Chemistry Education Research and Practice

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the RSC Publishing peer review process and has been accepted for publication. CERP is free to access thanks to sponsorship by the RSC's Education Division.

Accepted Manuscripts are published online shortly after acceptance, which is prior to technical editing, formatting and proof reading. This free service from RSC Publishing allows authors to make their results available to the community, in citable form, before publication of the edited article. This *Accepted Manuscript* will be replaced by the edited and formatted *Advance Article* as soon as this is available.

To cite this manuscript please use its permanent Digital Object Identifier (DOI®), which is identical for all formats of publication.

More information about *Accepted Manuscripts* can be found in the **Information for Authors**.

Please note that technical editing may introduce minor changes to the text and/or graphics contained in the manuscript submitted by the author(s) which may alter content, and that the standard **Terms & Conditions** and the **ethical guidelines** that apply to the journal are still applicable. In no event shall the RSC be held responsible for any errors or omissions in these *Accepted Manuscript* manuscripts or any consequences arising from the use of any information contained in them.

Chemistry Education Research and Practice

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

FULL PAPER

Metacognitive monitoring and learning gain in foundation chemistry

Kgadi C. Mathabathe,*[a] and Marietjie Potgieter[b]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

The ability to make realistic judgements of one's performance is a demonstration of the possession of strong metacognitive skills. In this study we investigate the relationship between accuracy of self-evaluation as an expression of metacognitive skill, and learning gain in stoichiometry. The context is an academic development programme at a South African University, offered for under-prepared students enrolled for science and engineering. These students generally exhibit unrealistically high levels of confidence in performance and this could potentially place them at risk by negatively affecting decisions regarding time management and self-regulation. We investigated whether overconfidence before instruction is corrected upon exposure to teaching. A three-tier stoichiometry test was used to collect qualitative and quantitative data before and after instruction. Findings indicate that the majority of the students were overconfident in the evaluation of their performance in both the pre- and posttests. Overconfidence was not a debilitating disposition when demonstrated in the pretest provided that it was corrected during teaching and learning. The most vulnerable students were those that judged their performance or lack thereof realistically in the pretest but became overconfident during the teaching and learning of stoichiometry. Our results suggest that under-prepared students are slow to develop accurate metacognitive monitoring skills within a classroom environment that did not include instruction focused on the development of such skills. We recommend a proactive and constructive response by educators which may reduce the incidence of failure and preserve the positive contribution of confidence, albeit excessively positive.

Keywords: learning gain, foundation chemistry, metacognition, stoichiometry

[a] *University of Pretoria, Department of Science, Mathematics and Technology Education, Pretoria, South Africa. Fax: +27865128641; Tel: +27 12 420 2758; E-mail: kgadi.mathabathe@up.ac.za*
[b] *University of Pretoria, Department of Chemistry, Pretoria, South Africa. Fax: +27 12 420 5441; Tel: +27 12 420 6472; E-mail:marietjie.potgieter@up.ac.za*

† Electronic Supplementary Information (ESI) available: The Stoichiometry test instrument. See DOI: 10.1039/b000000x/

Introduction

The need for the teaching of metacognitive skills is one of the main implications of research on the teaching and learning of science that has emerged during the past three decades, according to the editors of a recent monograph on metacognition in science education (Zohar and Dori, 2012). A small number of studies have been reported in this journal over the past five years focussing on the development of metacognition in problem solving (Cooper *et al*., 2008) and in the chemistry laboratory (Sandi-Urena *et al*., 2011). Metacognition is generally accepted to consist of both knowledge of cognition and regulation of cognition (Flavell, 1979), where knowledge of cognition refers to the ability to monitor learning or evaluate performance. In this study we explore the relationship between metacognitive monitoring and learning gain in stoichiometry.

Accurate performance evaluation is critical in decisions on the time required to study for a specific course, what study methods to employ as well as what topics to give the most attention to (Grimes, 2002; Nowell and Alston, 2007). Assessment in the form of test-taking plays a vital role in the development of important metacognitive skills such as accurate performance evaluation. Test-taking is a challenging academic requirement but it provides a valuable opportunity for students to learn how to regulate their own learning in a certain domain, to better monitor their performance in that domain and to make valid attributions of their failures and successes (Carvalho, 2009).

