot showing a comparison of the four pre-post performance evaluation subgroups in terms of average gain in performance

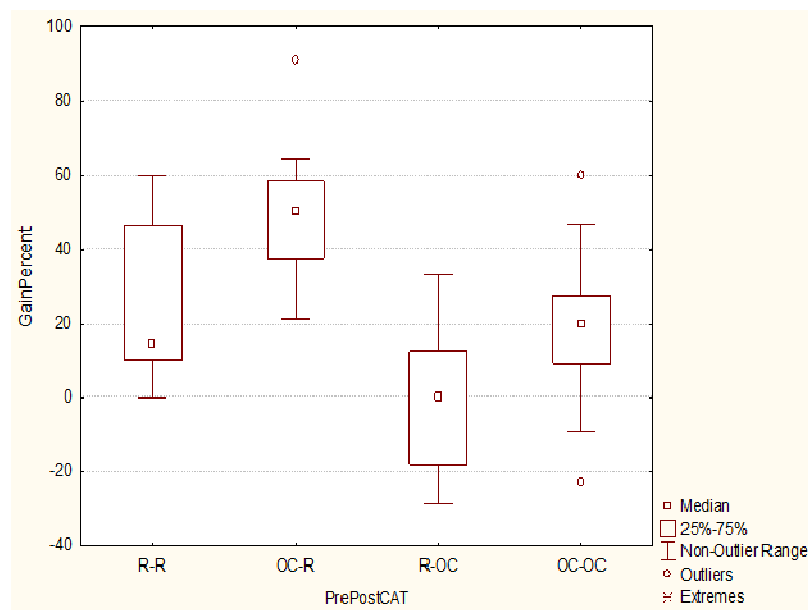


Table 4.16: Multiple comparisons (p-values) of average performance gain of students in the four pre-post performance evaluation subgroups

Kruskal-Wallis test: $H(3, N=89) = 31.53834$ $p = .0000$				
	R-R	OC-R	R-OC	OC-OC
R-R		0.063609	0.042951	1.000000
OC-R	0.063609		0.000000	0.000548
R-OC	0.042951	0.000000		0.013961
OC-OC	1.000000	0.000548	0.013961	

An overall p-value of 0.0000 obtained by the Kruskal-Wallis test indicates that the subgroups differ significantly in terms of performance gain. Looking at the finer details, Table 4.16 shows that the differences between all subgroups were significant except between R-R and OC-OC. The highest performance gain was observed for the OC-R subgroup, moderate and comparable gains were achieved by the R-R and OC-OC groups and virtually no gain was achieved by the R-OC subgroup.

4.3.3.1.8 Summary

To summarise the students in the four subgroups differed significantly in terms of how they performed in their posttests, their pre- and posttest average confidence scores and in performance gain. The difference in their performance in the first-semester chemistry

module, CMY 133, was marginal but insignificant. A significant difference was also not found with regard to pretest scores and performance in the CMY 143 end of semester examination. These findings confirmed that we were dealing with four discrete groups with different characteristics, but that care should be taken in the interpretation of results based on raw score data as shown in Table 4.11. The four subgroups had a comparable level of preknowledge as judged by their CMY 133 and pretest performance, but they differed significantly in terms of the learning gains demonstrated in posttest performance after having been taught the difficult topic of stoichiometry. However the respective learning gains achieved by the four subgroups in stoichiometry were not a predictor for end of semester performance in CMY 143.

4.3.4 Questionnaire responses

It became clear from studies reported in the literature (e.g. Gramzow *et al.*, 2003) that lack of knowledge may not be the only factor leading students to be biased in the evaluation of their performance in a test. Bias in performance evaluation may be associated with several psychological factors discussed in the literature review in Chapter 2. Gramzow *et al.* (2003) argued that overly-positive self-reports made by students with low actual grades were self-protection motivated while overly-positive self-reports by top performers were motivated by the need to self-enhance. Informed by this argument we designed a questionnaire to assist us to investigate the tendency of students in our sample to self-enhance or self-protect and to determine the relationship between bias in performance evaluation and these two constructs.

The questionnaire was designed according to the procedure explained in paragraph 3.5.2, and it consisted of statements with which students had to agree or disagree on a seven-category Likert scale. As explicated in paragraph 3.5.2.1 the questionnaire comprised ten items on self-enhancement and nine items on self-protection. Upon subjecting the instrument to statistical analysis to ascertain construct validity, the presence of three discrete constructs was revealed, i.e. two forms of self-enhancement labelled SEa and SEb as well as the construct of self-protection labelled SP. Items under the SEa construct were concerned with a desire to portray oneself to others as a hardworking, clever student. The SEb items measured the need to give a good impression about oneself to others and finally, items under the SP construct were concerned with the need to protect one's academic self-worth by attributing failure to external factors. The three factors and the items that loaded strongly onto them during factor

analysis are presented in Table 4.4. Moreover students were requested to report biographical nominal data such as gender and age in the questionnaire.

The entries that each student made on the Likert scales from one to seven were added to obtain a total score per construct. The total score per construct was interpreted as the strength of the students' endorsement of statements in that construct. Such data, together with data on gender, were used to answer research question four, which is to determine the relationship between:

- inaccuracy in performance evaluation and self-enhancement;
- inaccuracy in performance evaluation and self-protection and
- inaccuracy in performance evaluation and gender.

4.3.4.1 The relationship between bias in performance evaluation and the self-enhancement motivational factor

Two forms of self-enhancement were identified by factor analysis. The first form, abbreviated as SEa, measures the desire to portray oneself to others as a hardworking, clever student and the second form, abbreviated as SEb, measures the tendency to give a good impression about oneself to others. Figures 4.13 and 4.14 are scatterplots indicating the relationship between inaccuracy in performance evaluation in the pre- and posttests, respectively, and the self-enhancement motivational factor. Inaccuracy in performance evaluation is plotted on the Y-axis against the score for the self-enhancement motivational factor on the X-axis. The value on the Y-axis is determined for each student as the difference between expected performance and actual performance as explained in paragraph 3.10.1 The two scatterplots show that there is a weak, positive relationship between inaccuracy in performance evaluation and SEa scores, correlation coefficients were 0.08 and 0.01 for the relationship between the SEa scores and pre- and posttest inaccuracy in performance evaluation, respectively. However, these relationships were statistically insignificant as indicated by p-values of .425 and .931 for the pre- and posttests respectively.

Figure 4.13: Scatterplot showing the relationship between bias in performance evaluation in the pretest and SEa scores

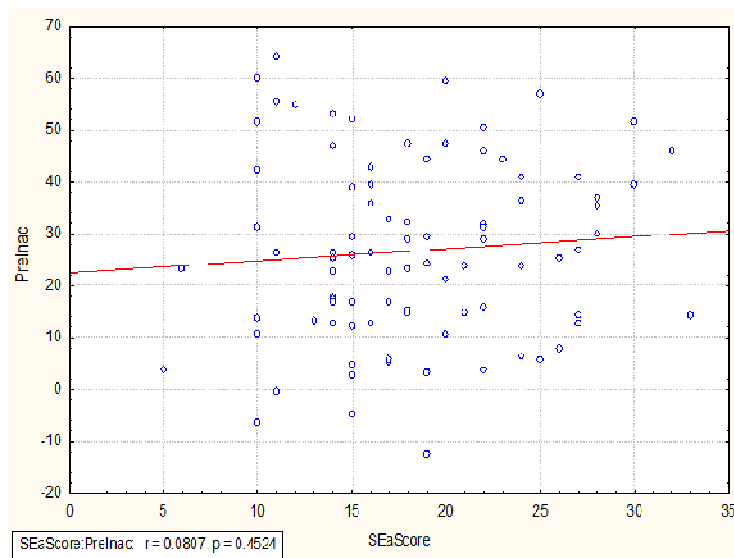
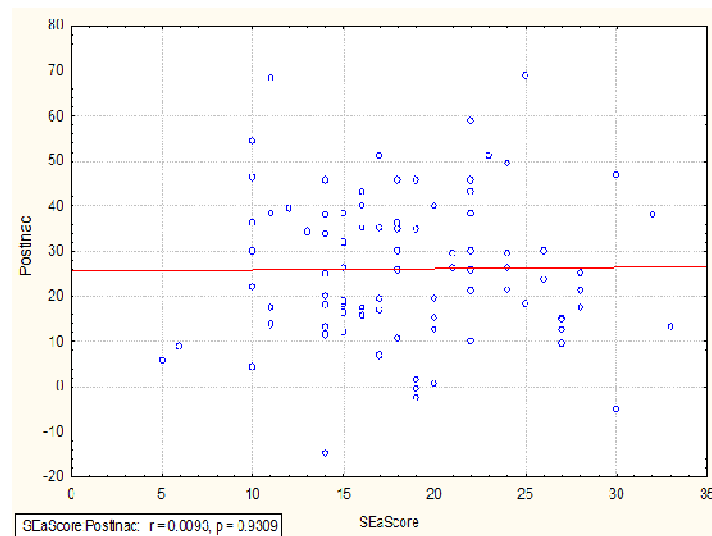


Figure 4.14: Scatterplot showing the relationship between bias in performance evaluation in the posttest and SEa scores



Figures 4.15 and 4.16 are scatterplots indicating the relationship between inaccuracy in performance evaluation in the pre- and posttest and the second form of the self-enhancement motivational factor. The scatterplots show a weak, negative relationship between the SEb scores and pre- and posttest inaccuracy in performance evaluation respectively. Correlation coefficients of -0.05 and -0.08 were calculated for the pre- and posttests respectively. P-values of .63 and .49 for the pre- and posttests, respectively, indicated an insignificant relationship between the two constructs.

Figure 4.15: Scatterplot showing the relationship between bias in performance evaluation in the pretest and SEb scores

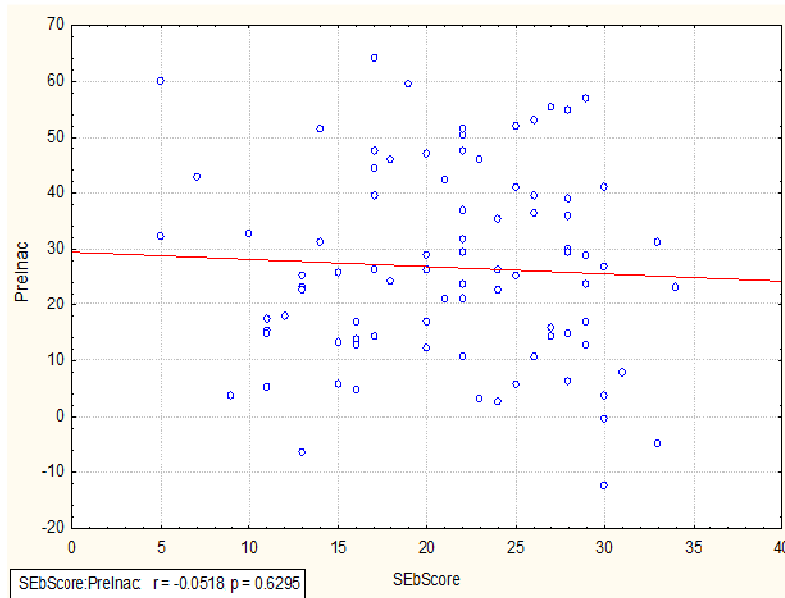
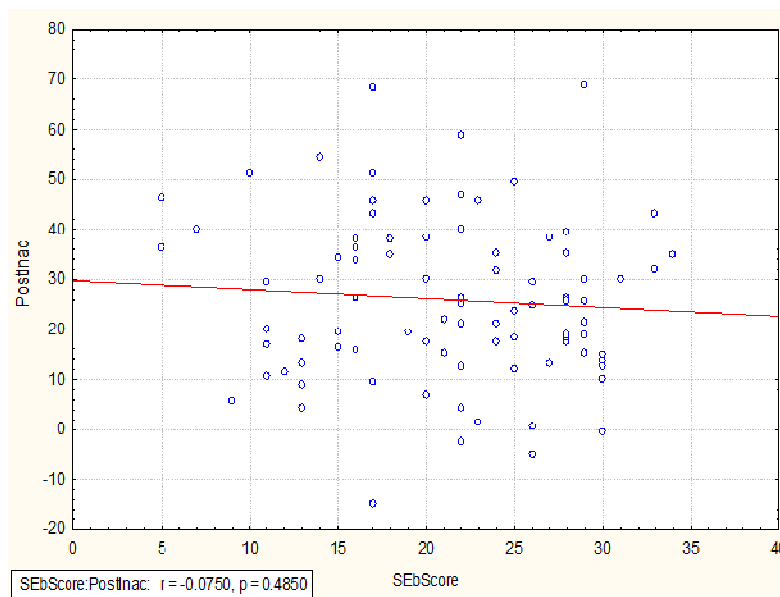


Figure 4.16: Scatterplot showing the relationship between inaccuracy in performance evaluation in the posttest and SEb scores



4.3.4.2 The relationship between bias in performance evaluation and the self- protection motivational factor

Figures 4.17 and 4.18 are scatterplots showing the relationship between inaccuracy in performance evaluation in the pre- and posttests and self-protection scores. The scatterplots show a weak, negative relationship between the pre- and posttest inaccuracy in performance evaluation and the self-protection scores. Correlation coefficients of -0.19 and -0.08 were

calculated for relationships between the SP scores and the pre- and posttest inaccuracy in performance evaluation respectively. The relationship between the two constructs was, however, not significant; p-values are .07 and .46 for the pre- and posttests, respectively.

Figure 4.17: Scatterplot showing the relationship between inaccuracy in performance evaluation in the pretest and SP scores

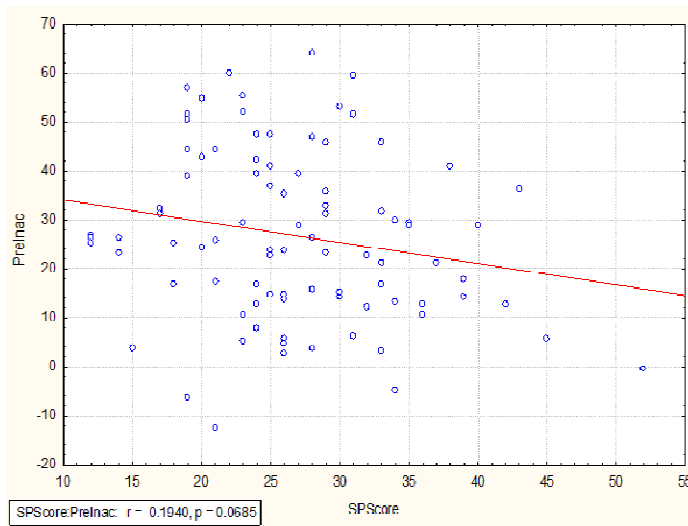
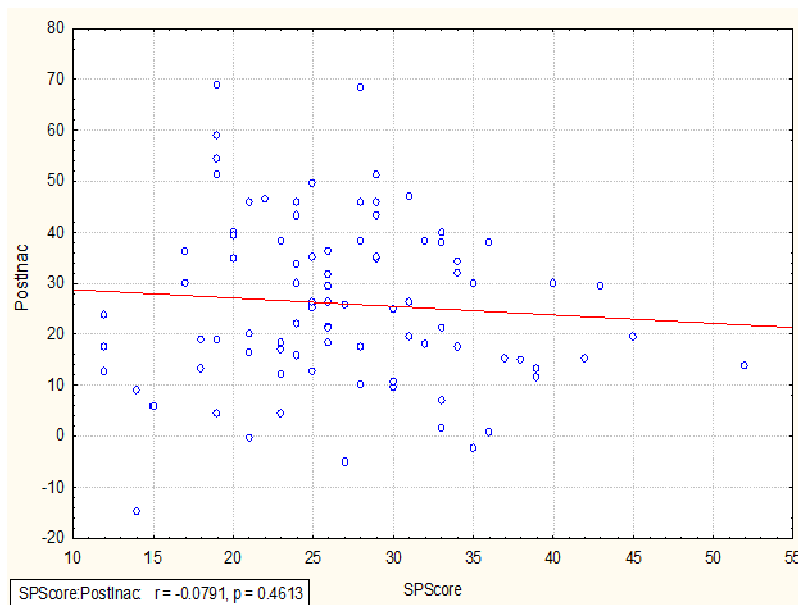


Figure 4.18: Scatterplot showing the relationship between inaccuracy in performance evaluation in the posttest and SP scores



4.3.4.3 The relationship between bias in performance evaluation and gender

Comparisons were made between accuracy of performance evaluation between male and female respondents. T-test p-values of .50 and .56 for the pre- and posttest, respectively, showed that males and females did not differ significantly in their inaccuracy of performance evaluation, i.e. males and females were equally inaccurate in the evaluation of their pre- and posttest performance.

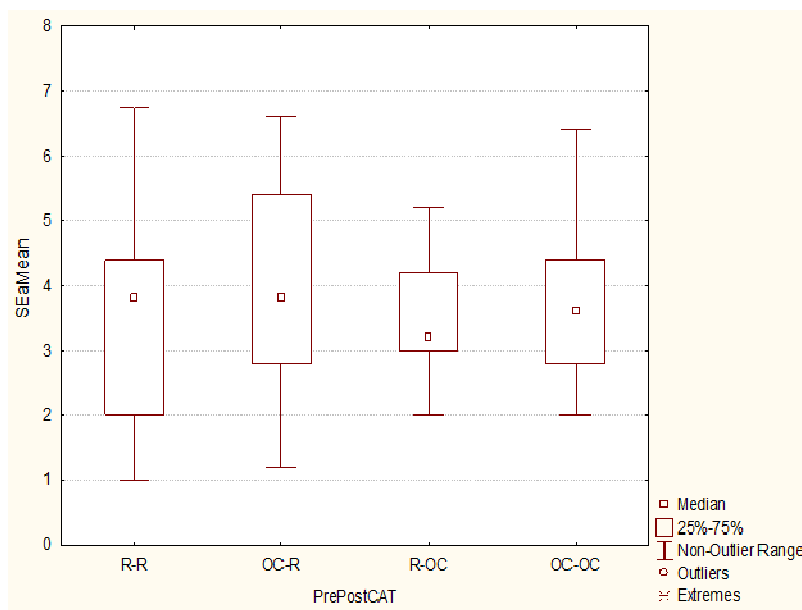
To summarise, the relationship between bias in performance evaluation and the first form of self-enhancement was found to be weak, positive but insignificant. The relationship between inaccuracy in performance evaluation and the second form of self-enhancement was also weak, negative but insignificant. The relationship between bias in performance evaluation and self-protection was negative, weak and insignificant. P-values greater than 0.05 in the pre- and posttest indicated that males and females were not significantly different in their bias in performance evaluation.

We have reported in paragraph 4.3.3.1.2 that students who were overconfident in the evaluation of their test performance after instruction, i.e. students in the OC-OC and R-OC subgroups performed poorly in the posttest (average of posttest scores of 45% and 43% for the OC-OC and R-OC subgroups respectively) and fewer than 50% of them passed in the pre- and posttests. According to Gramzow *et al.* (2003), inaccuracy of performance evaluation of students with low grades is motivated by the tendency to self-protect while inaccuracy in performance evaluation of students with high grades is self-enhancement motivated. Taking students in the OC-OC and R-OC subgroups as poor performers and taking the students in the R-R and OC-R subgroups as top performers based on their average of their posttest scores, we sought to determine if based on their performance evaluation subgroups, there would be any difference in the tendency to self-enhance or self-protect. Statistical measures were employed to determine whether there was a significant difference between students in the pre-post performance evaluation subgroups in terms of the three motivational factors measured in our questionnaire instrument namely, SEa, SEb and SP.

4.3.4.4 Comparison of the four performance evaluation subgroups in terms of the three motivational factors (SEa, SEb, SP)

The mean scores of the SEa, SEb and SP motivational factors of the four performance evaluation subgroups were computed and compared using the Kruskal-Wallis test. This analysis was done to check for the existence of a significant difference in the tendency to self-enhance or self-protect amongst the performance evaluation subgroups in both the pre- and posttests. Figure 4.19 is a boxplot generated after statistical analysis of the SEa scores of the four subgroups. A p-value of 0.8273 was obtained indicating that the four performance evaluation subgroups do not differ significantly in their tendency to self-enhance. Box plots of the R-R, OC-R and OC-OC subgroups show a wider range compared with the R-OC subgroup. The range of the R-OC subgroup is narrow and the median is smaller compared with that of the other three subgroups. Though the difference between the subgroups is not significant, subjects in the R-OC subgroup have a lower tendency to self-enhance. It would be interesting however, to see if a different outcome is obtained when using a bigger sample. It may be possible to obtain a clearer picture with a larger sample. It seems for all four subgroups the need to portray oneself to others as a hardworking, clever student is not correlated with their accuracy or inaccuracy in performance evaluation during the pretest and posttest.