Previous studies have reported the general occurrence of bias in performance evaluation in the form of overconfidence as well as the potentially negative consequences that it may have on academic success (Ochse, 2003; Potgieter *et al*., 2007). In the current study, overconfidence is defined as inflated levels of confidence that a student displays with regard to the accuracy of answers in a test. The context of this study is tertiary chemistry in an academic development programme at a South African university, specifically learning gains achieved in stoichiometry, a core content topic in the first-year syllabus. The majority of students in this programme come from disadvantaged backgrounds with inadequate preparation for tertiary science. Many of these students have been found to exhibit exaggerated levels of confidence in their mastery of foundational concepts and skills in chemistry (Potgieter *et al*., 2007). This study is an investigation of how accurately students in this programme evaluate their performance during test-taking and the influence of teaching on accuracy of performance evaluation.

Literature review

When we are prompted to make judgements on how we perceive our ability, how well we know something or how well we have performed a particular task, the judgements we report are called metacognitive judgements (Rosenthal, 2000; Dunlosky *et al*., 2005; Koriat and Bjork, 2005; Fernandez-Duque and Black, 2007). Metacognitive judgements have been extensively investigated partially due to the fact that mastery of the skill of accurately making them may result in the effective management of self-regulated study, which is necessary in a tertiary environment where an autonomous approach to studying is required (Dunlosky *et al.*, 2005).

Metacognition refers to the knowledge and experiences we have about our own cognitive processes (Flavell, 1979). Metacognition consists of metacognitive knowledge and metacognitive experiences. 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) that an individual performs to control cognitive activities to ensure that cognitive goals are met. 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. All the facets of metacognitive knowledge are necessary for one to self-regulate one's thinking and learning effectively (Hartman, 2001). 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. Making accurate judgements about one's performance and competence is a metacognitive process that people can use to regulate their behaviour towards successful learning (Hacker *et al.,* 2008). Students who can accurately assess the effectiveness of their learning strategies and their understanding of concepts in a particular subject area should be able to make informed decisions as to whether and when to intensify or redirect their studying for a test (Hacker *et al.,* 2008). Metacognitive skills should therefore differentiate a novice learner from an expert learner. An expert learner knows how to learn and also knows which strategies work best (Grimes, 2002; Nowel and Alston, 2007).

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). Although the two paths have remained separate, contemporary research was introduced to the construct of metacognition through the publication of Nelson and Narens (1990)'s 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). Nelson and Narens (1990) identified several types of metacognitive judgements namely ease-of-learning judgements, judgements of knowing or judgements of learning, feeling-of-knowing judgements and confidence judgements. The theoretical framework of Nelson and Narens (1990) describes three stages, namely the acquisition, retention and retrieval stages during which metacognitive judgements are made. 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. In our study we were particularly interested in the judgements made during test taking. In addition to exaggerated confidence judgements made based on feelings or incorrect information, several factors associated with bias in performance evaluation or monitoring, particularly overconfidence, have been identified and categorised by Carvalho (2009) as personal, task-related and environmental factors. Task related factors include the lack of skill (Kruger and Dunning, 1999; Ehrlinger. 2008); properties of the task (Lichtenstein and Fischhoff, 1997); format selected for evaluation (Carvalho, 2009) and the quality of feedback received (Carter and Dunning, 2008). Personal factors include factors such as the tendency to rely on chronic self-views to evaluate performance (Ehrlinger 2008); the need for self-protection and self-enhancement (Gramzow *et al*., 2003); theories of intelligence that respondents adhere to (Ehrlinger, 2008); personality traits (Campbell, Goodie and Foster, 2004) and gender (Beyer and Bowden, 1997). Some of these factors were investigated in this study but will be reported in a separate paper.