Figure 4.19: Box plot showing a comparison of the self-enhancement levels (SEa) of the four pre-post performance evaluation subgroups



The boxplots in Figure 4.20 represent a comparison of the four subgroups in terms of the distributions of their SEb scores. For the second type of self-enhancement factor (SEb), subjects were also not significantly different in their tendency to self-enhance. A p-value of 0.8370 was obtained in the Kruskal-Wallis test. Looking at the box plot in Figure 4.20, the R-OC subgroup has a low median compared with the other three subgroups which means that though the subgroups do not differ significantly the subjects in the R-OC subgroup demonstrate lower tendencies to self-enhance, i.e. the need of these students to give a good impression about themselves to others tends to be less compared with students in the other three performance evaluation subgroups. The same trend was observed for the R-OC subgroup in terms of the SEa factor, but the difference was also not statistically significant. It is interesting to note that the R-OC group showed the weakest performance in the posttest. In order to support the findings of Gramzow *et al.* (2003) this subgroup should show the highest tendency to self-protect.

Figure 4.20: Box plot showing a comparison of the self-enhancement levels (SEb) of the four pre-post performance evaluation subgroups

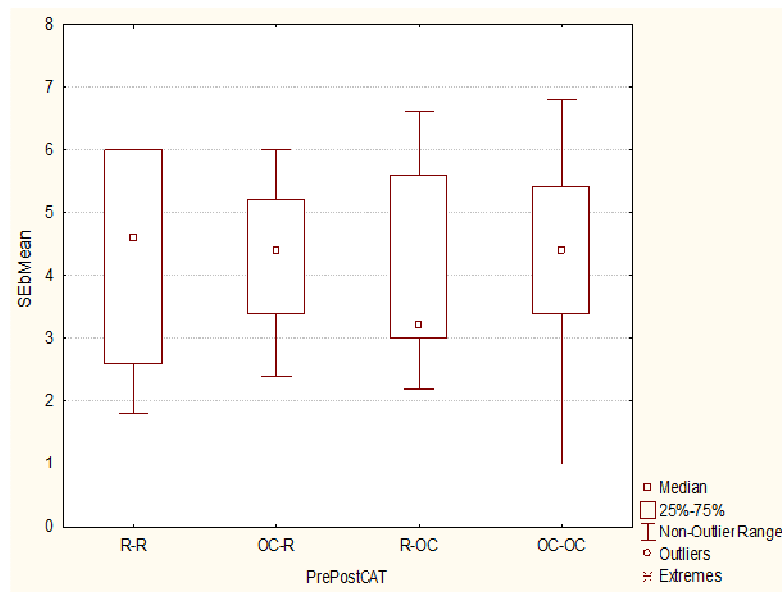
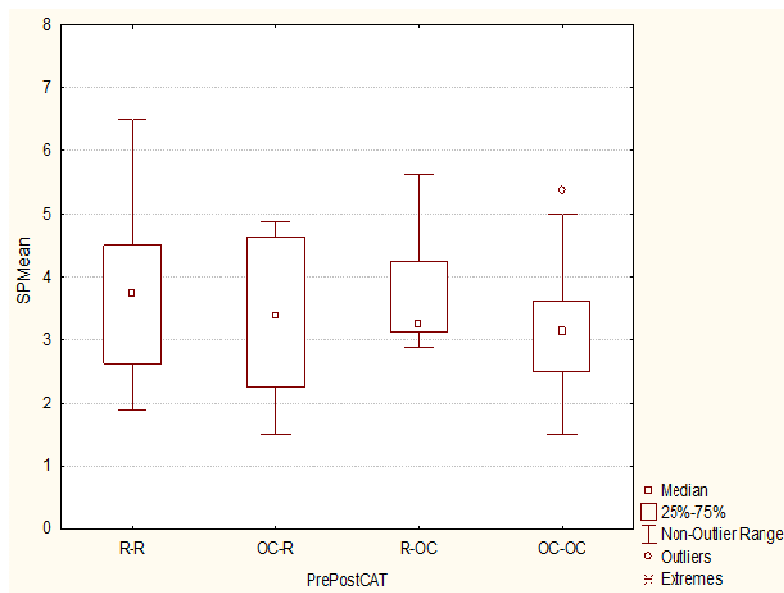


Figure 4.21 below shows boxplots which indicate how the four subgroups compare in their tendency to self-protect. In terms of the tendency to self-protect, the four performance evaluation subgroups do not differ significantly, $p = .1992$. In Figure 4.21 the OC-OC subgroup has the lowest median compared with the other three subgroups, however only by a small margin. Results show that students in the four performance evaluation subgroups do not significantly differ in the need to protect their academic self-worth. Noteworthy is that the

R-OC subgroup does not show a high tendency to self-protect as the findings of Gramzow *et al.* (2003) would have predicted.

Figure 4.21: Box plot showing a comparison of the self-protection levels (SP) of the four pre-post performance evaluation subgroups



To summarise, there is no significant difference amongst the four subgroups in all three motivational factors, namely SEa, SEb and SP. The tendency by the four performance evaluation subgroups to self-enhance or self-protect was not found to be statistically different and no inferences could be made from the results. We were therefore unable to confirm the findings of Gramzow *et al.* (2003) that the motivational factors underlying inaccuracy in self-evaluation differ for different performance groups. It is possible that significant differences may be observed when a larger sample is used.

4.3.5 Qualitative data analysis

A three-tier test instrument on the topic of stoichiometry was used to collect data on students' performance and accuracy of performance evaluation. In the third part of each item of the stoichiometry test, students were requested to motivate their choice of confidence rating in a free response format (see paragraph 3.5.1.2). The purpose of providing students with an opportunity to explain their choice of confidence rating was to collect data which could assist us in understanding the factors that have motivated their choice of confidence ratings in the tests. This qualitative data was instrumental in answering research question three:

Research question 3: What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?

Having realised that a global view of quantitative data did not enable us to identify students who had shown improvement with regard to accuracy of performance evaluation, we decided to separate the students into four subgroups based on whether they became realistic or remained overconfident in their evaluation (paragraph 4.3.3.1). In the analysis of the three categories for accuracy of performance evaluation, namely the Overconfident, Realistic and Underconfident, four subgroups emerged. The subgroups were labelled as the OC-OC, OC-R, R-R and the R-OC subgroups as explained earlier. Statistical measures were employed to examine whether the four subgroups were discrete groups which differed significantly in terms of some of the characteristics presented in Table 4.11 (see paragraphs 4.3.3.1.1 to 4.3.3.1.8). It was necessary, therefore, to explore the factors that students relied on in the making of their judgements of learning as well as the influence that teaching had on those factors, within these four pre-post performance evaluation subgroups. Students in the R-R and OC-R subgroups distinguished themselves by demonstrating the quality of being able to gain from the teaching and learning process (as shown by performance gain) and realistically evaluate their performance in the posttest. The OC-OC subgroup never mastered the metacognitive skill of accurate performance evaluation, yet they demonstrated moderate learning gain while the R-OC subgroup did not gain anything from the teaching and learning experience yet became excessively confident in their mastery of stoichiometry. In answering research question three we wanted to understand how students, who were consistently overconfident or became overconfident in the posttest, differed from the realistic students in terms of the factors they reported when motivating their choice of confidence ratings. We also wanted to explore the influence of teaching on these factors.

The method used to organise and code qualitative data will be described in the following paragraphs in order to demonstrate the rigour with which qualitative data were analysed. A software package called ATLAS.ti version 4.2 was used to systematically organise the students' free responses according to their pre-post performance evaluation subgroups, for coding and for categorising data into themes. Excel was used to capture each student's free responses. The free responses were copied and pasted into Microsoft paint documents so that they could be easily assigned into ATLAS.ti hermeneutic units. Four hermeneutic units were

created, namely OC-OC, OC-R, R-R and R-OC. The responses of all the students in the OC-OC subgroup were grouped together. The same was done for responses of all students in the OC-R, R-R and R-OC subgroups, respectively. This was done to make it possible to separately analyse the students' responses according to their pre-post performance evaluation subgroups.

To analyse the qualitative data, the responses were read and re-read and coded systematically in each hermeneutic unit. To avoid multiple assignments of codes to the same response, I had to read and re-read each response to find the message conveyed by the written response and code it accordingly. To reduce the codes into a reasonable number of categories, my supervisors and I first read through the codes to find themes that were recurring in the data. Related codes were combined into themes. Each theme was assigned an identifying name using descriptive words from the text to establish a category. To aid the validity of the study, the vital step was sounding out the labels to my supervisors to see whether the labels made sense (Nieuwenhuis, 2007). In this way categories were allowed to emerge from the codes. Through analysis of the identified codes, twenty-seven categories were identified. For greater conceptual clarity categories that were linked to the same concept were grouped together to create super-categories. The categories and super-categories that emerged through this method are listed in Table 4.17. While data is being coded, Atlas.ti counts the frequency with which the codes occur. This function of the software enabled the easy enumeration of responses. Information on the frequency with which codes per categories occurred enabled us to determine which super-categories and categories were prevalent or dominant in each pre-post performance evaluation subgroup first in the pre- and then the posttest. The process of enumeration involves counting the number of times a category is applied to the data to obtain a sense of how often a specific phenomenon appears in the data (Nieuwenhuis, 2007). This was also helpful in tracing changes in the factors students in the different pre-post performance evaluation subgroups relied on when judging their performance after instruction. Table 4.17 shows the categories and super-categories that were generated from the emerging codes.

Table 4.17: Categories and super-categories generated from emerging codes

Super category (SC)	Categories (C)
DECLARATIVE KNOWLEDGE (SC1)	C7: What the person thinks is required to solve a problem. E.g. <i>“used conversion factor method”</i>
	C9: Uncertainty due to deficit in general declarative knowledge. E.g. <i>“Not sure of theory”</i>
	C19: Perceived possession of declarative knowledge required to solve the problem. E.g. <i>“know theory”</i>
PROCEDURAL KNOWLEDGE (SC2)	C8: Uncertainty due to deficit in general procedural knowledge (approach and method). E.g. <i>“not sure how to work out”</i>
	C15: Certainty based on the perceived possession of specific procedural knowledge that can be named and demonstrated – (balancing/calculations). E.g. <i>“sure of calculation” or student has shown how answer has been worked out or provided a balanced equation.</i>
	C22: Lack of specific procedural knowledge. E.g. <i>“don’t know how to intermingle particles”</i>
	C23: Perceived possession of relevant skills to solve the problem. E.g. <i>“Can calculate limiting reactants”</i>
	C27: Perceived possession of test-taking skills. E.g. <i>“used process of elimination”</i>
GLOBAL EVALUATION OF ANSWER (SC3)	C1: Estimation of chance. E.g. <i>“50% chance of getting answer right”</i>
	C2: Subjective feeling of doubt or uncertainty. E.g. <i>“Doubtful”</i>
	C3: Informed guess. E.g. <i>“Guess”</i>
	C5: Lack of confidence. E.g. <i>“don’t feel confident”</i>
	C6: Vague judgement of how answer looks in comparison with given responses. E.g. <i>“it’s right”</i>
	C10: Certainty due to a subjective, vague feeling not based on any evidence. E.g. <i>“sure of answer”</i>
	C11: Certainty due to a subjective feeling based on perceived correctness of answer. E.g. <i>“believe answer is correct”</i>
	C12: Estimation of confidence in approach or answer. E.g. <i>“confident with work”</i>
	C13: How own answer compares with given multiple choice options. E.g. <i>“closest value to own answer”</i>
	C14: Certainty due to an appeal that answer makes sense or is logical. E.g. <i>“Answer makes sense”</i>
	C17: Unreflective, almost defensive evaluation. E.g. <i>“just know”</i>
C18: Global feeling that answer is right. E.g. <i>“this is how I feel”</i>	
FEELING OF UNPREPAREDNESS (SC4)	C4: Feeling of general inadequacy and unpreparedness. E.g. <i>“haven’t studied yet”</i>
EXTERNAL FACTORS (SC5)	C20: Deficits of the question. E.g. <i>“vague question”</i>
	C24: Familiarity with question or concept. E.g. <i>“done in high school”</i>
	C25: Unfamiliarity with question or concept. E.g. <i>“never seen such question before”</i>
METACOGNITIVE STRATEGIES (SC6)	C16: Metacognitive strategy. E.g. <i>“double checked solution”</i>
MATH SKILLS (SC7)	C21: Uncertainty due to inadequate mathematical skills. E.g. <i>“No idea how to calculate”</i>
LACK OF MEMORY (SC8)	C26: Lack of memory. E.g. <i>“Forgot how to work out”</i>

4.3.5.1 Qualitative data section of the three-tier instrument: Interrater reliability

Interrater reliability was calculated to determine the level of agreement between independent coders. Analysis of qualitative data obtained through the third tier of the chemistry test instrument was conducted through the method of thematic analysis, explained in more detail in paragraph 3.10.3. This is a method which entails assigning emerging codes to meaningful units of qualitative data and categorising these codes into themes which are in turn used for data interpretation.

The system employed for the identification of codes and the coding of data had to be also subjected to a statistical measure to ensure reliability. To ensure reliability of the method and the validity of the designated codes, an independent coder was asked to code some of the data. The independent coder was provided with raw text and a list of coding categories as well as super-categories and asked to randomly choose text data to code. The coding of the independent rater was then compared with my coding in order to determine inter-coder reliability. According to Nieuwenhuis (2007), inter-coder reliability is the consistency among different coders in terms of the assignment of a specific code or category to a specific piece of text. Qualitative research becomes more defensible when more than one coder is used and high inter-coder reliability is obtained (Nieuwenhuis, 2007). Interrater or inter-coder reliability is defined as a measure of the level of agreement between two raters or coders in the assignment of categories to text data. It provides valuable information regarding the effectiveness of an employed coding system. Cohen's kappa is a statistical measure of interrater reliability. It normally ranges from 0 to 1.0. Fleiss (1981) suggests that values of kappa less than 0.40 are a reflection of poor agreement. Kappa between 0.40 and 0.75 indicate fair to good agreement and kappa values above 0.75 indicate strong agreement.

After coding the students' free responses, 18 students were randomly chosen from the sample by the independent coder and their uncoded, free responses for each item in the pre- and posttests were coded by the coder using the provided list of categories and super-categories. My codes together with the codes of the second rater were recorded and the level of agreement between the two coding systems was determined by computing simple kappa values. Simple kappa's were used as opposed to weighted values of kappa because our codes were discrete codes with no relationship between them, i.e. if the response was to agree or disagree, the responses did not differ in terms of their degrees or extent of agreeing or disagreeing; they were discrete codes. Table 4.18 shows the item number, the value of

simple kappa for each item in the pre- and posttest as well as the level of agreement determined between the coding systems of the two coders according to the criteria in Fleiss (1981), for responses of all 18 students randomly chosen by the independent coder.

Table 4.18: Cohen's kappa values and the level of agreement observed between the coding systems of the two coders per item

Item*	Pretest values of Cohen's kappa	Agreement	Posttest values of Cohen's kappa	Agreement
1	0.44	Fair	0.52	Fair
3	0.55	Fair	0.75	Good
4	0.55	Fair	0.64	Good
5	0.82	Strong	0.52	Fair
6	0.54	Fair	0.54	Fair
7	0.73	Good	0.59	Fair
8	0.58	Fair	0.70	Good
9	0.78	Strong	0.35	Poor
10	0.73	Good	0.79	Strong
11	0.63	Good	0.73	Good
12	0.58	Fair	0.69	Good
13	0.62	Good	0.61	Good
14	0.53	Fair	0.72	Good
15	0.72	Good	0.70	Good
16	0.79	Strong	0.53	Fair
17	0.63	Good	0.49	Fair
18	0.68	Good	0.73	Good
19	0.71	Good	0.90	Strong
20	0.83	Strong	0.91	Strong

* Data for item 2 were not analysed as explained in paragraph 4.2.1.2

In the pretest the agreement between the coding of the two coders ranges from fair to strong with kappa values from 0.44 to 0.83. A fair agreement between the coders is observed in less than 50% of the items, i.e. seven items (items 1, 3, 4, 6, 8, 12 & 14). A good agreement is observed in eight items (items 7, 10, 11, 13, 15, 17, 18, 19). Kappa values above 0.75 are

observed for four items (items 5, 9, 16 & 20) which is an indication of strong agreement between the coders.

In the posttest the coders reached poor agreement ($\kappa = 0.35$) in only one item, i.e. item 9. This disagreement was handled as follows: For item 9 with poor agreement the coders looked at the individual student responses and the corresponding codes assigned by both coders to determine the most appropriate code that would be acceptable to both coders. Two instances in item 9 were observed where one coder assigned super-category 7 (Mathematical skills) and the other coder assigned super-category 2 (procedural knowledge). The commonality in the two student responses was that all the students referred to an inability to calculate or an uncertainty in the calculation and this was seen as a more specific form of procedural knowledge. The two coders agreed to assign super-category 7 (Mathematical skills) which is more specific than super-category 2 (Procedural knowledge) to both responses. Only one instance was found in which one coder assigned super-category 7 (Mathematical skills) and the other coder assigned super-category 3 (Global evaluation of answer). In the response the student referred to an uncertainty about the calculation and this was seen as a more specific response. The coders agreed to assign super-category 7 (Mathematical skills) to that response. In another instance (item 9) one coder assigned super-category 2 (Procedural knowledge) while the other assigned super-category 4 (Feeling of unpreparedness) to the same response. The coders agreed that the response “I am not sure how to approach the problem” was not indicating a lack of preparedness but rather a lack of procedural knowledge, i.e. how to go about solving the problem. Super-category 2 was assigned to this student response. The last instance in item 9 where there was no agreement in terms of assignment of codes was observed in one student’s response (“Do not understand the equation”), which was assigned super-category 4 (Feeling of unpreparedness) by one coder and super-category 1 (Declarative knowledge) by the other coder. The data was searched for instances in which the two coders agreed on an assigned super-category for a similar response, but such instance was not found. The coders then reached consensus that the response was an indication of the lack of declarative knowledge rather than a lack of preparation. Super-category 1 was then assigned to the response. The coders agreed in their assignment of super-categories to the rest of the remaining thirteen out of a total of fifteen responses in posttest item number 9. In item 9 all the responses for which an agreement was reached have been assigned super-category 3 (Global evaluation of answer). Table 4.18 above shows that apart from item 9 for which the coders reached poor agreement, a fair to good agreement indicated by values of κ

between 0.40 and 0.75 was reached for more than 50% of the items (items 1, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17 and 18) while a strong agreement indicated by values of kappa ranging from 0.79 to 0.91 was observed for 3 items (items 10, 19 and 20).

4.3.5.2 Presentation and interpretation of qualitative data

Table 4.19 captures student responses about how they would explain their choice of confidence indicator in the test. Responses are clustered by performance evaluation subgroups, i.e. all the responses of students in the OC-OC subgroup are grouped together, similarly for OC-R, R-OC and R-R subgroups. A thematic analysis was conducted to determine whether different patterns of responses would be evident for the four performance evaluation subgroups. The data were therefore analysed separately by pre-post performance evaluation subgroups. For each performance evaluation subgroup, the table shows the number of times a category emerged first in the pretest and then in the posttest. The number of times a category appears is then converted into a percentage, first in the pretest and then the posttest.

Some categories had very low incidences of occurrence and therefore in deciding whether the frequency with which a specific category or super-category was cited was significant or not, an arbitrary cut-off value of 5% was adopted. In the discussion that follows of how the response patterns of the students in the different pre-post performance evaluation subgroups compared in the pre- and the posttest, I will focus only on the categories and super-categories with a frequency of occurrence of 5% or higher. Based on this criterion, the super-categories labelled Declarative knowledge (SC1), Procedural knowledge (SC2), Global evaluation of answer (SC3), and External factors (SC5) emerged as major super-categories. The incidences of occurrence of super-category 4 (The feeling of unpreparedness) were of marginal prominence while the Metacognitive strategies (SC6), Mathematics skills (SC7) and Lack of memory (SC8) super-categories showed incidences of occurrence which were too scarce to be considered in our discussion of qualitative data. The super-category labelled 'No explanations' (SC9) covers all the entries where students omitted explanations of their choice of confidence judgement ratings. Super-category 10 labelled 'Others' represents all the responses that could not be coded such as incomplete sentences, incomplete equations, etc.

Table 4.19: Students' open-ended responses about how they would explain their choice of confidence judgement ratings

Categories (C) and Super-Categories (SC)		OC - OC (50)				OC - R(13)				R - OC(15)				R - R(11)			
		PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%
C7	What person thinks is required to solve problem	27	2.84	44	4.63	10	4.05	11	4.45	7	2.46	3	1.05	6	2.87	8	3.83
C9	Uncertainty: deficit in general declarative knowledge	9	0.95	12	1.26	0	0.00	2	0.81	3	1.05	5	1.75	3	1.44	4	1.91
C19	Perceived possession of declarative knowledge required to solve problem	12	1.26	19	2.00	1	0.40	1	0.40	7	2.46	9	3.16	2	0.96	11	5.26
SC1	DECLARATIVE KNOWLEDGE SUBTOTALS	48	5.05	75	7.89	11	4.45	14	5.67	17	5.96	17	5.96	11	5.26	23	11.00
C8	Uncertainty: deficit in general procedural knowledge (approach and method)	17	1.79	18	1.89	4	1.62	3	1.21	12	4.21	6	2.11	2	0.96	4	1.91
C15	Certainty: perceived possession of specific procedural knowledge that can be named or demonstrated	46	4.84	93	9.79	15	6.07	27	10.93	22	7.72	14	4.91	22	10.53	30	14.35
C22	Lack of specific procedural knowledge	6	0.63		0.00	5	2.02	4	1.62	7	2.46	4	1.40	5	2.39	2	0.96
C23	Perceived possession of relevant skills to solve the problem	312	32.84	215	22.63	85	34.41	51	20.65	61	21.40	54	18.95	61	29.19	39	18.66

Categories (C) and Super-Categories (SC)		OC - OC (50)				OC - R(13)				R - OC(15)				R - R(11)			
		PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%
C27	Perceived possession of test-taking skills	1	0.11	3	0.32	0	0.00	0	0.00	4	1.40	0	0.00	0	0.00	0	0.00
SC2	PROCEDURAL KNOWLEDGE SUBTOTALS	382	40.21	329	34.63	109	44.13	85	34.41	106	37.19	78	27.37	90	43.06	75	35.89
C1	Estimation of chance	0	0.00	3	0.32	0	0.00	0	0.00	1	0.35	0	0.00	0	0.00	0	0.00
C2	Subjective feeling of doubt and uncertainty	90	9.47	80	8.42	25	10.12	14	5.67	28	9.82	26	9.12	12	5.74	10	4.78
C3	Informed guess	40	4.21	39	4.11	18	7.29	11	4.45	24	8.42	14	4.91	33	15.79	11	5.26
C5	Lack of confidence	15	1.58	0	0.00	0	0.00	1	0.40	0	0.00	2	0.70	0	0.00	0	0.00
C6	Vague judgement: how answer looks in comparison with given responses	22	2.32	25	2.63	2	0.81	13	5.26	8	2.81	17	5.96	6	2.87	14	6.70
C10	Certainty: Subjective, vague feeling not based on any evidence	21	2.21	34	3.58	14	5.67	5	2.02	7	2.46	16	5.61	10	4.78	6	2.87
C11	Certainty: Subjective feeling based on perceived correctness of answer	21	2.21	76	8.00	11	4.45	11	4.45	15	5.26	14	4.91	9	4.31	9	4.31
C12	Estimation of Confidence in approach and answer	20	2.11	42	4.42	5	2.02	4	1.62	1	0.35	23	8.07	3	1.44	1	0.48

Categories (C) and Super-Categories (SC)		OC - OC (50)				OC - R(13)				R - OC(15)				R - R(11)			
		PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%
C13	How an answer compares with given multiple choice options	25	2.63	15	1.58	2	0.81	6	2.43	1	0.35	2	0.70	3	1.44	8	3.83
C14	Certainty: appeal that answer makes sense or is logical	31	3.26	20	2.11	8	3.24	13	5.26	9	3.16	4	1.40	4	1.91	5	2.39
C17	Unreflective, almost defensive evaluation	35	3.68	34	3.58	3	1.21	0	0.00	4	1.40	7	2.46	0	0.00	3	1.44
C18	Global feeling that answer is right	1	0.11	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
SC3	GLOBAL EVALUATION OF ANSWER SUBTOTALS	321	33.79	368	38.74	88	35.63	78	31.58	98	34.39	125	43.86	80	38.28	67	32.06
C4	Feeling of general inadequacy and unpreparedness	10	1.05	24	2.53	7	2.83	15	6.07	10	3.51	7	2.46	4	1.91	8	3.83
SC4	FEELING OF UNPREPAREDNESS SUBTOTALS	10	1.05	24	2.53	7	2.83	15	6.07	10	3.51	7	2.46	4	1.91	8	3.83

Categories (C) and Super-Categories (SC)		OC - OC (50)				OC - R(13)				R - OC(15)				R - R(11)			
		PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%	PRE	%	POST	%
C20	Deficits of the question	31	3.26	32	3.37	5	2.02	16	6.48	13	4.56	17	5.96	2	0.96	8	3.83
C24	Familiarity with question or concept	22	2.32	38	4.00	12	4.86	23	9.31	9	3.16	9	3.16	3	1.44	4	1.91
C25	Unfamiliarity with question or concept	19	2.00	4	0.42	8	3.24	4	1.62	8	2.81	1	0.35	3	1.44	1	0.48
SC5	EXTERNAL FACTORS SUBTOTALS	72	7.58	74	7.79	25	10.12	43	17.41	30	10.53	27	9.47	8	3.83	13	6.22
C16	Metacognitive Strategy	1	0.11	1	0.11	0	0.00	1	0.40	0	0.00	0	0.00	0	0.00	1	0.48
SC6	METACOGNITIVE STRATEGY SUBTOTALS	1	0.11	1	0.11	0	0.00	1	0.40	0	0.00	0	0.00	0	0.00	1	0.48
C21	Uncertainty: Inadequate mathematical skills	34	3.58	26	2.74	4	1.62	4	1.62	1	0.35	8	2.81	8	3.83	10	4.78
SC7	MATH SKILLS SUBTOTALS	34	3.58	26	2.74	4	1.62	4	1.62	1	0.35	8	2.81	8	3.83	10	4.78
C26	Lack of memory	34	3.58	9	0.95	2	0.81	1	0.40	11	3.86	3	1.05	7	3.35	4	1.91
SC8	LACK OF MEMORY SUBTOTALS	34	3.58	9	0.95	2	0.81	1	0.40	11	3.86	3	1.05	7	3.35	4	1.91
SC9	NO EXPLANATIONS	33	3.47	44	4.63	4	1.61	4	1.61	10	3.51	20	7.02	1	0.48	8	3.8
SC10	OTHERS	15	1.58	1	0.10	0	0	2	0.80	2	0.70	0	0	0	0	0	0
	TOTALS	950	100	950	100	247	100	247	100	285	100	285	100	209	100	209	100

Table 4.19 will subsequently be discussed in terms of trends observed in each performance evaluation subgroup. As mentioned before the discussion will focus only on responses with a prevalence of 5.00% or higher. In the discussion below all percentage values for response prevalence will be rounded off to one decimal place to simplify comparisons. In the interpretation of frequency data it will be helpful to bear in mind that some super-categories clearly reflect a subjective judgement (SC3 and SC4); some are rationally motivated (SC1, SC2, SC6, SC7 and SC8), whereas a single super-category seems to be based on both the test question as an object, feelings of not knowing due to unfamiliarity with the question or the features of the question itself, integral and cognitive feelings (SC5).

a. OC-OC subgroup

Responses whose incidence of occurrence prevails in the pretest are responses categorised as perceived possession of relevant skills to solve the problem (C23: 32.8%, e.g. “Because I still understand how, what is done when balancing a chemical equation”) and subjective feeling of doubt and uncertainty (C2: 9.5%, e.g. “Hope it is correct. Not 100% sure”, “Might be correct but still not sure”). Dominant responses in the posttest are observed in the same categories, i.e. C23 (22.6%) and C2 (8.4%), as well as two additional categories, i.e. certainty due to perceived possession of specific procedural knowledge that can be named or demonstrated (C15: 9.8%, e.g. “Because I balanced these equations practically and this is the answer I found”, “The amount of moles are right according to the ratio – $(8/10 \times 13.7)$: $(12/10 \times 13.7)$) and certainty due to subjective feeling based on perceived correctness of answer (C11: 8.0%, e.g. “It seems correct to me”).

When the focus is shifted to super-categories the results for the OC-OC subgroup indicate that judgements made on the basis of declarative knowledge (SC1) and those made based on external factors (SC5) increased marginally in the posttest (SC1 – Pre: 5.1%, Post: 7.9%; SC5 – Pre: 7.6%, Post: 7.8%). Examples of typical SC1 responses are “Because I know about limiting reactants” and “I just think so, but I know nothing about a mole”. The following statements are typical examples of SC5 responses: “Because the diagrams are a bit confusing”, “I have not encountered such a question before so I find it hard to answer this one”. The prevalence of objective judgments based on procedural knowledge (SC2) observed in responses like “I actually performed the calculations to find the exact answer” decreased from 40.2% in the pretest to 34.6% in the posttest and the global evaluation of answers observed in explanations such as “I chose 80% because I am not fully sure of the answer but I

believe it is so”, indicating subjective feeling of doubt and uncertainty, increased in the posttest as compared with the pretest (SC3, Pretest: 33.8%, Posttest: 38.7%).

b. OC-R subgroup

Dominant responses in the pretest are two rational judgements based on possession of procedural knowledge (C15: 6.1%, e.g. “I balanced the formulas and found out that my answer was correct”, “Because I did my calculations”) and possession of skills relevant to solve the problem (C23: 34.4%, e.g. “I think I can balance equations”, “I balanced the equation to make sure that what I have on the left-hand side is equal to what I have on the right hand side. One of the general rules in chemistry”), as well as a number of global judgements, i.e. a subjective feeling of doubt and uncertainty (C2: 10.1%, e.g. “Because I am not very sure of my answer”), judgements based on an informed guess (C3: 7.3%, e.g. “Took a calculated guess, but I think it is right”, “I am guessing and using logic”) and a subjective vague feeling not based on any evidence (C10: 5.7%, e.g. “I am sure and confident of my answer”). The dominant responses in the posttest are based on procedural knowledge (C15: 10.9%), perceived possession of skills (C23: 20.7%), external factors (C20: 6.5%, e.g. “The question was simple. I am guaranteed to get it right” and C24: 9.3%, e.g. “I have applied it to a lot of problems and that is what I do to balance so I trust my answer) and four subjective judgements for which the incidences of occurrence are just above the analysis threshold of 5% (C2: 5.7%, C6: 5.3%, e.g. “Seems like the right one but chemistry is never certain”, C14: 5.3%, e.g. “Seems logical” and C4: 6.1%, e.g. “Never really understood in class”).

The prevalence of judgements based on procedural knowledge decreased by almost 10% (SC2 – Pretest: 44.1%; Posttest: 34.4%), while that of subjective judgements remained stable. An example of a typical SC2 response is, “Because I did calculations to find the correct answer”. The prevalence of judgements based on a global evaluation of the answer (SC3 – Pretest: 35.6%; Posttest: 31.6%), observed in statements like “Just a feeling I have”, decreased by roughly the same margin as the increase documented for the feeling of unpreparedness (SC4 – Pretest: 2.8%; Posttest: 6.1%), observed in statements such as “We didn’t do this section in detail yet”. This subgroup showed the largest increase in the prevalence of SC5 responses, i.e. from 10.1% in the pretest to 17.4% in the posttest. Several examples are included for this super-category, External factors, in order to demonstrate the nuances of these statements: “Question not clearly understandable”, “Because we have done

such questions in the small group lecture and I understood”, “Did almost the same question this morning and I got it right”, “Quite sure about my answer, learnt it in high school as well as here” and “Answer was simple to find”. The increase in SC5 responses may indicate that after teaching the students in the OC-R subgroup, being the subgroup that showed the highest learning gain, were able to recognise and reveal when they were or were not familiar with the question, when they did or did not understand what was being asked or when the question was easy to solve. This awareness may have placed them in a position to better evaluate their performance during the posttest.

c. R-OC subgroup

Two categories of procedural knowledge emerged as dominant responses in the pretest, i.e. certainty based on procedural knowledge (C15: 7.7%, e.g. “Because Hg is about 12.5 times heavier than Oxygen, therefore $(100 \times 1250) \div 100 = \pm 1350$) and perceived possession of skills required to solve the problem (C23: 21.4%, e.g. “I have a good way of interpreting graphs”). The global judgements that emerged as dominant responses are a subjective feeling of doubt and uncertainty (C2: 9.8%, e.g. “I am not entirely sure about the answer”), reliance on an informed guess (C3: 8.4%, e.g. “This is a guess”), and certainty based on perceived correctness of answer (C11: 5.3%, e.g. “I am very sure it’s correct”) which was just above the analysis threshold of 5%. The dominant responses in the posttest were based on rational judgements (C23: 19.0%), a number of subjective judgements (C2: 9.1%; C6: 6.0%, e.g. “Looks right”; C10: 5.6%, e.g. “I am quite sure”; C12: 8.1%, e.g. “I am confident of my answer”), and external factors (C20: 6.0%, e.g. “I am not sure how to get it because I don’t have an equation”).

The prevalence of judgements based on procedural knowledge decreased (SC2 – Pre 37.2%; Post: 27.4%) and the global evaluation of answers increased by a similar margin (SC3 – Pre: 34.4%; Post: 43.9%). Examples of typical SC2 and SC3 responses are, “Balancing of that equation gives $4\text{Cu}_2\text{O}$ ” or “Can balance the equation” and “I am doubting” or “I am guessing” respectively. The prevalence of judgements based on external factors was fairly stable (SC5 – Pre: 10.5%, Post: 9.5%). Examples of typical SC5 responses for this subgroup are “Because two possible answers are present. They both have equal chance” and “There are just so many numbers to work with”.

d. R-R subgroup

Dominant responses were observed in the pretest for two categories of procedural knowledge, namely certainty based on the perceived possession of specific procedural knowledge (C15: 10.5%, “I balanced the equation and found out which answer is correct”) and perceived possession of skills required to solve the problem (C23: 29.2%, e.g. “I am very good at balancing equations and I know my answer is correct”), and two categories of global judgements, i.e. a subjective feeling of doubt and uncertainty (C2: 5.7%, e.g. “I am not sure if my answer is correct but it might be”) and reliance on an informed guess (C3: 15.8%, e.g. “I am not sure about the answer. Guessed.”). In general the same pattern of responses was observed in the posttest. The most noticeable difference was a sharp decline in the prevalence of C23 (18.7%) and C3 (5.3%).

The following shifts were observed in the super-categories of this subgroup. The prevalence of judgements based on declarative knowledge has more than doubled in the posttest (SC1 – Pre: 5.3%; Post: 11.0%). Examples of typical SC1 responses are “I know the definition” or “Because I know the law of conservation of mass and I also know how to balance an equation”. The large increase in SC1 responses was unique to the R-R subgroup. A small increase was also observed for judgements based on external factors (SC5 – Pre: 3.8%; Post: 6.2%), evident in responses such as “I am not familiar with reactions in diagrams” or “Because that is what I learned so far in my chemistry classes”. The prevalence of judgements based on procedural knowledge has decreased by 7% (SC2 – Pre: 43.1%; Post: 35.9%) and so did the prevalence of judgements based on a global evaluation of the answer (SC3 – Pre: 38.3%; Post: 32.1%). Examples of typical SC2 and SC3 are “I balanced the equation before looking at the given options and I found the answer I worked out” or “I balanced the equation first and therefore calculated the moles of reactants using the moles of the product” and “Because I have a feeling that it’s correct” respectively.

4.3.5.3 Discussion of qualitative results

Analysis of shifts in the prevalence of super-categories between the pre- and posttests across all subgroups could potentially provide information on the influence of teaching on the factors that each subgroup relies on in making confidence judgements (research question 3). The findings may also assist in the identification of specific patterns in metacognitive monitoring that were associated with higher or lower learning gain in stoichiometry. The OC-R subgroup showed the highest learning gain by a significant margin (49%), followed by

moderate learning gains demonstrated by the R-R and OC-OC groups (25% and 19%, respectively). The R-OC group did not achieve any learning gain at all (-1%). The difference in learning gain between the R-R and OC-OC subgroups was not significant whereas those between the R-OC and OC-R subgroups and all of the others were found to be (paragraph 4.3.3.1.7). From a teaching perspective it would be important to identify the metacognitive processes that are unique to the OC-R group suggesting a positive correlation with high learning gain, and those unique to the R-OC group which may not be conducive to learning.

The two super-categories that feature most prominently for all subgroups are SC2 which is rationally based and SC3 which is not. Overall, students in all four subgroups were less inclined to motivate their choice of confidence judgement ratings based on perceived possession of procedural knowledge (SC2) in the posttest. It seems that with exposure to teaching, students were more aware of what they do and do not know and were not too quick to claim the possession of procedural knowledge. However, mixed trends were found for SC3. Students who stayed or became realistic in their performance evaluation (R-R, OC-R) were less inclined to motivate their choice of confidence judgement ratings in the posttest in terms of vague, subjective feelings of certainty or uncertainty (SC3). An opposite trend was observed for the students who remained or became overconfident in their judgements of performance (OC-OC, R-OC). The largest increase in the prevalence of SC3 responses was recorded for the R-OC subgroup. This increase was almost twice that of the increase observed for OC-OC. This is significant, especially in the light of the poor performance of the R-OC subgroup.

Two other super-categories were also populated with responses by all subgroups but to a much lesser extent than SC2 and SC3, i.e. motivations based on declarative knowledge (SC1) and motivations based on external factors (SC5). Overall, all subgroups motivated some confidence choices on the presence or absence of declarative knowledge in the pretest (4 – 6%) and demonstrated a similar or higher prevalence for this motivation in the posttest (6 – 8%). The exception is the R-R subgroup where the prevalence of SC1 increased from 5% to 11% in the posttest. In the majority of responses coded under super-category 5, i.e. External factors, students were able to objectively identify the chemistry concept or the feature of the problem statement that gave rise to their choice of confidence indicator. These responses suggest a more differentiated level of knowing or not knowing as compared with a general feeling of competence or inadequacy. Also included in SC5 are statements which

may reflect cognitive feelings, i.e. familiarity or unfamiliarity with the question (C24). The increase of this factor from the pretest to the posttest for the OC-R and the R-R subgroups, as compared with a fairly stable presence for the OC-OC and the R-OC subgroups should be noted. The largest increase in SC5 was observed for the OC-R subgroup which is also the group that achieved the highest learning gain.

The picture that emerges when all of these results for shifts in the prevalence of major response super-categories are analysed together is the following: Accuracy in the evaluation of posttest performance was associated with both a reduction in the prevalence of vague subjective judgments and with higher performance gain. Inaccuracy in self-evaluation in the pretest did not seem to hamper learning for both the OC-OC and OC-R subgroups. Instead, an increase in the tendency to base metacognitive monitoring on vague global judgments of performance in the posttest was associated with reduced accuracy of self-evaluation and lower learning gain.

The results reported in Table 4.19 and the trends that were described in the foregoing paragraphs can be interpreted by drawing on the work of Nelson and Narens (1990). These authors have formulated a theoretical framework for the metacognitive monitoring that occurs when students are confronted with a learning task or with answering a question in a test. This theoretical framework was described in detail in Chapter 2, paragraph 2.2. Our literature review (paragraph 2.2) revealed that during the construction of a metacognitive judgement students search their metamemory for evidence that a studied item will be successfully acquired, retained or retrieved and then express the judgement based on the amount of evidence found. Koriat (2000) labels judgements made on the basis of information in metamemory as information-based metacognitive judgements. Judgements made on the basis of experience or how an individual feels at the time of making the judgement he labels experience-based metacognitive judgements. Therefore subjective responses given as motivations for the confidence judgements made in the test, i.e. responses listed under SC3 labelled “Global evaluation of answer”, and SC4 – Feeling of unpreparedness, indicate that those judgements were more experience- or feeling-based than information-based. Responses listed in the Declarative knowledge, Procedural knowledge, Metacognitive strategies, Mathematical skills and Lack of memory super-categories (SC1, SC2, SC6, SC7 and SC8) imply that the confidence judgements were informed by content information retrieved from

metamemory. Responses listed under SC5 labelled “External factors” consisted of a mixture of judgements which were experience and information-based.

It seems that during their construction of a confidence judgement students try to search and retrieve information which they can use as evidence that a task will be successfully executed or a question accurately answered. However, in the absence of such information, students may resort to feelings as reference. The trends observed in our sample showed that students who, in spite of teaching, still over-estimated their performance, relied more heavily on feelings whilst answering the test questions rather than on their possession or lack of information to guide their choice of confidence judgement rating. This probably explains why they remained (OC-OC subgroup) or became (R-OC subgroup) biased in their performance evaluation even after teaching.