Metacognition is but one component of self-regulated learning (Schraw *et al., 2006*). Self-efficacy, which is a subcomponent of the motivation component of self-regulated learning, is described as the extent to which an individual is confident that he or she can perform a specific task (Bandura, 1997). Self-efficacy is an important aspect of self-regulated learning as it affects the extent to which learners engage and persist at challenging tasks (Schraw *et al.*, 2006). Bandura (1997) argued that self-efficacy judgements that are slightly higher than actual accomplishments serve to increase individual's effort and persistence. However, self-efficacy should not be confused with confidence statements made after a task has been completed. Self-efficacy is an expression of confidence about a task that must be performed. Confidence judgements that are made *ex post facto* are an expression of metacognitive monitoring. While self-efficacy has been found to be a strong predictor of academic performance (Britner and Pajares, 2006), overly confident evaluation of performance is not.

Accuracy of metacognitive judgements has been studied extensively in cognitive psychology and educational psychology over the past three decades. In general, the perceptions people hold of either their overall ability or specific performance tend to be correlated only modestly with their performance, with better accuracy being correlated with better performance (Bol & Hacker, 2001). Mabe and West (1982) surveyed 55 self-evaluation studies with a combined population of 14,811 subjects across a variety of domains and found the average correlation between self-estimates and actual performance to be only 0.29. While performance estimates have been shown to be generally unrealistic, overestimation of performance was particularly prevalent (for a recent review, see Dunning, 2005). Despite much poorer performance, weaker students have been found to be particularly overly optimistic about the correctness of their answers in tests (Carvalho, 2009; Carvalho & Yuzawa, 2001; Kruger & Dunning, 1999; Kennedy, Lawton & Plumlee, 2002, Potgieter *et al.*, 2007). These students display poor judgement in the sense that many of the answers which they expected to be correct are indeed wrong.

In studies on accuracy in self-evaluation a number of different methods have been reported to quantify the extent of inaccuracy. These methods can be described as either direct or indirect probes to elicit an expression of perceived performance, where direct methods refer to studies where respondents were asked to predict their total score before or after completing the task. The predicted score is then compared with actual performance on the test to obtain a calibration score (*e.g.* Dunning *et al*., 2003; Hacker *et al*., 2008). In indirect methods respondents are asked to report their confidence in their performance on each test item immediately after completion of the task. Confidence judgments are reported on a Likert scale and are interpreted as expressions of the likelihood of getting the answer correct. The average of confidence ratings over all test items is interpreted as perceived performance.[a] The statistic used in many studies on overconfidence in psychology is called the bias score which is the difference between average confidence ratings over all test items and the percentage of correct responses on the test (Pallier *et al*., 2002; Schraw, 2009). Confidence ratings are typically grouped into discrete categories on a Likert scale, but may also be unstructured. Schaefer *et al*. (2004) studied overconfidence using two-option fixed-choice questions on general knowledge followed by a seven category rating of confidence in the answer provided starting from 50% (50 – 52%, 53 – 60%, 61 – 70%, 71 – 80%, 81 – 90%, 91 – 97%, 98 – 100%). In his study, Carvalho (2009) asked psychology majors (N = 129) to indicate their confidence in the accuracy of their responses on a Likert scale ranging from 0 to 100 per cent. He reported that these students were less accurate in performance evaluation on multiple-choice test items than on short answer questions. We have found that direct methods for eliciting perceived performance were too insensitive a probe in our context and opted for requesting a confidence rating after every item on an eleven point Likert scale with 10% intervals between 0 and 100%.

Ochse (2003) conducted a study at a South African university where accuracy of self-evaluation was used to categorise third-year psychology students as overestimators, realists or underestimators in order to investigate the academic success of the different groups. Before their final examination, students were asked to indicate the 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 examination, actual scores were compared with expected scores. A difference of less than nine percentage points between actual and expected scores was considered a realistic estimate. Students who overestimated their actual score by nine or more percentage points were categorised as overestimators and those who underestimated their actual score by nine or more percentage points were categorised as underestimators. Overestimators, on the whole, expected higher marks than realists and underestimators, were significantly more confident about the accuracy of their expected scores, and perceived themselves to have higher ability, but however obtained the lowest scores of the three groups. On average the overestimators failed the course.