Our literature review enabled us to identify four types of feelings people tended to rely on when they construct their metacognitive judgements. Moreover three metamemory hypotheses used to explain the construct of “feelings of knowing” were identified, namely the cue familiarity hypothesis, the accessibility hypothesis and the competition hypothesis. The four types of feelings include Feelings of knowing (FOK), Feelings of not knowing (FOnK), Affective feelings and Cognitive feelings. Judgements made on the basis of cognitive feelings may seem similar to information-based metacognitive judgements because both involve the retrieval of information. However judgements made on the basis of cognitive feelings rely on the ease-of-retrieval as source of information rather than directly and solely on content information whereas information-based judgements rely on retrieved content information as a source. Next I will discuss each type of feeling and show how it emanates from our qualitative data.

i. Feelings of knowing

Koriat (2000) defines a FOK as a feeling people may refer to as intuitive, a hunch or “just knowing” and a feeling that requires no justification. This definition helps explain the kind of responses that we have encountered and that are listed under categories labelled C1 (Estimation of chance, e.g. “Because there is 60% that other answers are right”), C3 (Informed guess, e.g. “ I am guessing”), C10 (Certainty due to a subjective vague feeling not based on any evidence, e.g. “sure of answer”), C11 (Certainty due to a subjective feeling based on perceived correctness of answer”), C12 (Estimation of confidence in approach and

answer, e.g. “confident with work”), C14 (Certainty due to an appeal that answer makes sense or is logical, e.g. “Answer makes sense”), C17 (Unreflective, almost defensive evaluation, e.g. “Just know”) and C18 (Global feeling that answer is right, e.g. “this is how I feel”) in Table 4.17.

ii. Feelings of not knowing

According to Jing *et al.* (2003), FOnK are accurate negative FOK predictions that accurately anticipate ‘not knowing’. The accessibility hypothesis posits that people may base their JOKs on retrieved information. When little or no information is retrieved, people prefer a judgement of “I don’t know”, as was observed in responses listed under C26 (Lack of memory e.g. “Forgot how to work out”). The cue familiarity hypothesis on the other hand suggests that if the information in the question asked seems unfamiliar people may be quick to conclude that the information is not present in their metamemory, like responses observed under C25 (Unfamiliarity with questions or concept, e.g. “never seen such question before”).

iii. Affective feelings

Greifeneder *et al.* (2010) define affective feelings as experiences that may or may not be linked to an object. These feelings are more sensitive to an individual’s moods and attitude at the time of constructing a JOK. Affective feelings may be incidental or integral to the object such as a test being taken. Incidental feelings are elicited by an external source rather than the target being judged and integral feelings are elicited by features of a target object whether the features are real, perceived or imagined. In the context of our study, incidental feelings may be understood as being elicited by how a student felt while he/she was answering a test question and integral feelings, as feelings they might have been elicited by real, perceived or imagined features of the test such as level of difficulty, format of distractors in a multiple choice test or deficiencies in the test question. Responses listed under categories C2 (Subjective feeling of doubt and uncertainty, e.g. “Doubtful), C5 (Lack of confidence, e.g. “don’t feel confident”) and C18 (Global feeling that answer is right, e.g. “this is how I feel”) are consistent with the definition of incidental feelings. Responses listed under C6 (Vague judgement of how answer looks in comparison with given responses, e.g. “it’s right”), C13 (How own answer compares with given multiple choice options, e.g. “closest value to own answer”) and C20 (Deficits of the question, e.g. “vague question”) are consistent with the definition of integral feelings.

iv. Cognitive feelings

According to Greifeneder *et al.* (2003), a student may interpret familiarity with the content information as an indication of the ease with which the target information may be retrieved. Based on the cue familiarity hypothesis an individual may find the information familiar and as a result get the feeling that he/she knows the work. Therefore the individual would more likely to judge that he/she knows the answer when he/she is familiar with the information. This explanation is consistent with responses we observed and labelled as C24 (Familiarity with question or concept, e.g. “done in high school”). According to the competition hypothesis, the problem with relying on familiarity is that individuals may mistakenly assume familiarity with information because of its similarity with the target information resulting in inaccurate metacognitive judgements.

4.3.5.4 Conclusion

Intuitively a science educator would expect students to apply logical reasoning when required to motivate their confidence in the correctness of an answer in a science test. The sample in our study consisted of weak, under-prepared students who were found to exhibit high levels of confidence when asked to evaluate their performance. The students in our sample were expected to evaluate their performance in a test on a content topic which is difficult and which lends itself to misconceptions. Misconceptions are structures strongly held by students that are different from the accepted understanding by experts in the field (Hasan *et al.*, 1999). Hasan *et al.* (1999) interpreted highly exaggerated confidence levels as an indication of the presence of strong misconceptions which cause students to be confident of their answers even when these are incorrect. The assumption made here was that when students were very confident about their understanding and their confidence was unjustified because of flawed understanding, any one of a number of rationally based misconceptions listed in Chapter 2, paragraph 2.6.3, would be revealed in the explanations that they provided to justify their level of confidence in the correctness of their answers.

Solving stoichiometry problems usually requires predominantly procedural knowledge. Stoichiometry has a very minimal component of memory and recall problems. It is for this specific topic necessary to go beyond recall of declarative knowledge. In fact it was not surprising to find the prevalence of responses in the SC2 (Declarative knowledge) to be so small across all subgroups in both the pre- and posttest. The prevalence of this super-category stayed at a level of 5 or 7% and only for the R-R subgroup did it ever attain a level of 11%. In

Chapter 2, paragraph 2.6.1, a typical stoichiometry problem was cited to show that both formal reasoning and the use of multistep mathematical operations were required to solve the problem. When stoichiometry problems include pictorial representations at the atomic or molecular level in their problem statements, students would also have to demonstrate the successful manipulation of the submicro and symbolic levels of thinking as described by Johnstone (1991) (Chapter 2, paragraph 2.6.2). Solving such stoichiometry problems will require students to demonstrate representational competence and would challenge them to reveal their conceptual understanding. Furthermore it was expected that the nature of the topic, which is explained in detail in Chapter 2, paragraph 2.6, and hence the type of questions on the topic would elicit rational rather than subjective responses. We expected that students would base their choice of confidence indicators on the possession or lack of formal reasoning, mathematical skills, conceptual understanding, declarative knowledge, or the procedural knowledge that may be required to solve the problem. Instead, our results revealed that even in a science test in the specific format that we have used, students believed an answer to be correct based on feelings rather than on rationally motivated judgements. A super-category which clearly reflects a subjective judgement, SC3, (Global evaluation of answer) constituted a substantial 30 to 40% of all the confidence judgements in the pre- and posttest responses for all the subgroups and this was unexpected (OC-OC – Pre: 33.8%, Post: 38.7%; OC-R – Pre: 35.6, Post: 31.6%; R-OC – Pre: 34.4%, Post: 43.9%; R-R – Pre: 38.3%, Post: 32.1%).

With the help of relevant literature we were able to understand that responses such as “I believe that I am correct” may be an indication of judgements made by students on the basis of feelings of knowing characterised by an intuitive feeling or a hunch which requires no justification. Feelings of not knowing observed in statements like “I don’t know”, may be observed when little or no information could be retrieved. The danger here is that according to the familiarity hypothesis, people may be quick to make an “I don’t know” judgement when they find the information in the question unfamiliar. This may be the instance when a student would state that he/she does not know the answer to a question because he/she has never come across a stoichiometry question with a pictorial presentation of atoms or molecules as part of the problem statement and then make a statement like “I am not familiar with the reactions in diagrams”. Incidental feeling-based judgements were observed in responses such as “this is how I feel” or “I am not confident at all ‘cause I have never done this. I don’t even know how to calculate the answer”. Integral feeling-based judgements

could be observed in responses such as “confused by the question” or “Because all or some of the options are confusing, made me doubt my answer”.

To conclude the interpretation of qualitative results provided another basis for comparing students in the subgroups and it confirmed the finding that we were dealing with four distinct groups with unique properties. The qualitative results assisted in the identification of the factors, whether rational or subjective, that students relied on in making confidence judgements. Observation and analysis of shifts in the prevalence of response categories across all subgroups between the pre- and posttests provided valuable information on changes in metacognitive monitoring after the chemistry content had been taught. These findings complement those derived from the analysis of quantitative data by providing insight into the metacognitive processes associated with performance evaluation, especially those that are associated with higher learning gain.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

CONTENTS	PAGE
5.1 Introduction	130
5.2 Overview of the study	130
5.3 Summary of the findings	133
5.4 Educational implications of findings	143
5.5 Contributions and significance of the study	144
5.6 Limitations of the study	146
5.7 Recommendations	149
5.8 Areas for further research	150

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 INTRODUCTION

The chapter commences with the presentation of the overview of the study followed by a summary of the findings with respect to the research questions. The implications that the findings may have for teaching are presented next. Further highlights are the contributions and significance of the study. The limitations of the study with respect to relevant strengths and weaknesses are delineated. Finally recommendations of the research are detailed and directions for further research close the chapter.

5.2 OVERVIEW OF THE STUDY

To investigate bias in performance evaluation in a group of students in the University of Pretoria's BSc Four-year programme (BFYP) a case study following a mixed methodological approach was conducted over a period of three years. The embedded experimental design largely based on a quantitative approach with the qualitative approach taking a secondary, supplementary role within the overall design was followed. A detailed description of this design was presented in paragraph 3.3.1. The sample of our study ($N = 91$) comprised 35 males and 55 females with a median age of 19 years. Students in our sample were admitted to the BFYP because they had failed to meet mathematics and science entry requirements required to gain access into mainstream university science courses. As a result, the sample comprised students who were weak or under-prepared or both. The teaching strategies and activities in the BFYP described in paragraph 1.2.1 are strategically employed to address the academic under-preparedness for tertiary studies of such students. In previous studies (Hasan *et. al.*, 1999; Ochse, 2003; Potgieter *et. al.*, 2007) poor-performing students had been found to exhibit high levels of overconfidence when evaluating their performance. Nowel and Alston (2007) suggested that overly optimistic assessments of how much one knows and understands might lead one to study less than if one had accurate perception. Mastering the skill of metacognitive monitoring on the other hand, may result in effective regulation of self-paced study, which is necessary in a tertiary environment where an independent approach to studying is required (Dunlosky *et al.*, 2005). It became important therefore, to establish whether students in the BFYP also exhibited high levels of confidence which were not justified by good performance. In our attempt to determine the presence of bias in performance evaluation, we chose a difficult topic in the first-year chemistry syllabus

anticipating that a more challenging topic would be instrumental in differentiating better between students who were accurate in their performance evaluation and those who were not. Stoichiometry was chosen among other topics which were found in literature as topics that gave first-year chemistry students the most difficulty (Huddle & Pillay, 1996). The following research questions were therefore formulated to guide us in the attempt to investigate bias in self-evaluation in our sample before and after instruction as well as the factors underlying bias in performance evaluation.

Research question 1: How accurately do BFYP students evaluate their performance in a stoichiometry test?

Research question 2: What is the influence of teaching of stoichiometry in the BSc Four-year programme on performance and accuracy of performance evaluation?

Research question 3: What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?

Research question 4: What is the relationship between bias in performance evaluation and self-enhancement, self-protection and gender?

Quantitative and qualitative data were collected as an attempt to answer the four research questions. Two data collection instruments had to be developed and subjected to several measures, most of which were statistical, to gather enough evidence to ensure that the results and conclusions based on the data obtained through them would be valid and reliable. A 20-item stoichiometry test instrument and a 19-item questionnaire were developed for data collection purposes. The stoichiometry test instrument was used as pre- and posttest to enable the investigation of the influence of teaching on both performance and accuracy in performance evaluation (research question 2). The questionnaire was used to investigate the association between bias in performance evaluation and the tendency to self-enhance or self-protect (research question 4) and hence was only used once to collect data, i.e. during the pretest.

Educators (three high school teachers and one first year university lecturer) were consulted and requested to comment on the appropriateness of the chemistry test questions. Among other comments and recommendations made by the educators, question 2 was identified as

ambiguous, but was however retained until sufficient statistical evidence could be gathered to determine its possible removal from the instrument. Upon calculation of item-total correlations, item 2 was found to have a very weak item-total correlation in the pretest and a negative correlation in the posttest. Exclusion of item 2 during statistical analysis improved the pretest Cronbach's alpha marginally from 0.63 to 0.64 and that of the posttest from 0.67 to 0.69, which served as confirmation of the concerns raised by the educators. Item 2 was therefore omitted from the data set in all the subsequent analyses.

In addition to the chemistry test questions the instrument had additional two tiers per item, resulting in a three-tier test instrument. Quantitative data in the form of confidence judgement ratings were collected in the second tier. Qualitative data in the form of free-response explanations for the justification of the choice of confidence judgement rating made in the second tier were collected in the third tier. Quantitative data collected by means of the first and second tiers of the instrument were used to determine the accuracy with which students evaluated their performance in the pre- and posttests (research question 1). Paragraph 3.10.1 provided a detailed description of the procedure used to determine accuracy of performance evaluation as well as how these results were used to categorise students as either overconfident, realistic or underconfident. The free-response explanations provided by the students in the third tier constituted qualitative data and these were used to investigate factors underlying bias in performance evaluation (research question 3). Free-responses provided after instruction were analysed to determine the influence of teaching on the factors underlying bias in performance evaluation (research question 3). Coding of these responses led to the formulation of 27 response categories and eight super-categories. The prevalence or categories were computed both for the pre- and posttest responses. The shifts in the prevalence of responses between the pre- and posttests were noted and analysed.

Factor analysis of the questionnaire instrument items revealed that the items were measuring three as opposed to two constructs, i.e. two forms of self-enhancement and one of self-protection. Correlation coefficients less than 0.25 indicated a poor correlation between the three factors which was confirmation that these were three discrete factors measuring different constructs. Five items loaded strongly onto the first form of self-enhancement, labeled SEa. Items in this factor expressed the desire to portray oneself as a hardworking, clever student. Five items loaded strongly onto the second form of self-enhancement, labelled SEb, and items in this factor expressed the need to convey a good impression about oneself to

others. Nine items which expressed the need to protect one's academic self-worth by attributing failure to external factors loaded strongly onto the self-protection factor, labeled SP. Details of results obtained through factor analysis were provided in paragraph 4.2.2.1. Cronbach's alpha coefficients of 0.66, 0.74 and 0.60 were computed for the items in the SEa, SEb and SP factors respectively, providing sufficient statistical evidence that data and results obtained by means of these items were reliable and valid.

5.3 SUMMARY OF THE FINDINGS

The previous paragraphs served as a presentation of the overview of the study. The discussion from here onwards will focus on the results obtained as well as how these results were used to answer the four research questions.

Research question 1:

How accurately do BFYP students evaluate their performance in a stoichiometry test?

Taking into consideration the difficulty of the topic, the level of preparedness of the students in our sample and the format of the test (multiple-choice), an acceptable margin of error had to be defined. Setting this margin at 3 out of 19 answers judged incorrectly in terms of correctness of answer translated into an error of 15.8%. The choice of confidence indicator was interpreted as an indication of expectation that the chosen answer would be correct, i.e. an indication of expected performance. Table 5.1 below shows the categorization of students based on the accuracy with which they evaluated their performance in the pre- and posttests. The performance evaluation of the majority of students was found to be inaccurate in both the pre- and posttests, i.e. 69% of the students were found to be overconfident in the pretest and 71% in the posttests. This means that approximately 70% overestimated their actual performance by more than 15.8% as implicated by the confidence that they expressed in the correctness of their answers in the test. Less than a third of the students in our sample were able to evaluate their performance within the margin of 15.8% error, in the pre- and posttests, i.e. 31% in the pretest and 26% in the posttest.

Table 5.1: Summary of categorization of students in terms of accuracy of performance evaluation in the pre- and posttest

	PRETEST				POSTTEST		
	Quantity	Male	Female		Quantity	Male	Female
OC	63* (69%)	25	37	OC	65* (71%)	25	39
R	28 (31%)	10	18	R	24 (26%)	9	15
UC	0	0	0	UC	2 (2%)	1	1
Totals	91	35*	55*	Totals	91	35*	55*

* One record without gender information omitted.

Research question 2:

What is the influence of teaching of stoichiometry in the BSc Four-year programme on performance and accuracy of performance evaluation?

This question will be answered based on results that were reported in Chapter 4, paragraphs 4.3.1, 4.3.2, 4.3.3 and 4.3.3.1. Table 5.2 below shows a summary of the results in terms of performance and confidence scores obtained in the pre- and posttests. P-values of 0.00 indicated a statistically significant difference between the pre- and posttest performance and between pre- and posttest average confidence scores.

Table 5.2: Summary of student performance and average confidence scores in the pre- and posttest

	Performance (test scores)		Average confidence scores	
	Pretest	Posttest	Pretest	Posttest
Sample size	91	91	91	91
Mean	7.0	9.6	63.0	75.7
STD deviation	2.9	3.4	17.4	13.5
Minimum	2 (Max = 19)	3 (Max = 19)	16.3	40.5
Maximum	15 (Max = 19)	18 (Max = 19)	94.7	99.5

In the posttest students made significant advances in terms of mastery of solving stoichiometry problems but they did not improve in terms of their accuracy of performance evaluation. It is important to note that the posttest results presented in Table 5.2 are not directly attributed to the teaching of stoichiometry alone. Other factors such as the commitment of students to work harder may have played a role. The mean performance improved from 7.0 out of 19 (37%) in the pretest to 9.6 out of 19 (51%) in the posttest, while the mean of average confidence scores increased from 63 % in the pretest to 76 % in the posttest. The improvement in average performance in the chemistry test was 14%, but this was accompanied by an increase of 13% in average estimated performance. These results are disconcerting because based on suggestions in the literature (Kruger & Dunning, 1999) we assumed that the acquisition of content knowledge would expose students to the processing demands of the topic and therefore play a role in guiding them to make more realistic and accurate evaluations of their performance. The increase in average confidence may be due to the fact that in the period between administration of the pre- and posttests a team of lecturers taught stoichiometry without making any explicit attempts towards addressing accuracy of performance evaluation. Kruger and Dunning (1999) suggested that increasing the content knowledge of poor performers would improve their metacognitive ability in terms of the accuracy with which they evaluate their performance. It was upon this suggestion that it was anticipated that with the teaching of stoichiometry, the inaccuracy of performance evaluation would be corrected. Figure 4.3 in Chapter 4, which is a depiction of how students in the performance-based quartile rankings compared in terms of the accuracy with which they evaluated their performance in the posttests, showed that even when students were allowed the scope to make errors in their judgement by a margin of 15.8%, only the top performing students were able to evaluate their performance in the posttests within that margin.

The global view of results as shown in Table 5.2 obscured the finer details about students who were able to show an improvement in terms of accuracy in performance evaluation after instruction and those whose ability to do so may have deteriorated. It became necessary to probe the data further and identify patterns in how students shifted in their performance evaluation subgroups after the teaching and learning experience. The two-way frequency table shown in Chapter 4, Table 4.10, assisted in defining five subgroups on the basis of accuracy of performance evaluation in the pre- and posttests. The five groups, labelled first by their pretest category and then their posttest category, were the OC-OC, OC-R, R-R,

R-OC and the R-UC subgroups. Table 5.3 shows the five subgroups, their codes and the number of students in each subgroup.

Table 5.3: Five pre-post performance evaluation subgroups and the number of students in each subgroup

Performance evaluation subgroups	Number of students in each subgroup
OC-OC	50
OC-R	13
R-R	11
R-OC	15
R-UC	2

As mentioned in Chapter 4, paragraph 4.3.3.1, a large percentage of the 63 students who were overconfident in the pretest showed no improvement in their accuracy of performance evaluation, i.e. 50 remained overconfident (OC-OC subgroup) and only 13 were able to make accurate judgements in the posttest (OC-R subgroup). In the group of students who were realistic in their pretest performance evaluation (28 students), less than half remained realistic in their judgement in the posttest (11 students: R-R subgroup) and 15 became overconfident (R-OC subgroup). Two of the 28 students who were realistic in their pre-test performance evaluation became underconfident in the posttest (R-UC subgroup). None of the students in our sample were underconfident in the pretest.

For the convenience of the reader Table 4.11 in Chapter 4 is presented as Table 5.4 in this chapter. Table 5.4 is an overview of how the four subgroups compared in terms of their performance in the pre- and posttests as well as their gain from the teaching and learning experience. The R-UC subgroup is not represented in Table 5.4 because of its small size.

Having established that the four subgroups differed in terms of the accuracy with which they evaluated their performance in the pre- and posttests, we wanted to determine whether accuracy of performance evaluation was associated with higher learning gain and with the better end of semester performance. Such results would indicate whether accuracy of performance evaluation as a metacognitive skill is a desired attribute for the learning of

chemistry. For this purpose it was important to determine whether the four subgroups (OC-OC, OC-R, R-R and R-OC) were comparable in terms of ability and prior knowledge in stoichiometry, as judged by their performance in the first semester module, CMY 133, and their pretest performance respectively. If not, then an argument could be made that some subgroups were predisposed towards better performance and higher learning gain because of higher ability or a stronger foundation in chemistry.

Table 5.4: Pre- and posttest performance data according to performance evaluation subgroups

	POST OC		POST R	
PRE OC	Number	50	Number	13
	Av. Pretest performance (%)	33	Av. Pretest performance (%)	38
	Av. Posttest performance (%)	45	Av. Posttest performance (%)	68
	% Pass Pretest	10	% Pass Pretest	23
	% Pass Posttest	40	% Pass Posttest	77
	Av. Performance Gain (%)	19	Av. Performance Gain (%)	49
PRE R	Number	15	Number	11
	Av. Pretest performance (%)	41	Av. Pretest performance (%)	45
	Av. Posttest performance (%)	43	Av. Posttest performance (%)	61
	% Pass Pretest	27	% Pass Pretest	36
	% Pass Posttest	27	% Pass Posttest	91
	Av. Performance Gain (%)	-1	Av. Performance Gain (%)	25

The results indicated that there was no statistical difference between pretest performances of the subgroups and a marginal, but significant difference between their performance in CMY 133. The CMY 133 performance of the OC-OC subgroup was lower than that of OC-R, but there was no significant difference in all other pair-wise comparisons. The four subgroups could therefore be assumed to be comparable in terms of prior knowledge in stoichiometry based on pretest performance, but the OC-R subgroup seemed to have been more able or somewhat better prepared than the OC-OC subgroup for the challenges of the content.

The results presented in Table 5.4 above were analysed to determine whether differences were statistically significant. It was established that the students in the four subgroups differed significantly in terms of performance in the posttest, their pre- and posttest average

confidence scores and in performance gain. A significant difference was not found with regard to performance in the CMY 143 end of semester examination. These findings confirmed that we were dealing with four discrete groups with different characteristics. The four subgroups had a comparable level of preknowledge as judged by their pretest performance, but they differed significantly not only in their accuracy of performance evaluation, but also in terms of the learning gains demonstrated in posttest performance after having been taught the difficult topic of stoichiometry. The OC-R subgroup achieved the highest learning gain by a significant margin. Moderate learning gains were demonstrated by the R-R and OC-OC subgroups and the R-OC subgroup did not achieve any learning gain at all.