In studies with psychology students, Dunning and coworkers

(Kruger and Dunning, 1999; Dunning *et al*., 2003) have consistently found that the mismatch between perceived and actual performance was the largest for poor performing students, that this mismatch was smaller for students with better performance and that the best performing students were not only best calibrated but often slightly underestimated their performance. They have called this finding the dual burden of incompetence, to be unskilled and unaware of it (Kruger and Dunning, 1999), and argued that one of the reasons for overconfidence is the fact that the same knowledge and skills that are required for good performance are also required for accurate self-evaluation of performance (Dunning *et al.*, 2003). Making informed and accurate judgements of performance in a test requires the learner to have a good knowledge and understanding of the content and metacognitive knowledge of task and strategy variables with regard to that content or subject matter. How else can one judge if the problem was solved successfully when one does not even recognise which approach is required to solve the problem in the first place? For this study we chose a chemistry topic, stoichiometry, for which prior knowledge was known to be limited, so that there would be a large scope for improvement, in order to see whether growth in knowledge and understanding would be accompanied by improvement in the accuracy of metacognitive judgments. Success in solving stoichiometry problems requires representational competence, formal reasoning and being able to correctly apply multistep mathematical operations (Johnstone, 1991; Huddle and Pillay, 1996). Previous research has shown that students are able to solve one-step problems such as finding the molar mass of a compound or calculating the number of moles corresponding to a given mass of the compound, but they are unsuccessful when a problem requires the stringing together of such steps to solve a more complex task (Lazonby *et al*., 1985; Huddle and Pillay, 1996). Herron (1990) suggests that failure on the more complex tasks may be due to poorly developed metacognitive skills that govern the organisation of work, sequencing of tasks, and checking of results. For example, when a learner is asked to solve a stoichiometry problem in chemistry, the learner must be able to:

a. recognise and understand what is being asked or what is expected of him or her;
b. 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);
c. perceive and acknowledge when he or she cannot solve the problem;
d. reflect on the reasons why he or she cannot solve the problem and make changes to his or her strategies in order to improve.

The problem arises when there is a mismatch between students' perception of mastery and actual performance. Overconfidence could potentially hamper learning by negatively affecting decisions regarding the regulation and control of cognition (Schraw *et al*., 2006). However, overconfidence could also motivate students to take on challenges that would otherwise be too imposing (Pajares, 1996; Zimmerman, 2000; Ehrlinger, 2008). This dichotomy has prompted us to ask the question whether overconfidence is necessarily undesirable within the context of first year chemistry in an academic development programme.

The context of this study

This study was conducted in partnership with educational psychology to investigate the relationship between accuracy of self-evaluation as an indication of metacognitive skill, and learning gain in stoichiometry, and whether overconfidence before instruction is reduced upon exposure to teaching of the topic. In addition we wanted to investigate the underlying reasoning informing metacognitive judgments on content mastery. The study was therefore guided by the following questions:

1. How accurately do BSc Four-year programme (BFYP) students evaluate their performance in a stoichiometry test?
2. How does accuracy of performance evaluation change upon teaching of the topic?
3. How does accuracy of performance evaluation vary as a function of item type?
4. How is learning gain associated with shifts in accuracy of performance evaluation?

Quantitative data was required to measure performance and learning gain, and to determine accuracy of performance evaluation. Qualitative data, on the other hand, was collected to enrich our understanding of the metacognitive factors underlying judgements of performance (Nelson and Narens, 1990). The current article reports only the analysis of quantitative data whereas the qualitative work will be reported later.

Methods

An embedded experimental mixed methods approach was followed to answer the above-mentioned research questions (Maxwell and Loomis, 2003; Harrits, 2011; Leech & Onwuegbuzie, 2011).

Participants

The sample consisted of 91 participants who voluntarily participated in the study and for whom complete records were obtained, *i.e.* 35 male students (38%) and 55 female students (60%), and one record with gender information omitted. The ages of participants ranged between 17 and 25 years (M = 19). These students shared commonalities in that they were all taught stoichiometry by the same lecturer in large group lectures and they completed compulsory computerised quizzes on the topic. In addition, they attended three fifty minute long small group tutorial sessions per week over a period of three weeks, where they had plenty of problem-solving and feedback opportunities. Even though they had different lecturers with different teaching styles in the small group sessions, the lecturers worked collaboratively in terms of the material used in these sessions as well the quantity, format and content of tests and tasks given to the students. Stoichiometry is taught extensively at high school

level, specifically at grades 10 and 11. By the time these students were in their first year undergraduate courses after completing grade 12, it had been a year since they received instruction on the topic. Participants were duly informed of the objectives of the study and were promised complete anonymity and that the results obtained in the study would not affect their grades in any way. The study was repeated the following year with a sample of 300 students from the new cohort of students enrolled in the BFYP.

Research setting

The study took place in the second semester of the university's academic development programme, the BSc Four-year programme (BFYP). The minimum duration of BSc programmes at the particular South African university is three years but in this programme an extra year is added at the foundational level to address under-preparedness of incoming first year students. The study received ethical clearance from the institution where it was conducted.

Data collection instrument

Pretest data were collected before a 3-week period of instruction on stoichiometry and posttest data after formal instruction was completed and ample opportunity was provided for guided and unguided problem-solving. The posttest was used for summative purposes about which students were informed beforehand. Pretest and posttest data were collected through a 20-item test instrument. Each item in the instrument consisted of three tiers to enable the simultaneous collection of both quantitative and qualitative data. The first tier consisted of a multiple choice question on stoichiometry followed by the second tier, a Likert scale from 0% to 100% with 10% intervals, on which participants were asked to indicate their confidence in the accuracy of the chosen response in the first question. Lastly, in the third tier participants were asked to explain their choice of confidence indicators. The common structure of test items is demonstrated in Figure 1.

1.1 Given the equation 3A + B → 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?

Fig. 1 Example of a three-tier test question

The test instrument was piloted and refined. The face and content validity were checked by tertiary educators and Grade 12 teachers and small changes were made as recommended by them. Cronbach alpha coefficients were calculated with respect to content items to determine the reliability of the test items. Item 2 was excluded *post hoc* based on validation feedback and poor item performance in the pretest. Posttest Cronbach's alpha coefficients of 0.69 and 0.70 were obtained in successive years of implementation. The test instrument is included as Appendix I.

The test instrument comprised nineteen items of a range of difficulties chosen and adapted from current literature and first year chemistry textbooks. Eight items measured procedural knowledge, formal reasoning and numeric problem-solving skills using multistep mathematical operations. Two items assessed declarative knowledge. To probe for conceptual understanding, eight items incorporated sub-microscopic representations or particulate drawings of atoms and molecules. One item assessed students' ability to match the symbolic representation of a balanced chemical equation with a graphical representation of the reaction.

Results and discussion

The sum of all correctly answered items on the stoichiometry test was used as a measure of performance with correct answers scored one (1) and incorrect answers zero (0). The confidence judgement ratings reported on a scale of 0% to 100% per item were used to determine the average of confidence judgements for each individual student and for the sample as a whole. Table 1 below presents a summary of the results in terms of performance and confidence scores obtained in the pre- and posttests for the first year of the study. The mean performance improved from 37% in the pretest to 51% in the posttest as would be expected, while the mean of average confidence scores increased by a similar margin from 63% in the pretest to 76% in the posttest. *P-values* ($p < 0.05$) confirmed that there was a statistically significant difference between the pre- and posttest performance and between pre- and posttest average confidence scores.