It was surprising that the respective learning gains achieved by the four subgroups in stoichiometry were not found to be predictive for end-of-semester performance in CMY 143. As mentioned in Chapter 4, paragraph 4.3.3.1.6, stoichiometry is an important, but minor component of the CMY 143 syllabus. It requires analytical reasoning as well as mathematical skills. Students who struggled to master stoichiometry, which is a difficult topic, could perform well in other topics.

Research question 3:

What are the factors that students rely on when making performance evaluations and what shifts, in terms of reliance on these factors, are observed after the teaching of stoichiometry?

This question will be answered based on qualitative results that were reported in Chapter 4, paragraph 4.3.5. Qualitative data were carefully analysed and coded as described in paragraph 4.3.5. By means of this process twenty-seven response categories were identified which were grouped together in eight super-categories (Table 4.19). Some of these super-categories clearly reflect a subjective judgement (SC3 and SC4); some seem to be objective or rationally motivated (SC1, SC2, SC6, SC7 and SC8), while the basis of a single super-category consisted of a combination of information-based and feeling-based motivations (SC5). The two super-categories that featured most prominently for all subgroups in both the pre-test and the post-test were “Procedural knowledge” (SC2), which motivated the choice of confidence indicator in terms of the possession or lack of possession of specific procedural knowledge or skills, and “Global evaluation of answer” (SC3), which mainly

reflected vague, subjective feelings or judgments. Less common, but still important, were choices motivated by perceived possession or lack of declarative knowledge (SC1, “Declarative knowledge”) and “External factors” (SC5), which ascribed confidence choices to perceived deficits in the test question or familiarity or lack of familiarity with the question. The prevalence of motivations in the other six categories was too low (less than 5%) to make a meaningful contribution towards the understanding of the metacognitive judgments that students made in this study.

It is clear from the discussion earlier that five distinct subgroups emerged from the analysis of accuracy of performance evaluation. Not only did these subgroups differ in their pre-post self-evaluation, but also in the learning gain that was demonstrated after the teaching of stoichiometry. The fact that one subgroup achieved a learning gain significantly higher than the rest and one subgroup did not achieve any gain at all, highlighted the need for information that could assist in explaining the strengths and deficiencies in metacognitive skills that gave rise to such a difference. The qualitative data proved to be a rich resource for this purpose. Qualitative data were analysed and interpreted according to subgroups and will be discussed as such.

As expected, different patterns of results were observed for the four subgroups. The students who remained overconfident, i.e. the OC-OC subgroup, predominantly used a global evaluation of answers (SC3: 33.8%) as well as procedural knowledge (SC2: 40.2%) to motivate their choice of confidence, with other types of explanations appearing considerably less often in the pretest. The same pattern of responses was observed for this subgroup in the posttest, however with a shift of ca. 5% away from claiming procedural knowledge (SC2: 34.6%) towards a global evaluation of answers (SC3: 38.7%).

The pattern observed for students in the R-OC subgroup was very similar to that of the OC-OC group. They also predominantly reported motivations based on global evaluation of answers (SC3: 34.4%) and procedural knowledge (SC2: 37.2%) in the pretest, with other types of explanations appearing less often. However, the shift away from SC2 towards SC3 in the posttest was more pronounced. The decrease in how frequently they motivated their choice of confidence indicator in terms of possession or deficiency of procedural knowledge (SC2: 27.4%) matched the increase in the frequency of motivations based on subjective judgements (SC3: 43.9%), but the shift was twice the magnitude found for the OC-OC

subgroup. This finding is significant since this was the subgroup that was most vulnerable in the sense that they did not demonstrate any learning gain in stoichiometry (Table 5.4).

In the pretest students in the R-R subgroup predominantly reported motivations based on procedural knowledge (SC2: 43.1%) and the global evaluation of answers (SC3: 38.3%), with other types of explanations appearing less often. This subgroup was exceptional in the extent to which they indicated that their answer was an “informed guess” (C3: 15.8%), a motivation that featured prominently only in the pretest. It seems that admitting that they did not know the work and that they had only guessed the answer, proved easier for this subgroup of students than for other subgroups where this category appeared less frequently (OC-OC: 4.2%, OC-R: 7.3%, R-OC: 9.8%). The decrease in the prevalence of motivations based on procedural knowledge (SC2: 35.9%) was similar to the decrease in the frequency of motivations based on subjective judgements (SC3: 32.1%) in the posttest. The more than twofold increase in prevalence of judgements based on declarative knowledge that was observed in the posttest (SC1 – Pre: 5.3%; Post: 11.0%) was unique to the R-R subgroup.

In the pretest students in the OC-R subgroup predominantly reported motivations based on procedural knowledge (SC2: 44.1%) and the global evaluation of answers (SC3: 35.6%), with other types of explanations appearing less often. The prevalence of judgements based on procedural knowledge decreased by almost 10% (SC2 – Pretest: 44.1%; Posttest: 34.4%), whereas that of judgements based on a global evaluation of the answer decreased by 4% (SC3 – Pretest: 35.6%; Posttest: 31.6%). This subgroup showed the largest increase in the prevalence of SC5 responses, i.e. from 10.1% in the pretest to 17.4% in the posttest.

There were two unique features of the posttest responses of the OC-R subgroup which will be explored further in an attempt to contribute to a better understanding of the metacognitive processes revealed by this group of students. The first unique feature is their inclination to admit when they felt unprepared in the posttest (SC4, 6.1%), which represents the highest incidence of SC4 responses recorded for any of the subgroups. This was evident in statements like “Never really understood in class” or “We did not do this section in detail yet” or “I don’t know the answer”. Secondly, their inclination to make judgements based on external factors (SC5) was high in the pretest and it almost doubled in the posttest. Several examples are included for this super-category in order to demonstrate the nuances of these statements: “Question not clearly understandable”, “Because we have done such questions in the small

group lecture and I understood”, “Did almost the same question this morning and I got it right”, “Quite sure about my answer, learnt it in high school as well as here” and “Answer was simple to find”. Considering that the students in this subgroup showed the highest learning gain, the type of SC4 and SC5 responses provided by this subgroup suggest that they have acquired a more differentiated knowledge of the topic during teaching. The increase in SC4 and SC5 responses may indicate that after teaching the students in the OC-R subgroup improved in their ability to recognise and reveal when they were or were not familiar with the question, when they did or did not understand what was being asked or when the question was easy to solve. This awareness may have enabled them to apply more effective metacognitive monitoring during the teaching and learning of stoichiometry, resulting in a mastery of the content that was superior to that of all of the other subgroups.

The two super-categories that feature most prominently for all subgroups are SC2 which is rationally based and SC3 which is not. Overall, students in all four subgroups were less inclined to motivate their choice of confidence judgement ratings based on perceived possession of procedural knowledge (SC2) in the posttest. It seems that with exposure to teaching, students were more aware of what they did and did not know and were not too quick to claim the possession of procedural knowledge. However mixed trends were found for SC3. Students who stayed or became realistic in their performance evaluation (R-R, OC-R) were less inclined to motivate their choice of confidence judgement ratings in the posttest in terms of vague, subjective feelings of certainty or uncertainty (SC3). An opposite trend was observed for the students who remained or became overconfident in their judgements of performance (OC-OC, R-OC). The largest increase in the prevalence of SC3 responses was recorded for the R-OC subgroup. A sharp increase in the motivations based on external factors was seen for the OC-R and the R-R subgroups as compared with a fairly stable presence for the OC-OC and the R-OC subgroups.

Research question 4:

What is the relationship between bias in performance evaluation and self-enhancement, self-protection and gender?

Research question 4 was concerned with the determination of the relationship that exists between bias in performance evaluation and several motivational factors. Among the factors

listed in the literature which could potentially be associated with such a bias, we have chosen to investigate self-enhancement, self-protection and gender as possible factors.

Research question four was explored for the sample as a whole as well as for the four subgroups as separate entities. In terms of the results obtained for the sample as a whole, the relationship between bias in performance evaluation and the first form of self-enhancement (SEa) was found to be weak, positive but insignificant. The relationship between inaccuracy in performance evaluation and the second form of self-enhancement (SEb) was also weak, negative but insignificant. The relationship between bias in performance evaluation and self-protection was negative, weak and insignificant. P-values greater than 0.05 in the pre- and posttests indicated that males and females were not significantly different in their bias in performance evaluation. When research question 4 was explored for the four subgroups, the statistical analysis results showed that the four subgroups did not significantly differ in terms of any of the three motivational factors, namely SEa, SEb and SP. The tendency by the four performance evaluation subgroups to self-enhance or self-protect was not found to be statistically different; hence no inferences could be made from the results.

Students in the OC-OC and R-OC subgroups performed poorly in the posttest and fewer than 50% of them passed the pre- and posttests. According to Gramzow *et al.* (2003), exaggerated performance evaluation by students with low grades is often motivated by the tendency to self-protect while exaggerated performance evaluation by students with high grades is often self-enhancement motivated. Taking students in the OC-OC and R-OC subgroups as poor performers and the students in the R-R and OC-R subgroups as top performers based on the average of their posttest scores and pass rates the findings of Gramzow *et al.* (2003) could not be confirmed by the results obtained for this study. While it must still be established in future studies whether significant differences are observed when a larger sample is used, it is expected that the findings of such a study will not be different. Gramzow and co-workers (2003) used as sample of undergraduate psychology students at a university in the USA for their study while our study was conducted among science students who were weak or under-prepared and who were primarily second language English speakers in an African context. It is likely that our sample of science students was less self-aware and less experienced in reflection and self-analysis than the psychology students who routinely engage in such practices as part of their training. In addition, the influence of cultural factors

on the willingness of students to disclose feelings of self-doubt or inadequacy is likely to be different between the two samples.

5.4 EDUCATIONAL IMPLICATIONS OF FINDINGS

Based on the results, the educational implications of this study may be summarised as follows;

- Developing one's metacognitive skills may be a critical ingredient for successful learning. The explanations students made to justify their choice of confidence rating revealed the factors that underlie the metacognitive monitoring skills of students during test-taking.
- Our study revealed that even in a science test in the specific format that we used, on a specific topic like stoichiometry, which requires predominantly procedural knowledge and formal reasoning, students believed an answer to be correct, based on feelings or a global evaluation of answer rather than on rationally motivated judgements. Our contribution to science education in this regard is that science educators should know that not every answer provided by students may be rationally based.
- The picture that emerged when all of the results for shifts in the prevalence of major response super-categories were analysed together is the following: Accuracy in the evaluation of posttest performance was associated with both a reduction in the prevalence of vague subjective judgments and with higher performance gain. An increase in the tendency to base metacognitive monitoring on vague global judgments of performance in the posttest was associated with reduced accuracy of performance evaluation and lower learning gain. This finding represents the most significant contribution of our study to current knowledge in science education.
- Students who do not develop or display the ability of making realistic judgements of their own mastery of new material during teaching and learning gain less from the experience and are more likely to fail in the posttest. There is therefore a possibility that weak students may be unskilled or incompetent and unaware of it (Kruger & Dunning, 1999). Metacognition involves monitoring one's progress as one learns and making changes and adapting one's strategies when one realises that these

strategies are not effective in order to achieve academic success. The danger arises when students are inaccurate in the monitoring of their progress during learning. If students are unaware of their poor performance in tests, they may be unlikely to realise the level of their incompetence, their limitations, deficiencies and misconceptions in a particular subject and consequently fail to gain from the teaching and learning experience and regulate their learning by changing ineffective strategies.

- When students do not know what they do not know and then incorrectly and naively assume that they have mastered a particular cognitive domain when they have not, educators have first to teach and make the students aware that they do not know something and only then can the educators teach the particular domain. Educators normally teach the domain without knowing that it is also important to teach students that they do not actually know something (Kennedy *et al.*, 2002).
- Knowledge of the nature of the task and the type of processing demands it will place on the individual (task variables), declarative or factual knowledge, contextual knowledge and procedural knowledge of a particular subject or topic, may assist learners to make more realistic judgements of their performance in a test. However, contrary to what was posited by Kruger and Dunning (1999), teaching of content material alone has little or no effect on the students' ability to accurately evaluate their performance.

5.5 CONTRIBUTIONS AND SIGNIFICANCE OF THE STUDY

Our study makes several theoretical and methodological contributions which are discussed below.

- **Theoretical contributions**
Kruger and Dunning (1999) suggested and Kennedy *et al.* (2002) found that the best way to raise the metacognitive ability of students in terms of accuracy of performance evaluation of their own work is to increase their content knowledge and training. Our results have shown that increasing their competence does not automatically reduce their bias in performance evaluation.

Our results showed that students who came into the field confident and then became realistic in their evaluation after instruction, i.e. the OC-R subgroup, actually gained the most in terms of learning stoichiometry. This indicated that there might be an element of bias in performance evaluation that seems to be beneficial. It seems that there are cases when confidence may be beneficial and some when it may be detrimental. Students who remained overconfident in the posttest, i.e. in the OC-OC subgroup, did not gain from the learning experience as much as those who entered overconfident but became better calibrated later. Those who entered tentatively as realists and then with a little exposure became completely unrealistic in their performance evaluation were shown to be the most vulnerable based on their lack of learning gain. Inaccuracy in performance evaluation in the pretest did not seem to hamper learning for both the OC-OC and OC-R subgroups. In fact students who were over-optimistic about their performance in the pretest may have been less intimidated by the challenges of the new content material than those who were better calibrated (R-R and R-OC subgroups).

The findings of Gramzow *et al.* (2003) that exaggerated performance evaluation by students with low grades is often motivated by the tendency to self-protect while exaggerated performance evaluation by students with high grades is often self-enhancement motivated could not be confirmed by the results obtained in this study. While it is possible that significant differences may be observed when a larger sample is used in future studies, it is expected that the findings of such a study will not be different due to the reasons stated before, i.e. that Gramzow and co-workers (2003) used as sample of undergraduate psychology students at a university in the USA for their study whereas our study was conducted with science students who were weak or unprepared and who were primarily second language English speakers in an African context. It is likely that our sample of science students was less self-aware and less experienced in reflection and self-analysis than the psychology students who routinely engage in such practices as part of their training. Moreover, the influence of cultural factors on the willingness of students to disclose feelings of self-doubt or inadequacy is likely to be different between the two samples. This finding is an indication that the findings of Gramzow *et al.* (2003) are unlikely to be generalizable to our kind of sample. The findings of this study suggest that Gramzow *et al.* (2003) should repeat their study in another context in order to substantiate their findings.

- **Methodological contributions**

Qualitative data analysis brought a richness to our understanding of factors underlying the metacognitive judgements of students, which would never have been obtained if a mixed method design had not been followed. Intuitively, a science lecturer would say “I want all my students to make rational judgements about the extent to which they have mastered the topic or whether they arrived at a correct answer”, because science educators are typically overly reliant on rational thinking and analytical reasoning. We consider chemistry to be a very systematic subject which requires logical reasoning, hence rational judgements. The analysis of qualitative data in this project revealed that even in a subject like chemistry which requires formal and objective reasoning, students may base judgments of the accuracy of answers on global evaluations, like FOKs, FOnK, affective feelings or cognitive feelings which are subjective judgements.

It became clear that the five distinct subgroups which emerged from the analysis of accuracy of performance evaluation not only differed in their pre-post self-evaluation, but also in the learning gain that was demonstrated after the teaching of stoichiometry. The fact that one subgroup achieved a learning gain significantly higher than the rest and one subgroup did not achieve any gain at all, highlighted the need for information that could assist in explaining the strengths and deficiencies in metacognitive skills that gave rise to such a difference. The qualitative data proved to be a rich resource for this purpose.

5.6 LIMITATIONS OF THE STUDY

The study has certain limitations which should be taken into consideration when interpreting the results. The limitations concern the following;

- The advantage of the embedded mixed methods design used in the study is that it can be used when a researcher does not have enough time or resources to commit extensively to quantitative and qualitative data collection because one data type is given less priority than the other. However, there are drawbacks to such an approach. The fact that two types of data sets are collected, the one type of data may take on a secondary role when there is insufficient time to commit to extensive data collection

and analysis of both sets of data. In a primarily qualitative design, a statement like “It is the compound that has carbon and oxygen, so it’s pretty clear”, provided by a student after indicating that they are 100% sure that (A) is the correct answer (which is in fact incorrect) to question 20 of the test instrument (Appendix I) shown below would not be viewed as significant on the basis of its incidences or frequencies of occurrence but for the fact that the statement had been made even when the answer was wrong. In such a case more time would be dedicated to uncovering the reasons why the students reasoned in that manner to justify their choice of confidence rating. A more deliberate and elaborate qualitative analysis and interpretation of each of the students’ responses could potentially reveal possible misconceptions or even enable the researcher to draw up questions that could be used to probe further during a follow-up interview. These findings could serve to enlighten science educators or even lecturers in academic development programmes in terms of misconceptions which could potentially be revealed by student responses as well as factors which are not necessarily rational, that students may use to assert their understanding or even the correctness of their answers in a test.

Figure 5.1: Question 20 of the stoichiometry test instrument shown in Appendix I

20. Use the following equation:

$$\text{CaCO}_3 + 2\text{HCl} \rightarrow \text{CO}_2 + \text{CaCl}_2 + \text{H}_2\text{O}$$

Calcium carbonate hydrochloric acid Carbon dioxide calcium chloride water

If 14 g of calcium carbonate react with 0.2 moles of hydrochloric acid, which reactant(s) do you use in your calculations to find the mass of carbon dioxide produced?
 (the molecular weight of calcium carbonate is 100.087g/mole)

- CaCO₃
- HCl
- Any of the two reactants
- None of the two reactants
- Both CaCO₃ and HCl

- The limitation of a case study design is that data obtained cannot be used to arrive at a generalising conclusion, i.e. findings based on data from the sample cannot be generalised to the entire population of students in academic development programmes. A case study design however, enabled, through the use of multiple data

collection techniques, the investigation of bias in performance evaluation in a group of BFYP students within a real classroom context. Understanding and insight gained from the factors underlying bias in performance evaluation, specifically subjective judgements of performance, may serve to inform staff at tertiary institutions who are involved in academic development programmes. The findings of this study may be used to inform the design, monitoring and presentation of a curriculum and assessment strategies unique to such programmes in order to achieve improved pass rates and therefore increased access for students into mathematics and science fields.

- Only 94 complete records were obtained in our study. Incomplete records resulted in a small sample. As the project developed it became clear that data analysis would have to be conducted for subsets rather than the whole sample. This sample was not big enough for that, especially since subgroups were not equally populated.
- Although the posttest results strongly suggest that teaching might have had a significant effect on performance and accuracy of performance evaluation of the students we cannot say that this was the case as there were other confounding factors which had not been controlled. In a future study the influence of teaching may be investigated by means of a control group which could write the pre- and posttest without exposure to teaching as an intervention.
- Only two students, one male and one female, were realistic in their performance evaluation in the pretest and became underconfident in the posttests. The female student obtained a score of 10 out of 19 (53%) and 14 out of 19 (74%) in the pre- and posttest, respectively which translated into a 40% learning gain. The male student obtained a score of 6 out of 19 (32%) in the pretest and 15 out of 19 (79%) in the posttest, translating into a 69% learning gain. The posttest performance of the two students placed them as top performers and finding them in the R-UC subgroup confirmed the findings of Ochse (2003) and what Kruger and Dunning (1999) had suggested, that top-performing students were more cautious and modest in their self-evaluations relative to their peers. Although our results confirmed what was found and suggested in the literature, reliable inferences could not be made from a sample of two students. It seems however, that this tendency to be more cautious and modest may be observed for top-performing students as was the case in our study and

in the literature. Our sample was made up of weak, poor-performing students and our results showed that this kind of sample would not include enough students with attributes similar to those observed in the R-UC subgroup, hence it cannot be anticipated that if this study is repeated with a bigger, similar group of students, different results would be obtained.

5.7 RECOMMENDATIONS

The following recommendations arise from the study:

- Assessment in the form of test-taking should provide students with an opportunity to assess their understanding of the material as well as the effectiveness of their study and learning skills. Understanding factors that underlie students' planning and monitoring during test-taking may contribute to the creation of evaluation practices and conditions that promote learning during the evaluation process (Carvalho, 2007). In an effort to understand these factors, the practice of expecting students to judge their performance in each test item and to provide an explanation for the judgements made, should be adopted in chemistry classes.
- Kennedy *et al.* (2002) suggested that those who are less competent are less able to reflect accurately on their ability in a given domain. Wade, Trathen and Schraw (1990) demonstrated by means of examples of students' reflections on their thinking while they were reading that the readers' reflections fostered the planning, monitoring, evaluation and use of available information to make sense of what they read. They suggested that as conventional descriptions of metacognition, such reflections served to unveil judgements about the readers' thinking processes. Chemistry students should be encouraged to train and develop their metacognitive skills by routinely engaging in such reflective practices as part of their learning.
- The quality and intervals of feedback provided by the educators during assessment activities as well as grading practices should be improved with the aim of making them instrumental in informing students of what they know and do not know.

- Tests should consist of tasks that require higher cognitive demand and construction of responses, requiring deeper engagement which may force students to critically and realistically judge their performance.

5.8 AREAS FOR FURTHER RESEARCH

The schooling system in South Africa mainly produces students who are under-prepared for tertiary studies. To facilitate access for such students to science and mathematics-related programmes and ultimately, careers, tertiary institutions have put academic development programmes in place as an intervention. However, under-prepared or weak students have been found to exhibit high levels of overconfidence when evaluating their performance. Overly-optimistic bias in performance evaluation can potentially have serious consequences, as it may lead students to study less than if they had accurate perceptions (Grimes, 2002; Nowell & Alston, 2007). The majority of students in our sample were overconfident in both the pre- and posttest. Students who remained and became overconfident had a low or no learning gain. A similar study may be conducted to investigate bias in performance evaluation and its association to learning gain in academic development programmes of other institutions and confirm the current findings.

The posttest results strongly suggested that teaching might have had a significant effect on the performance of the students but we cannot say that this was the case because of other confounding factors which were not controlled. The study should be repeated on a larger sample so that the results can be verified and hopefully generalised to other similar groups.

In order to make the findings of this study accessible to practitioners there is more work that needs to be done. Patterns were observed but these were not clear and conclusive enough to be presented as information which could be made available to and implemented by practitioners, e.g. the typical chemistry lecturers. There is a new realization of factors such as those we have found underlying the metacognitive judgements of students that are important, but it is not yet known how these could be dealt with. More research should be conducted to explore these factors further before the findings of such a study can benefit the day-to-day teaching of chemistry.

REFERENCES

1. Ben-Zvi, R., Eylon, B. & Silberstein, J. (1988). Theories, principles and laws. *Education in Chemistry*, 89 – 92.
2. Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. *Sex Roles*, 41(314), 279 – 296.
3. Beyer, S. & Bowden, E. M. (1997). Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias. *Personality and Social Psychological Bulletin*, 23, 157 – 180.
4. Bogdan, R. C. & Biklen, S. K. (1982). *Qualitative research for education: An introduction to theory and methods*. Boston: Allyn and Bacon.
5. Boujaoude, S. & Barakat, H. (2000). Secondary school students' difficulties with stoichiometry. *School Science Review*, 81(296), 91 – 98.
6. Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77 – 101.
7. Caracelli, V. J. & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 15(2), 195 – 207.
8. Carvalho, M. K. F. & Yuzawa, M. (2001). The effects of social cues on confidence judgements mediated by knowledge and regulation of cognition. *The Journal of Experimental Education*, 69 (4), 325 – 343.
9. Carvalho, M. K. F. (2007). Confidence judgements in real classroom settings: Monitoring performance in different types of tests. *International Journal of Psychology*, 1 – 16.
10. Cavaye, A. L. M. (1996). Case study research: A multi-faceted research approach. *Information Systems Journal*, 6, 227 – 242.
11. Cohen, L., Manion, L. & Morrison, K. (2007). *Research methods in Education* (6th ed.). USA & Canada: Routledge.
12. Creswell, J. W. & Plano, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.
13. Dunlosky, J., Serra, M. J., Matvey, G. & Rawson, K. A. (2005). Second-Order Judgements About Judgements of Learning. *The Journal of General Psychology*, 132(4), 335 – 346.
14. Dunning, D., Johnson, K., Ehrlinger, J. & Kruger, J. (2003). Why people fail to recognise their own incompetence. *American Psychological Society*, 12 (3), 83 – 87.

15. Dykstra, D., Boyle, C. & Monarch, I. (1992). Studying conceptual change in learning physics. *Science Education*, 76, 615 – 662.
16. Ehrlinger, J. (2008). Skill level, self-views and self-theories as sources of error in performance evaluation. *Social and Personality Psychology Compass*, 2(1), 382 – 398.
17. Ehrlinger, J. & Dweck, C. S. (forthcoming) cited in Ehrlinger (2008). Carter, T. J. & Dunning, D. (2008). Faulty performance evaluation: Why evaluating one's own competence is an intrinsically difficult task. *Social and Personality Compass*, 2(1), 346 – 360.
18. Falconer, D. J., & Mackay, D. R. (1999). *Ontological problems of pluralist research methodologies*. Retrieved March 22, 2011, from aisel.aisnet.org/cgi/viewcontent.cgi?article=1573&context=amcis1999
19. Fernandez-Duque, D. & Black, S. E. (2007). Metacognitive judgment and denial of deficit: Evidence from frontotemporal dementia. *Judgment and Decision Making*, 2(5), 359 – 370.
20. Field A. (2009). *Discovering Statistics using SPSS (3rd ed.)*. Los Angeles, London, New Delhi, Singapore & Washington DC: SAGE.
21. Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. Resnick (Ed.), *The nature of intelligence* (pp. 231-236). Hillsdale, NJ: Erlbaum
22. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry. *American Psychologist*, 34, 906 – 911.
23. Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. 2nd Ed. New York: Wiley.
24. Frazer, M. J. & Servant, D. (1986). Aspects of stoichiometry titration calculations. *Education in Chemistry*, 23, 54 – 56.
25. Frazer, M. J. & Servant, D. (1987). Aspects of stoichiometry, where do students go wrong? *Education in Chemistry*, 24, 73 – 75.
26. Furió, C., Azcona, R. & Guisasola, J. (2002). The learning and teaching of the concepts 'Amount of substance' and 'Mole': A review of the literature. *Chemistry Education: Research and Practice in Europe*, 3(3), 277 – 292.
27. Gauchon, L. & Méheut, M. (2007). Learning about stoichiometry: from students' preconceptions to the concept of limiting reactant. *Chemistry Education Research and Practice*, 8(4), 362 – 375.

28. Glucksberg, S. & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 311 – 325.
29. Goodie, A. S. (2003). The effects of Control on Betting: Paradoxical Betting on Items of High Confidence with Low value. *Journal of Experimental Psychology*, 29(4), 598 – 610.
30. Gramzow, R. H., Elliot, A. J., Asher, E. & McGregor, H. A. (2003). Performance evaluation bias and academic performance: Some ways and some reasons why. *Journal of Research in Personality*, 37, 41 – 61.
31. Greifeneder, R., Bless, H. & Pham, M. T. (2010). When do people rely on affective and cognitive feelings in judgement? A review. *Personality and Social Psychology Review*, XX(X), 1 – 35.
32. Grimes, P. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *Journal of Economic Education*, 33(1), 15 – 30.
33. Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 105 – 117). Thousand Oaks, CA: Sage.
34. Hake R. R. (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66, 64 – 74.
35. Halter, J. SDSU Department of Educational Technology.
<http://coe.edu/eet/Articles/metacognition/start.htm> accessed on 2008/08/29.
36. Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56, 208 – 216.
37. Hartman, H. J. (2001). *Metacognition in Learning and Instruction: Theory Research and Practice*. Dordrecht: Kluwer Academic Publishers.
38. Hasan, S., Bagayoko, D. & Kelley, E. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, 34, 249 – 299.
39. Haun, D. E., Zeringue, A., Leach, A. & Foley, A. (2000). Assessing the competence of specimen-processing personnel. *Laboratory medicine*, 31, 633 – 637.
40. Healy, M., & Perry, C. (2000). Comprehensive criteria to judge validity and reliability of qualitative research within the realism paradigm. *Qualitative Market Research – An international Journal*, 3(3), 118 – 126.

41. Heller, P. & Finley, F. (1992). Variable uses of alternative conceptions: A case study in current electricity. *Journal of Research in Science Teaching*, 29, 259 – 275.
42. Higgins, E & Tatham, L. (2003). Exploring the potential for multiple-choice questions in assessment. *Assessment 2* (4.shtml) ISSN 1477 – 1241. Retrieved on 12th May 2010, from <http://www.Itu.mmu.ac.uk/Itia/Issue 4/higginstatham.shtml>
43. Huddle, P. A. & Pillay, A. E. (1996). An In-Depth Study of Misconceptions in Stoichiometry and Chemical Equilibrium at a South African University. *Journal of Research in Science Teaching*, 33(1), 65 – 77.
44. Ivankova, N. V., Cresswell, J. W. & Clark, V. L. P. (2007). Foundations and approaches to mixed methods research. In Maree, K. (Ed.). *First steps in research*. Pretoria: Van Schaik Publishers.
45. Jing, L., Kazuhisa, N. & Yuejia, L. (2003). Neural correlates of “feeling-of-not-knowing”: evidence from functional MRI. *Chinese Science Bulletin*, 48(2), 144 – 147.
46. Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of Computer assisted learning*, 7(2), 75 – 83.
47. Kennedy, E. J., Lawton, L. & Plumlee, L. (2002). Blissful ignorance: The problem of unrecognised incompetence and academic performance. *Journal of Marketing Education*, 24(3), 243 – 252.
48. Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). London: Routledge.
49. Kolb, D. (1978). The Mole. *Journal of Chemical Education*, 55, 728 – 732.
50. Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, 9, 149 – 171.
51. Koriat, A. & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 187 – 194.
52. Krauss, S. E. (2005). Research paradigms and meaning making: a primer. *The Qualitative Report*, 10 (4), 758 – 770.
53. Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated performance evaluations. *Journal of Personality and Social Psychology*, 77 (6), 1121 – 1134.
54. Laugier, A. & Dumon, A. (2000). as cited in Gauchon, L. & Méheut, M. (2007).
55. Lichtenstein, S. & Fischhoff (1997). Do these who know more also know more about how much they know? *Organisational behaviour and Human Performance*, 20, 159 – 183.

56. Livingston, J. A. (1997). Metacognition: An overview. Unpublished manuscript, State University of New York at Buffalo.
57. Loftus, E. F. & Wagenaar, W. A. (1988). Lawyers' predictions of successes. *Jurimetrics Journal*, 29, 437 – 453.
58. Mabila, T. E., Malatje, S. E., Addo-Bediako, A., Kazeni, M. M. M. & Mathabatha, S. S. (2006). The role of foundation programmes in science education: The UNIFY programme at the University of Limpopo, South Africa. *International Journal of Educational Development*, 26, 295 – 304.
59. Maki, R. H. (1999). The roles of competition, target accessibility, and cue familiarity in metamemory for word pairs. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25, 1011 – 1023.
60. Maree, K. & Pietersen, J. (2007). Surveys and the use of questionnaires. In Maree, K. (Ed.). *First steps in research*. Pretoria: Van Schaik Publishers.
61. Martin, A.J., Marsh, H.W. & Debus, R.L. (2003). Self-handicapping and defensive pessimism: A model of self-protection from a longitudinal perspective. *Contemporary Educational Psychology*, 28, 1 – 36.
62. Martin, M. O., Mullis, I.V.S., Gonzalez, E.J. & Chrostowski, S. J. (2004). TIMSS 2003 international science report: Findings from IEA's trends in International Mathematics and Science study at the fourth and eighth grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
63. McFate, C. & Olmsted III, J. (1999). Evaluating Student Preparation through Placement Tests. *Journal of Chemical Education*, 76(4), 562 – 565.
64. Metcalfe, J. (2000). Metamemory: Theory and data. In Tulving, E. & Craik, F. I. M. (Eds.), *The Oxford Handbook of Memory*, 197 – 211. New York: Oxford University Press.
65. Mullis, I.V.S., Martin, M. O., Gonzalez, E.J. & Chrostowski, S. J. (2004). TIMSS 2003 international mathematics report: Findings from IEA's trends in International Mathematics and Science study at the fourth and eighth grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
66. Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In Bower, G. H. (Ed.). *The psychology of learning and motivation: advances in research and theory*. San Diego, California: Academic Press, Inc.
67. Nieuwenhuis, J. 2007. Qualitative research designs and data gathering techniques. In Maree, K. (Ed.). *First steps in research*. Pretoria: Van Schaik Publishers.

68. Novak, J. D. (1988). Learning science and the science of learning. *Studies in Science Education*, 15, 77 – 101.
69. Nowell C. & Alston M. R. (2007). I thought I Got an A! Overconfidence Across the Economics Curriculum. *Journal of Economic Education*, 131 – 142.
70. Nunnally, J. C. (1967). *Psychometric Theory*. New York: McGraw-Hill.
71. Ochse, C. (2003). Are positive self-perceptions and expectancies really beneficial in an academic context? *South African Journal of Higher Education*, 17 (1), 6 – 73.
72. Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L. & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgements. *The Journal of General Psychology*, 129, 257 – 299.
73. Pietersen, J. & Maree, K. 2007. Standardisation of a questionnaire. In Maree, K. (Ed.). *First steps in research*. Pretoria: Van Schaik Publishers.
74. Pietersen, J. & Maree, K. 2007. Statistical analysis I: descriptive statistics. In Maree, K. (Ed.). *First steps in research*. Pretoria: Van Schaik Publishers.
75. Potgieter, M., Davidowitz B. & Mathabatha S. (2007). Do they know that they don't know? The relationship between confidence and performance of first year chemistry students at three tertiary institutions in South Africa. Presented at the 38th annual conference of the Australasian Science Education Research Association (ASERA) in Fremantle, WA.
76. Reder, L. M. & Ritter, F. E. (1987). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18 (3), 435 – 451.
77. Ridley, D. S., Schutz, P. A., Glanz, R. S. & Weinstein, C. E. (1992). Self-regulated learning: the interactive influence of metacognitive awareness and goal-setting. *Journal of Experimental Education*, 60 (4), 293 – 306.
78. Rosenthal, D. M. (2000). Consciousness, Content, and Metacognitive Judgments. *Consciousness and Cognition*, 9, 203 – 214.
79. Schaefer, P. S., Williams, C. C., Goodie, A. S. & Campbell, W. K. (2004). Overconfidence and the Big Five. *Journal of Research in Personality*, 38, 473 – 480.
80. Schwartz, B. L. & Perfect, T. J. (2002). Introduction: toward an applied metacognition. In Schwartz, B. L. & Perfect, T. J. (Eds.). *Applied Metacognition*. Cambridge University Press.
81. Seifert, T. L. (2004). Understanding student motivation. *Educational Research*, 46(2), 137 – 149.

82. Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Journal of Instructional Psychology*, 22(1), 77 – 88.
83. Strecher, V. J., Kreuter, M. W. & Kobrin, S. C. (1995). Do cigarette smokers have unrealistic perceptions of their heart attack, cancer, and stroke risks? *Journal of Behavioural medicine*, 18, 45 – 54.
84. Tashakkori, A. & Teddlie, C. 1998. Mixed methodology: combining qualitative and quantitative approaches. *Applied Social Research Methods Series*, 46. Thousand Oaks:Sage.
85. Wade, W., Trathen, W. & Schraw, G. (1990). An analysis of spontaneous study strategies. *Reading Research Quarterly*, 25, 147 – 166.
86. Winnie, P. H. & Nesbit, J. C. (2010). The psychology of academic achievement. *Annual Review of Psychology*, 61, 653 – 678.
87. Woolfolk, A. E. (1998). *Educational Psychology* (7th ed.). Needham Heights: Allyn & Bacon.
88. Yun, S. & Takeuchi, R. (2007). Employee Self-Enhancement Motives and Job Performance Behaviors: Investigating the Moderating Effects of Employee Role Ambiguity and Managerial Perceptions of Employee Commitment. *Journal of Applied Psychology*, 92(3), 745 – 756.

APPENDIX I: THE TEST INSTRUMENT

STOICHIOMETRY TEST



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Natural and Agricultural
Sciences
Chemistry department
Tel: (012) 420 4905
Email: kgadi.mathabathe@up.ac.za

Chemistry test for BSc Four Year Programme (BFYP) students

Thank you for being willing to participate in this test.

Please answer each question honestly and accurately by circling the alphabet of the multiple choice question corresponding to the option you have chosen as correct. E.g. **a.**

This information will be treated confidentially.

You should have your calculator and a periodic Table with you, but your textbook should be packed away. Thank You.

The results of this test are very important to us for research purposes.

For Office Use Only		
Actual score		
Av. Confidence		

Student Number:

--	--	--	--	--	--	--	--	--	--

Question 1

Given the equation $3A + B \rightarrow C + D$, if 4 moles of A reacted with 2 moles of B, which of the following is true?

- a. The limiting reactant is the one with the higher molar mass.
- b. A is the limiting reactant because you need 6 moles of A to react with 2 moles of B.
- c. B is the limiting reactant because three A molecules react with every one B molecule.
- d. B is the limiting reactant because there are only 2 moles of B available.
- e. Neither reactant is limiting.

1.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
----------------	-----------	-----------	-----------	-----------	-----------------	-----------	-----------	-----------	-----------	------------------

1.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q1.1: _____

Q1.2: _____

Question 2

A mole ratio is:

- a. A fraction.
- b. A ratio.
- c. A conversion factor.
- d. All of the above.
- e. both a ratio and a conversion factor

2.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

2.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q2.1: _____

Q2.2: _____

Question 3

Given the following balanced equation: $\text{N}_2 + 3\text{H}_2 \rightarrow 2\text{NH}_3$, which of these is an **INCORRECT** mole ratio?

a. $\frac{3 \text{ moles H}_2}{1 \text{ mole N}_2}$

b. $\frac{2 \text{ moles NH}_3}{3 \text{ moles H}_2}$

c. $\frac{6 \text{ moles NH}_3}{3 \text{ moles N}_2}$

d. $\frac{1 \text{ mole N}_2}{3 \text{ moles NH}_3}$

3.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

3.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q3.1: _____

Q3.2: _____

Question 4

Balancing a chemical equation is achieved by:

- setting the coefficients equal to one and adjusting subscripts in the formulas
- adjusting the coefficients to the smallest possible whole number ratio
- adjusting the number of elements produced
- adjusting the formula of a compound.
- writing in appropriate coefficients to ensure mass balance then adjusting the coefficients to the smallest possible whole number ratio.

4.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

4.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q4.1: _____

Q4.2: _____



Question 5

When the equation: $\text{Cu}_2\text{O} + \text{CH}_4 \rightarrow \text{H}_2\text{O} + \text{Cu} + \text{CO}_2$ is correctly balanced the coefficient in front of the formula for copper (I) oxide (Cu_2O) is:

- a. 1
- b. 2
- c. 3
- d. 4
- e. none of the above.

5.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------





5.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q5.1: _____

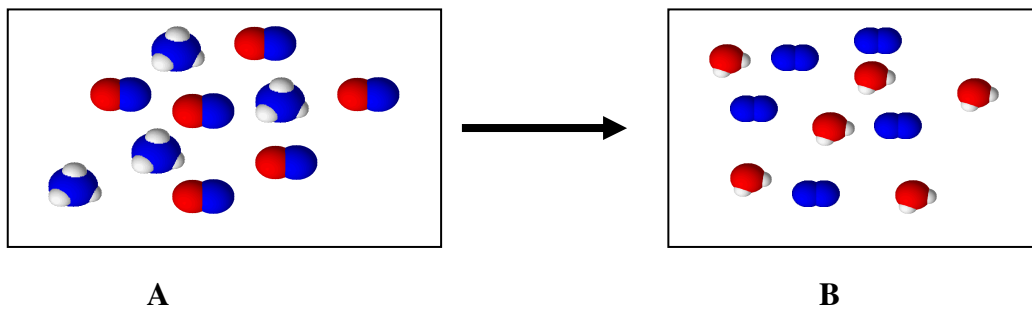
Q5.2: _____

Question 6

Ammonia () reacts with Nitrogenmonoxide () to form Nitrogen gas () and water ().

Consider the mixture of Ammonia and Nitrogenmonoxide in a closed container before (A) and after (B) the reaction has occurred. All reactants were used up during the reaction.

How many moles of each reactant were there if 13.7 moles of $N_2(g)$ is produced?



- 10.96 moles $NH_3(g)$ and 16.44 moles $NO(g)$
- 2.74 moles $NH_3(g)$ and 16.44 moles $NO(g)$
- 3.43 moles $NH_3(g)$ and 5.15 moles $NO(g)$
- 54.8 moles $NH_3(g)$ and 82.2 moles $NO(g)$

6.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

6.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q6.1: _____

Q6.2: _____

Question 7

How many moles of methane (CH_4) are required to produce one mol of copper (Cu) by the reaction given in **question 5**?

- a. 0.5 moles
- b. 1.0 moles
- c. 1.5 moles
- d. 0.25 moles
- e. 0.125 moles
- f. none of the above.

7.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

7.3 Why did you choose that specific confidence indicator?

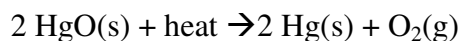
For Office Use Only

Q7.1: _____

Q7.2: _____

Question 8

Joseph Priestley discovered oxygen in the eighteenth century by using heat to decompose mercury(II) oxide:



What mass of mercury(II) oxide would be required to produce 100 g of O₂.