Table 1 Summary of student performance and average confidence scores in the pre- and posttest

	Performance (test scores, %), $N = 91$		Average confidence scores (%), $N = 91$	
	Pretest (19 items)	Posttest (19 items)	Pretest	Posttest
Mean score	7.0 (37%)	9.6 (51%)	63.0	75.7
Standard deviation	2.9 (15%)	3.4 (18%)	17.4	13.5
Minimum	2 (11%)	3 (16%)	16.3	40.5
Maximum	15 (79%)	18 (95%)	94.7	99.5

Test scores out of a total mark of 19 were converted to a percentage to represent *actual performance*. The average confidence scores represent the *perceived performance* of the students. The comparison between the actual and perceived performance of students are presented in Figures 2 and 3 for the pretest and posttest, respectively, in terms of the mean values for confidence ratings and test scores for students per performance quartile. This method of analysis and presentation corresponds with that of Dunning *et al.* (2003) and enables direct comparison

of findings. The results presented in Figures 2 and 3 show that students in our sample overestimated their performance before and after instruction. Our results corroborated with those of Dunning *et al.* (2003) in that the students in the bottom quartile misjudged their performance by the biggest margin. One would have expected students to judge their performance better after instruction, but this happened only to a limited extent for the third and top performance quartiles. The mismatch between actual and perceived performance increased marginally for the bottom quartile in the posttest. Dunning *et al.* (2003) reported that the gap between actual and perceived performance narrowed as performance improved and a small underestimation of performance was demonstrated by the top quartile. This did not happen in our case where there was still a 9% overestimation of performance for the top quartile in the posttest.

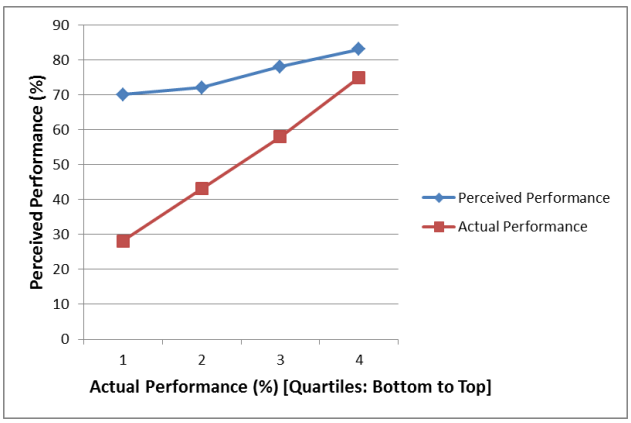

Fig. 2 Perceived *versus* actual pretest performance of BFYP students in the four performance quartiles

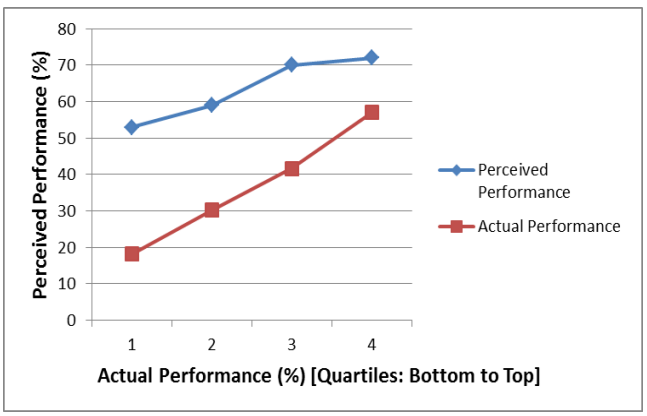

Fig. 3 Perceived *versus* actual posttest performance of BFYP students in the four performance quartiles

The accuracy of judgements made in the pre- and posttest (Research questions one and two)

It is clear from the results presented in Table 1 and Figures 2 and 3 that the students were unrealistic in their judgement of performance both in the pretest and the posttest. Despite getting on average only 37% of answers correct in the pretest, they were 63% certain that the answers were correct. Similarly, despite achieving a mere 51% in the posttest they were on average 76% certain of a correct answer. These results provide a general answer to the first two research questions. On average, the performance improved after teaching of stoichiometry but the accuracy of performance evaluation did not. However, a wealth of information is hidden behind mean values which requires further unpacking.