- a. 100 g
- b. 200 g
- c. 627 g
- d. 677 g
- e. 1354 g
- f. 2700 g

8.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

8.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q8.1: _____

Q8.2: _____

Question 9

What mass of calcium carbonate (CaCO_3) is needed to react completely with 50.00 mL of 0.383 M sulfuric acid (H_2SO_4) according to the following balanced chemical equation? $\text{CaCO}_3 + \text{H}_2\text{SO}_4 \rightarrow \text{CaSO}_4 + \text{CO}_2 + \text{H}_2\text{O}$

- a. 19.2 g
- b. 0.958 g
- c. 1.92 g
- d. 9.58 g
- e. 767 g
- f. 13.1 g

9.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------



9.3 Why did you choose that specific confidence indicator?

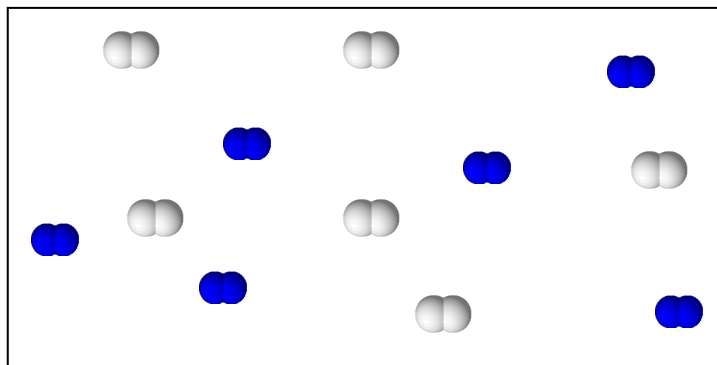
For Office Use Only

Q9.1: _____

Q9.2: _____

Question 10

Nitrogen (N_2) and hydrogen (H_2) react to form ammonia (NH_3). Consider the mixture of N_2 () and H_2 () in a closed container as illustrated below:



In your opinion, the chemical reaction **stops** when:

- all the Nitrogen is used up.
- all the Nitrogen and all the Hydrogen are both totally used up.
- all the Hydrogen is used up.
- all the Nitrogen or all the Hydrogen is used up.
- I do not know.

10.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

10.3 Why did you choose that specific confidence indicator?

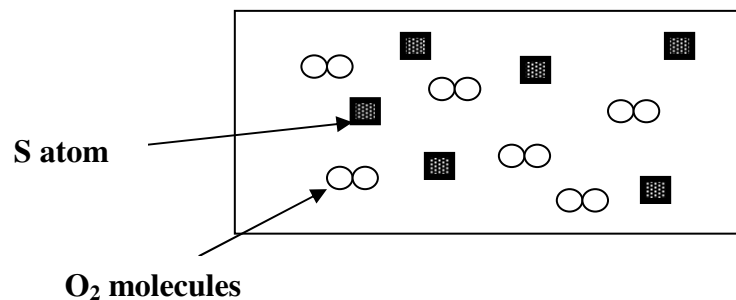
For Office Use Only

Q10.1: _____

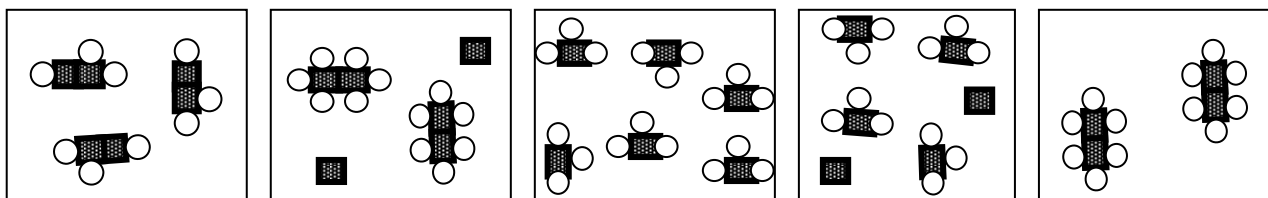
Q10.2: _____

Question 11

The diagram represents a mixture of S atoms and O₂ molecules in a closed container.



Which diagram shows the results after the mixture reacts as completely as possible according to the equation $2S + 3O_2 \rightarrow 2SO_3$?



a.

b.

c.

d.

e.

11.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

11.3 Why did you choose that specific confidence indicator?

For Office Use Only

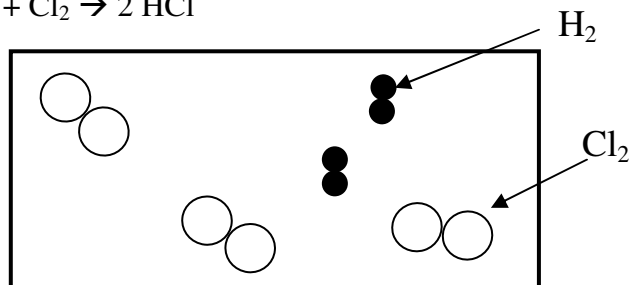
Q11.1: _____

Q11.2: _____

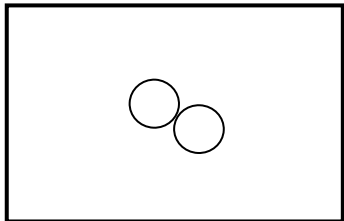
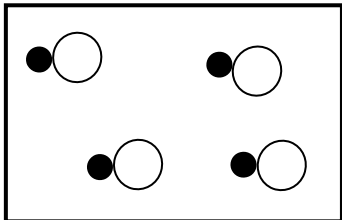
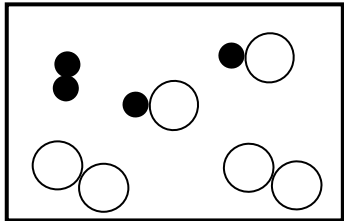
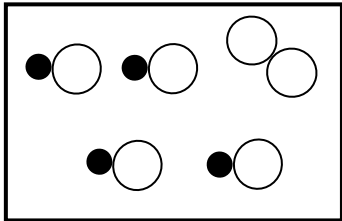
Question 12

12.1 Here is a picture of a container with three Cl_2 molecules and two H_2 molecules. A chemical reaction occurs until the maximum amount of HCl has been produced.

The reaction is $\text{H}_2 + \text{Cl}_2 \rightarrow 2 \text{HCl}$



The picture of the container after the reaction looks like:

- a. 
- b. 
- c. 
- d. 
- e. none of these

12.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

12.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q12.1: _____

Q12.2: _____

Question 13

13.1 Consider the following generic chemical reaction:



How many moles of B would you need to react completely with 5 moles of A?

- a. 1.2
- b. 1.5
- c. 2
- d. 3.3
- e. none of the above

13.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

13.3 Why did you choose that specific confidence indicator?

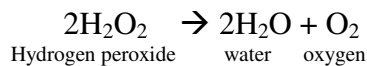
For Office Use Only

Q13.1: _____

Q13.2: _____

Question 14

14.1 Hydrogen peroxide will decompose to form water and oxygen gas according to the following equation:

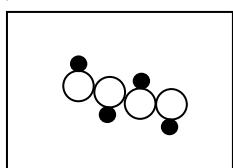


Use the following key for the diagrams:

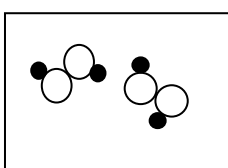
Oxygen ○

Hydrogen ●

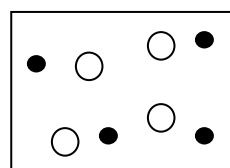
Which diagram is the best representation of the hydrogen peroxide before it decomposes?



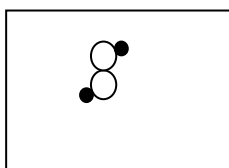
a.



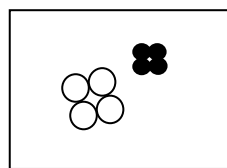
b.



c.



d.



e.

14.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

14.3 Why did you choose that specific confidence indicator?

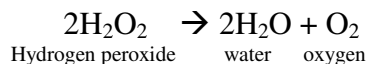
For Office Use Only

Q14.1: _____

Q14.2: _____

Question 15

15.1 Hydrogen peroxide will decompose to form water and oxygen gas according to the following equation.

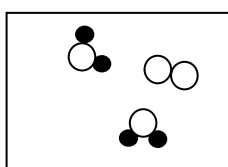


Use the following key for the diagrams:

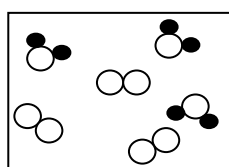
Oxygen ○

Hydrogen ●

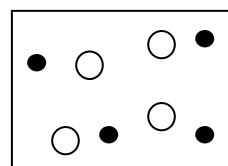
Which diagram is the best representation of the products after hydrogen peroxide decomposes?



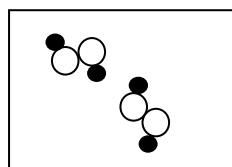
a.



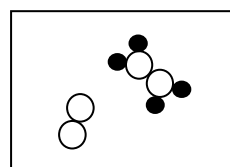
b.



c.



d.



e.

15.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

15.3 Why did you choose that specific confidence indicator?

For Office Use Only

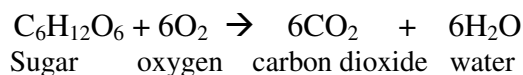
Q15.1: _____

Q15.2: _____



Question 16

16.1 Your body reacts sugar with oxygen to form carbon dioxide and water according to the following chemical equation.



How many million oxygen atoms would be needed to react completely with one million sugar molecules?

- a. 3
- b. 6
- c. 9
- d. 12
- e. none of the above

16.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

16.3 Why did you choose that specific confidence indicator?

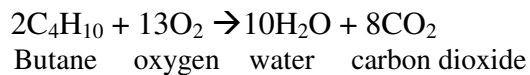
For Office Use Only

Q16.1: _____

Q16.2: _____

Question 17

17.1 Butane is combusted completely with excess oxygen to form water and carbon dioxide.



The reaction yielded 1 mole of water. How many moles of carbon dioxide were produced?

- a. 0.8
- b. 1.25
- c. 4
- d. 8
- e. none of the above

17.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

17.3 Why did you choose that specific confidence indicator?

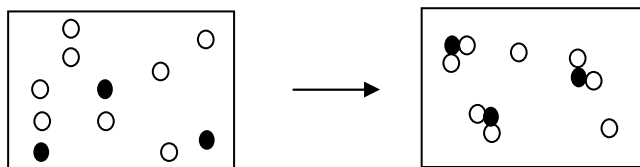
For Office Use Only

Q17.1: _____

Q17.2: _____

Question 18

18.1 The reaction of element X (●) with element Y (○) is represented in the following diagram.



Which equation describes this reaction?

- a. $3X + 8Y \rightarrow X_3Y_8$
- b. $3X + 6Y \rightarrow X_3Y_6$
- c. $X + 2Y \rightarrow XY_2$
- d. $3X + 8Y \rightarrow 3XY_2 + 2Y$
- e. $X + 4Y \rightarrow XY_2$

18.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

18.3 Why did you choose that specific confidence indicator?

For Office Use Only

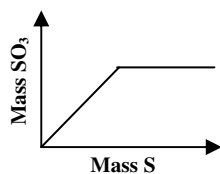
Q18.1: _____

Q18.2: _____

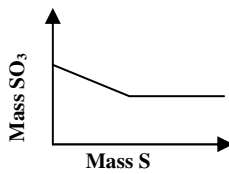


Question 19

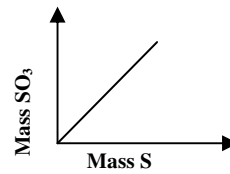
19.1 For the chemical reaction: $2S + 3O_2 \rightarrow 2SO_3$, which of the following graphs best represents the formation of SO_3 , if S is added indefinitely (or in excess) to a fixed amount of O_2 ?



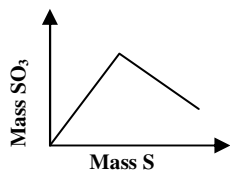
a.



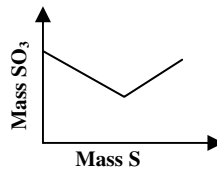
b.



c.



d.



e.

19.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

19.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q19.1: _____

Q19.2: _____



Question 20

20.1 Use the following equation:



If 14g of calcium carbonate react with 0.2 moles of hydrochloric acid, which reactant(s) determines the mass of carbon dioxide produced?
(the molecular weight of calcium carbonate is 100 g/mole)

- a. CaCO_3
- b. HCl
- c. Any of the two reactants
- d. None of the two reactants
- e. Both CaCO_3 and HCl

20.2 How **confident/sure** are you that the answer you have chosen is correct?

0% sure	10	20	30	40	50% sure	60	70	80	90	100% sure
---------	----	----	----	----	----------	----	----	----	----	-----------

20.3 Why did you choose that specific confidence indicator?

For Office Use Only

Q20.1: _____

Q20.2: _____

Finished!

Thank you for your participation!!



Periodic Table of Elements

1 H 1.01																	2 He 4.00
3 Li 6.94	4 Be 9.01											5 B 10.81	6 C 12.01	7 N 14.01	8 O 16.00	9 F 19.00	10 Ne 20.18
11 Na 22.99	12 Mg 24.31											13 Al 26.98	14 Si 28.09	15 P 30.97	16 S 32.07	17 Cl 35.45	18 Ar 39.95
19 K 39.10	20 Ca 40.01	21 Sc 44.96	22 Ti 47.87	23 V 50.95	24 Cr 52.00	25 Mn 54.94	26 Fe 55.85	27 Co 58.93	28 Ni 58.69	29 Cu 63.55	30 Zn 65.39	31 Ga 69.72	32 Ge 72.61	33 As 74.92	34 Se 78.96	35 Br 79.90	36 Kr 83.80
37 Rb 85.47	38 Sr 87.62	39 Y 88.91	40 Zr 91.22	41 Nb 92.91	42 Mo 95.94	43 Tc 98.91	44 Ru 101.07	45 Rh 102.91	46 Pd 106.42	47 Ag 107.87	48 Cd 112.41	49 In 114.82	50 Sn 118.71	51 Sb 121.76	52 Te 127.60	53 I 126.90	54 Xe 131.29
55 Cs 132.91	56 Ba 137.33	57 La 138.91	72 Hf 178.49	73 Ta 180.95	74 W 183.84	75 Re 186.21	76 Os 190.23	77 Ir 192.22	78 Pt 195.08	79 Au 196.97	80 Hg 200.59	81 Tl 204.38	82 Pb 207.20	83 Bi 208.98	84 Po 208.98	85 At 209.99	86 Rn 222.01
87 Fr 223.02	88 Ra 226.03	89 Ac 227.03	104 Rf 261.11	105 Db 262.11	106 Sg 263.12	107 Bh 262.12	108 Hs 265	109 Mt 266									
58 Ce 140.12	59 Pr 140.91	60 Nd 144.24	61 Pm 144.91	62 Sm 150.36	63 Eu 151.97	64 Gd 157.25	65 Tb 158.93	66 Dy 162.50	67 Ho 164.93	68 Er 167.26	69 Tm 168.93	70 Yb 173.94	71 Lu 174.97				
90 Th 232.04	91 Pa 231.04	92 U 238.03	93 Np 237.05	94 Pu 244.06	95 Am 243.06	96 Cm 247.07	97 Bk 247.07	98 Cf 251.08	99 Es 252.08	100 Fm 257.10	101 Md 258.10	102 No 259.10	103 Lr 262.11				

APPENDIX II: STUDENT QUESTIONNAIRE (PILOT STUDY)

Please answer the following Questionnaire based on the test.

1. Are there any ambiguities or potential language barriers that may cause a second language respondent to misunderstand? If any, kindly indicate below with a brief explanation if possible.

2. Please comment on the following:

2.1 Clarity of instructions (i.e. will a student be able to understand what is being asked and hence be able to answer the questions the way the examiner wants them answered?)

2.2 Vocabulary and terminology used in the test (i.e. is it appropriate for the level of students for which the test is intended?)

Finished!

Thank you for your participation!!

APPENDIX III: EDUCATOR INSTRUCTION SHEET (PILOT STUDY)

Dear Teacher,

Thank you for being willing to complete this task. Your honest response in this regard is highly appreciated and valuable to the outcome of this study. The purpose of this task is to establish the validity of a chemistry test on stoichiometry, which will be written by first-year BSc four year programme students of the University of Pretoria. The results of the test will be used in an MSc research project. You are advised to carefully go through instructions below before commencing with the task.

Instructions

1. You are provided with two documents, the students' copy of the test as well as the teachers' copy of the test. Go through the students' copy first. Attempt to answer the questions in the test. At the end, please record the time it took you to complete the test.
2. **DO NOT** go through the teachers' copy before answering the questions in the students' copy.

APPENDIX IV: EDUCATOR QUESTIONNAIRE: SECTION B
(PILOT STUDY)

SECTION B

Please answer the following Questionnaire based on the students' copy of the test.

1. Are there any ambiguities or potential language barriers that may cause a second language respondent to misunderstand? If any, kindly indicate below with a brief explanation if possible.

2. Is it reasonable to expect a grade 11 or 12 learner to be able to answer the questions in this test? (cross the relevant option)

YES	NO
-----	----

3. Please indicate on the test, question(s) which you think is/are too easy (**E**) or too difficult (**D**) by circling the number of the question(s) e.g. **(5) D** or **(5) E**

4. Please comment on the following:

4.1 Overall presentation of the test (i.e. Format)

4.2 Clarity of instructions (i.e. will a student be able to understand what is being asked and hence be able to answer the questions the way the examiner wants them answered?)



4.3 Soundness of the chemistry content in the items (questions).

4.4 Vocabulary and terminology used in the test (i.e. is it appropriate for the level of students for which the test is intended?)

FINISHED, YOU MAY PROCEED TO SECTION C.

APPENDIX V: EDUCATOR QUESTIONNAIRE: SECTION C
(PILOT STUDY)

Question 1

Given the equation $3A + B \rightarrow C + D$, if 4 moles of A reacted with 2 moles of B, which of the following is true?

- The limiting reactant is the one with the higher molar mass.
- A is the limiting reactant because you need 6 moles of A to react with 2 moles of B.
- B is the limiting reactant because three A molecules react with every one B molecule.
- B is the limiting reactant because there are only 2 moles of B available.
- Neither reactant is limiting.

This item intends to measure the following:

- The student is able to determine the limiting reactant.
- The student can execute numerical problem solving to extract quantitative information on the reactant that limits product formation.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 2

A mole ratio is:

- a. A fraction
- b. A ratio
- c. A conversion factor
- d. All of the above
- e. Both a ratio and a conversion factor

This item intends to measure the following:

- The student knows what a mole ratio is.
- The student knows that a mole ratio is in a form of a fraction i.e. written as a ratio of moles and is used as a conversion factor in solving stoichiometry problems.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 3

Given the following balanced equation: $\text{N}_2 + 3\text{H}_2 \rightarrow 2\text{NH}_3$, which of these is an

INCORRECT mole ratio?

a. $\frac{3 \text{ moles H}_2}{1 \text{ mole N}_2}$

b. $\frac{2 \text{ moles NH}_3}{3 \text{ moles H}_2}$

c. $\frac{6 \text{ moles NH}_3}{3 \text{ moles N}_2}$

d. $\frac{1 \text{ mole N}_2}{3 \text{ moles NH}_3}$

This item intends to measure the following:

- The student can identify the correct mole ratio from a symbolic representation of a balanced chemical equation.





Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

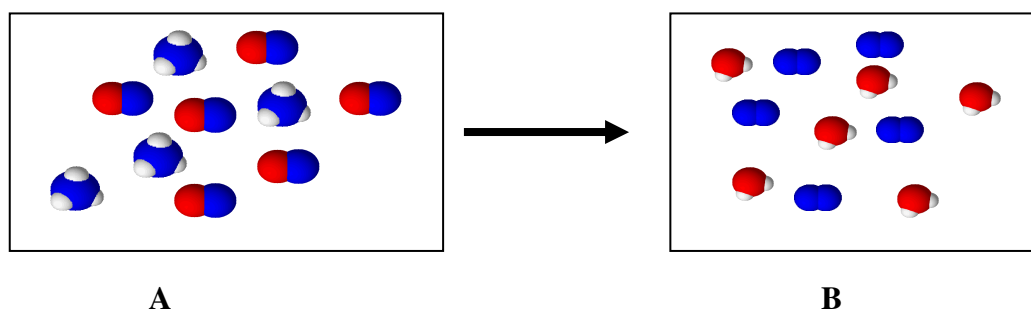
Any Comment or Suggestion?

Question 4

Ammonia () reacts with Nitrogenmonoxide () to form Nitrogen gas () and water ().

Consider the mixture of Ammonia and Nitrogenmonoxide in a closed container before (A) and after (B) the reaction has occurred. All reactants were used up during the reaction.

How many moles of each reactant were there if 13.7 moles of $N_2(g)$ is produced?



- 10.96 moles $NH_3(g)$ and 16.44 moles $NO(g)$
- 2.74 moles $NH_3(g)$ and 16.44 moles $NO(g)$
- 3.43 moles $NH_3(g)$ and 5.15 moles $NO(g)$
- 54.8 moles $NH_3(g)$ and 82.2 moles $NO(g)$

This item intends to measure the following:

- The student is able to convert particulate pictures (visual representations) into symbolic representation of a balanced chemical equation.
- The student is able to use the symbolic balanced equation to solve mathematical problems associated with stoichiometry.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 5

Balancing a chemical equation is achieved by:

- setting the coefficients equal to one and adjusting subscripts in the formulas
- adjusting the coefficients to the smallest possible whole number ratio
- adjusting the number of elements produced
- adjusting the formula of a compound.
- writing in appropriate coefficients to ensure mass balance then adjusting the coefficient to the smallest possible whole number ratio.

This item intends to measure the following:

- The student knows the theory behind balancing a chemical equation.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 6

When the equation: $\text{Cu}_2\text{O} + \text{CH}_4 \rightarrow \text{H}_2\text{O} + \text{Cu} + \text{CO}_2$ is correctly balanced the coefficient in front of the formula for copper (I) oxide (Cu_2O) is:

- a. 1
- b. 2
- c. 3
- d. 4
- e. none of the above.

This item intends to measure the following:

- The student knows how to balance a chemical equation.
- The student understands the meaning of subscripts and coefficients in a balanced chemical equation.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 7

How many moles of methane (CH_4) are required to produce one mol of copper (Cu) by the reaction given in **question 6**?

- a. 0.5 moles
- b. 1.0 moles
- c. 1.5 moles
- d. 0.25 moles
- e. 0.125 moles
- f. none of the above.

This item intends to measure the following:

- The student is able to take the knowledge of symbolic representation of atoms and molecules in a balanced chemical equation and transfer that knowledge to numerical methods to extract quantitative information e.g. identify the mole ratio and use the relevant method (algorithm) to solve the problem.

Do you agree that this item measures what it intends to measure?

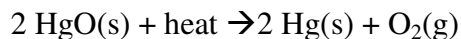
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 8

Joseph Priestley discovered oxygen in the eighteenth century by using heat to decompose mercury(II) oxide:



What mass of mercury (II) oxide would be required to produce 100. g of O₂.

- a. 100. g
- b. 200. g
- c. 627 g
- d. 677 g
- e. 1350 g
- f. 2700 g

This item intends to measure the following:

- The student is able to convert between grams and moles.
- The student is able to take the knowledge of symbolic representation of atoms and molecules in a balanced chemical equation and transfer that knowledge to numerical methods to extract quantitative information e.g. identify the mole ratio and use the relevant method (algorithm) to solve the problem.
- The student can determine the molecular mass of HgO from the information given on a periodic Table.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 9

What mass of calcium carbonate (CaCO_3) is needed to react completely with 50.00 mL of 0.383 M sulfuric acid (H_2SO_4) according to the following balanced chemical equation? $\text{CaCO}_3 + \text{H}_2\text{SO}_4 \rightarrow \text{CaSO}_4 + \text{CO}_2 + \text{H}_2\text{O}$

- a. 19.2 g
- b. 0.958 g
- c. 1.92 g
- d. 9.58 g
- e. 767 g
- f. 13.1 g

This item intends to measure the following:

- The student is able to convert volume and concentration units to moles.
- The student is able to directly use mole ratios in stoichiometry calculations.
- The student is able to convert moles to grams.
- The student can determine the molecular mass of CaCO_3 from the information given on a periodic Table.



Do you agree that this item measures what it intends to measure?

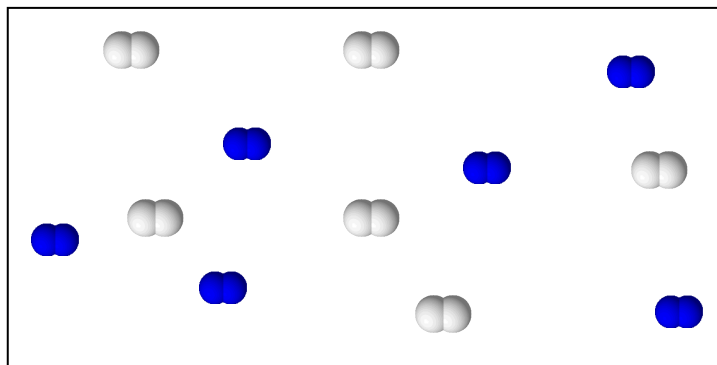
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 10

Nitrogen (N_2) and hydrogen (H_2) react to form ammonia (NH_3). Consider the mixture of N_2 () and H_2 () in a closed container as illustrated below:



In your opinion, the chemical reaction **stops** when:

- all the Nitrogen is used up.
- all the Nitrogen and all the Hydrogen are both totally used up.
- all the Hydrogen is used up.
- all the Nitrogen or all the Hydrogen is used up.
- I do not know.

This item intends to measure the following:

- The student is able convert particulate pictures (visual representations) into symbolic representation of a balanced chemical equation.
- The student is able to determine the limiting reactant.

Do you agree that this item measures what it intends to measure?

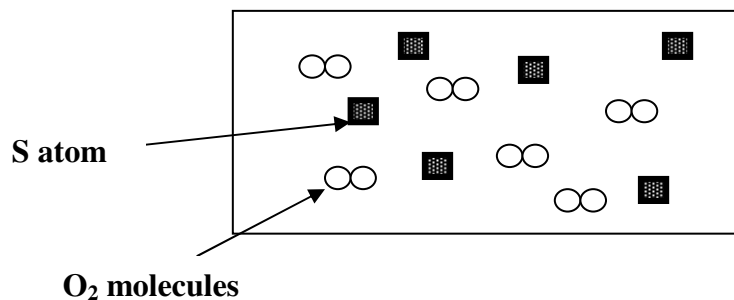
(cross the relevant option)

YES	NO
-----	----

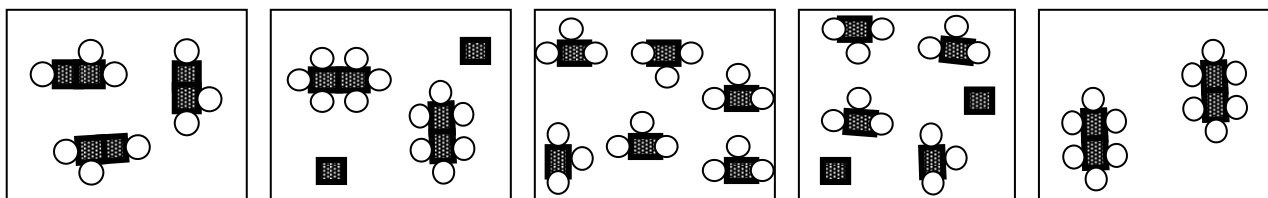
Any Comment or Suggestion?

Question 11

The diagram represents a mixture of S atoms and O₂ molecules in a closed container.



Which diagram shows the results after the mixture reacts as completely as possible according to the equation $2S + 3O_2 \rightarrow 2SO_3$?



a.

b.

c.

d.

e.

This item intends to measure the following:

- The student is able to analyse a reaction in which one reactant is present in a limited supply at the molecular level and can make predictions based on the balanced equation.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

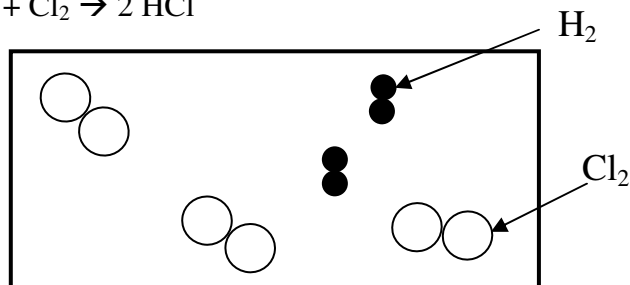
YES	NO
-----	----

Any Comment or Suggestion?

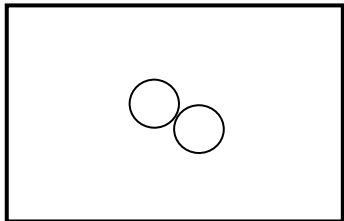
Question 12

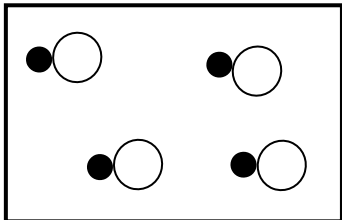
12.1 Here is a picture of a container with three Cl₂ molecules and two H₂ molecules. A chemical reaction occurs until the maximum amount of HCl has been produced.

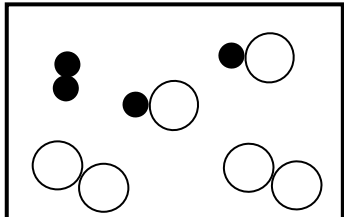
The reaction is $\text{H}_2 + \text{Cl}_2 \rightarrow 2 \text{HCl}$

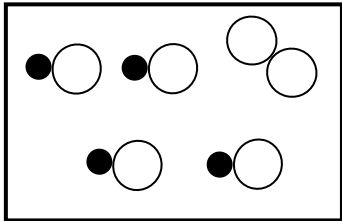


The picture of the container after the reaction looks like:

a. 

b. 

c. 

d. 

e. none of these

This item intends to measure the following:

- The student is able to analyse a reaction in which one reactant is present in a limited supply at the molecular level and can make predictions based on the balanced equation.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 13

13.1 Consider the following generic chemical reaction:



How many moles of B would you need to react completely with 5 moles of A?

- a. 1.2
- b. 1.5
- c. 2
- d. 3.3
- e. none of the above

This item intends to measure the following:

- The student is able to directly use mole ratios in stoichiometry calculations.

Do you agree that this item measures what it intends to measure?

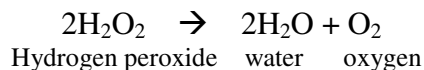
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 14

14.1 Hydrogen peroxide will decompose to form water and oxygen gas according to the following equation:

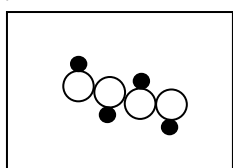


Use the following key for the diagrams:

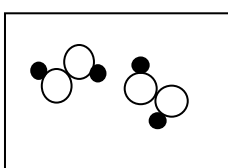
Oxygen ○

Hydrogen ●

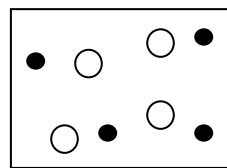
Which diagram is the best representation of the hydrogen peroxide before it decomposes?



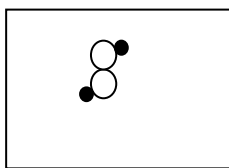
a.



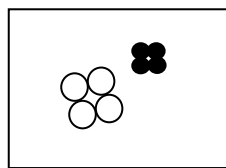
b.



c.



d.



e.

This item intends to measure the following:

- The student is able to interpret meanings of coefficients and subscripts in a particulate or visual representation.

Do you agree that this item measures what it intends to measure?

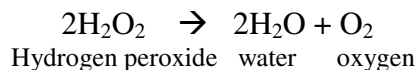
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 15

15.1 Hydrogen peroxide will decompose to form water and oxygen gas according to the following equation.

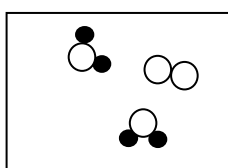


Use the following key for the diagrams:

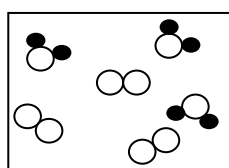
Oxygen ○

Hydrogen ●

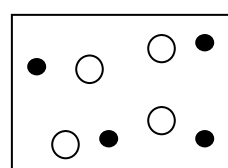
Which diagram is the best representation of the products after hydrogen peroxide decomposes?



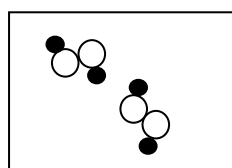
a.



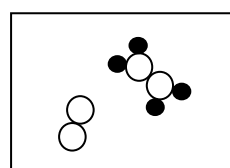
b.



c.



d.



e.

This item intends to measure the following:

- The student is able to analyse a reaction in which one reactant is present in a limited supply at the molecular level and can make predictions based on the balanced equation.
- The student is able to interpret the meanings of subscriptions and coefficients in a visual representation.

Do you agree that this item measures what it intends to measure?

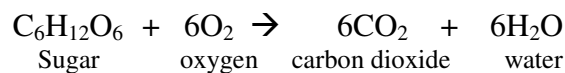
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 16

16.1 Your body reacts sugar with oxygen to form carbon dioxide and water according to the following chemical equation.



How many million oxygen atoms would be needed to react completely with one million sugar molecules?

- a. 3
- b. 6
- c. 9
- d. 12
- e. none of the above

This item intends to measure the following:

- The student understands when coefficients and subscripts are used in stoichiometry calculations.
- The student can distinguish between oxygen atoms and oxygen molecules.

Do you agree that this item measures what it intends to measure?

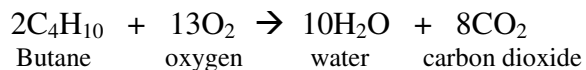
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 17

17.1 Butane is combusted completely with excess oxygen to form water and carbon dioxide.



The reaction yielded 1 mole of water. How many moles of carbon dioxide were produced?

- a. 0.8
- b. 1.25
- c. 4
- d. 8
- e. none of the above

This item intends to measure the following:

- The student is able to directly use mole ratios in stoichiometry calculations.

Do you agree that this item measures what it intends to measure?

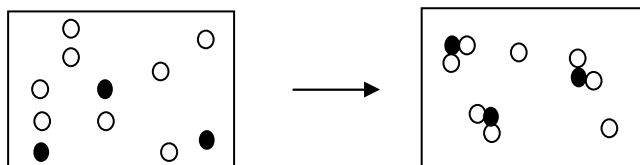
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 18

18.1 The reaction of element X (●) with element Y (○) is represented in the following diagram.



Which equation describes this reaction?

- $3X + 8Y \rightarrow X_3Y_8$
- $3X + 6Y \rightarrow X_3Y_6$
- $X + 2Y \rightarrow XY_2$
- $3X + 8Y \rightarrow 3XY_2 + 2Y$
- $X + 4Y \rightarrow XY_2$

This item intends to measure the following:

- The student is able to convert particulate pictures (visual representations) into symbolic representation of a balanced chemical equation.
- The student realises that excess reactants are not reported in a chemical reaction.
- The student knows that a reaction equation is expressed with the smallest possible whole numbers as coefficients.

Do you agree that this item measures what it intends to measure?

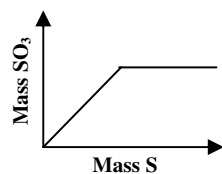
(cross the relevant option)

YES	NO
-----	----

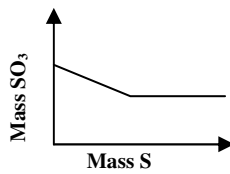
Any Comment or Suggestion?

Question 19

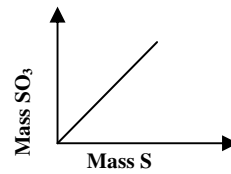
19.1 For the chemical reaction: $2S + 3O_2 \rightarrow 2SO_3$, which of the following graphs best represents the formation of SO_3 , if S is added indefinitely (or in excess) to a fixed amount of O_2 ?



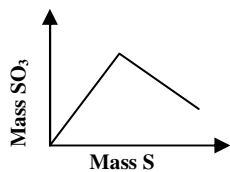
a.



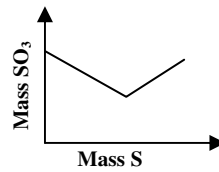
b.



c.



d.



e.

This item intends to measure the following:

- The student is able to use knowledge of symbolic representation of atoms and molecules in a balanced chemical equation to interpret graphical representation of a chemical reaction.

Do you agree that this item measures what it intends to measure?

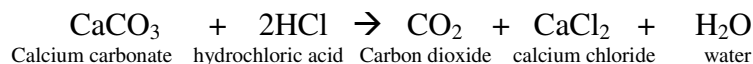
(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Question 20

20.1 Use the following equation:



If 14g of calcium carbonate react with 0.2 moles of hydrochloric acid, which reactant(s) determines the mass of CO₂ produced?

(the molecular weight of calcium carbonate is 100.087g/mole)

- a. CaCO₃
- b. HCl
- c. Any of the two reactants
- d. None of the two reactants
- e. Both CaCO₃ and HCl

This item intends to measure the following:

- The student is able to convert from grams to moles.
- The student understands that the limiting reactant limits the amount of product produced.
- The student is able to determine the limiting reactant.

Do you agree that this item measures what it intends to measure?

(cross the relevant option)

YES	NO
-----	----

Any Comment or Suggestion?

Finished!

Thank you for your participation!!

APPENDIX VI: QUESTIONNAIRE USED IN MAIN STUDY
QUESTIONNAIRE

**University of
Pretoria**

Faculty of Agricultural and
Natural Sciences

Chemistry department

Tel: (012) 420 4905

Email:

kgadi.mathabathe@up.ac.za

**Questionnaire for BSc Four Year Programme (BFYP)
students**

Thank you for being willing to complete this questionnaire. The purpose of this questionnaire is to establish a profile of the students in the BFYP programme participating in this study.



Accuracy of chemistry performance evaluation of BSc Four-year programme students: a case study

Instructions

It is important that you answer all the questions as honestly as possible.

Your answers to this questionnaire will be treated confidentially.

Please answer each question by **circling** the appropriate number in the shaded box.

The information will be treated confidentially.

Respondent's student number

1. Gender

Male	1
Female	2

2. Age

17 years	1
18 years	2
19 years	3
20 years	4
21 years	5
22 years	6
23 years	7
24 years	8
25 years	9
Other (specify)	10

Item	Strongly disagree	Disagree	Disagree somewhat	Undecided	Agree somewhat	Agree	Strongly agree
1. I intend to change my behaviours to create a good impression to others.	1	2	3	4	5	6	7
2. When I perform poorly in a test there are usually external circumstances that are to blame.	1	2	3	4	5	6	7
3. I am satisfied with my performance at University.	1	2	3	4	5	6	7
4. Studying makes me feel good.	1	2	3	4	5	6	7
5. I do not set goals that are hard to reach, because failure is painful.	1	2	3	4	5	6	7
6. I try to modify my behaviours to give good images to others.	1	2	3	4	5	6	7
7. When I get poor results in a test I just want to get rid of the script and not look at it again.	1	2	3	4	5	6	7
8. It is important to me to give a good impression to others.	1	2	3	4	5	6	7
9. My studies provide me with much satisfaction.	1	2	3	4	5	6	7
10. Doing other things (spending time with friends, socialising) are more appealing than studying Chemistry.	1	2	3	4	5	6	7



Accuracy of chemistry performance evaluation of BSc Four-year programme students: a case study

Item	Strongly disagree	Disagree	Disagree somewhat	Undecided	Agree somewhat	Agree	Strongly agree
11. I try to create the impression that I am a "good" student.	1	2	3	4	5	6	7
12. I do not want my friends to know about it when I have failed a test.	1	2	3	4	5	6	7
13. Often I get poor results in courses because the teacher has failed to make them interesting.	1	2	3	4	5	6	7
14. There are other subjects that would be more fulfilling than Chemistry.	1	2	3	4	5	6	7
15. I like to present myself to others as being a clever person.	1	2	3	4	5	6	7
16. It is unrealistic to expect good grades for maths and science because these are hard subjects.	1	2	3	4	5	6	7
17. I have spent many hours on my studies.	1	2	3	4	5	6	7
18. I am sensitive to the impression about me that others have.	1	2	3	4	5	6	7
19. It is likely that I will choose not to finish my studies.	1	2	3	4	5	6	7
20. I set difficult goals for myself so people can see I am serious about my work.	1	2	3	4	5	6	7

Item	Strongly disagree	Disagree	Disagree somewhat	Undecided	Agree somewhat	Agree	Strongly agree
21. Sometimes my success on exams depends on luck.	1	2	3	4	5	6	7
22. I have put a lot of work and effort into my studies that I would lose if I were to give up.	1	2	3	4	5	6	7
23. It is important to me that others see me as being the best in my class.	1	2	3	4	5	6	7
24. My performance does not reflect my ability: I was just unlucky not to be taught by a better teacher.	1	2	3	4	5	6	7
25. I am committed to finishing my studies.	1	2	3	4	5	6	7
26. It's important to me that others think I work hard.	1	2	3	4	5	6	7
27. It is better to expect poor results and to be surprised than to be disappointed when your expectations are not met.	1	2	3	4	5	6	7
28. I feel very involved with my studies.	1	2	3	4	5	6	7
29. I work harder than I normally do when I know someone is watching me.	1	2	3	4	5	6	7
30. If I did not have to take Chemistry, I would be better off.	1	2	3	4	5	6	7

APPENDIX VII

Consent Form

I understand that:

1. The purpose of this study is to determine factors that influence metacognitive judgements made about performance in a chemistry test, the influence of inflated metacognitive judgements on subsequent performance and the influence of teaching on metacognitive judgements and performance of students in the BSc Four Year Programme.
2. As part of this study I will have to participate in more than one activity i.e. writing a chemistry test and answering a questionnaire.
3. Any personal information about me that is collected during the study will be held in the strictest confidence and will not form part of my permanent record at the university.
4. I am not waiving any human or legal rights by agreeing to participate in this study.
5. My participation in this study is completely voluntary.

I verify, by signing below, that I have read and understand the conditions listed above.

Signature : _____

Date : _____



APPENDIX VIII: LETTER OF APPROVAL FROM THE UNIVERSITY OF PRETORIA'S ETHICS COMMITTEE



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

18 September 2008

Faculty of Natural and Agricultural Sciences
Ethics Committee
Tel: +2712 420 4107
Fax: +2712 420 3290
E-mail: ethics.nas@up.ac.za

Dr M Potgieter
Department of Chemistry
University of Pretoria
Pretoria
0002

Dear Dr Potgieter

RE: EC080818-029 Causes of error in self assessment: BSc Four Year Programme students assessing their performance in a stoichiometry test.

Your project conforms to the requirements for the Ethics Committee: Faculty of Natural and Agricultural Sciences.

Kind regards
Prof NH Casey
(Chairman: Ethics Committee, Faculty of Natural and Agricultural Sciences)