Accuracy of performance evaluation as a function of item type (Research question three)

Test items were classified according to the problem solving skills they required. Shifts in accuracy of judgement per item after instruction were also investigated. Items were categorised into four types, namely type A requiring multi-step mathematical operations, type B requiring declarative knowledge, type C requiring representational competence, and type D requiring interpretation of graphs. Type A was further divided into two subcategories, types A1 and A2. Type A1 represented all the items that required *simple* multistep mathematical operations to solve while Type A2 represented all questions that required the use of *complex* multistep mathematical operations. Accuracy of judgement as represented by the difference between *perceived* and *actual* performance was plotted for each item against item difficulty as reflected by average performance on that item. The results are shown in Figures 4 and 5 for the pre- and posttests, respectively. Different symbols are used to indicate item types.

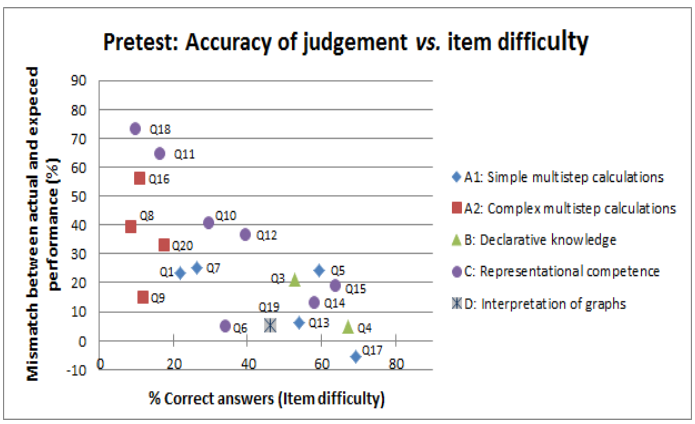

Fig.4 Accuracy of judgement *versus* item difficulty for the pretest

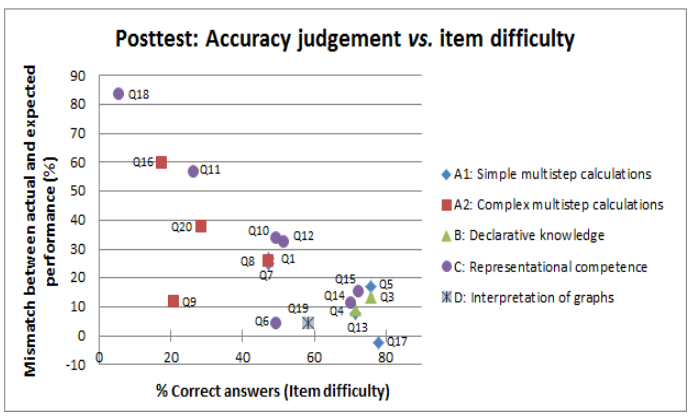

Fig.5 Accuracy of judgement *versus* item difficulty for the posttest

The first observation that can be made from Figures 4 and 5 is that accuracy of judgement does not seem to depend on item type. All five item types are represented in the band where the mismatch between actual and perceived performance is no more than 15 - 20%, *i.e.* where the judgement is fairly accurate. Similarly, three of the five item types are present in the band where judgement is poor or very poor, *i.e.* where the mismatch between actual and perceived performance is more than 20%. There were not enough items of types B and D in the instrument to be able to make general inferences about tasks requiring declarative knowledge or interpretation of graphs. The second observation is that good performance is accompanied by accurate judgement, but the reverse is not true. Students seem to have been aware of task properties associated with items 6, 9 and 19 in both tests and realised that they were unlikely to have answered them correctly. The third observation is that performance increased from the pretest to the posttest on all items except Q18, but accuracy of judgement did not. The shift in the position of items on the graphs from Figure 4 to Figure 5 is horizontal and not downwards to the right as one would have wished. The error in judgement in the posttest is therefore not reflecting lack of exposure or practice, but rather factors inherent in the task required for each item.

We now turn to the relationship between the nature of task properties and accuracy of performance judgment. For this purpose we compare item 9 with items 11, 16, 18 and 20. The difficulty of these items was comparable, with performance varying between 5% and 29%, but accuracy of judgment differed greatly (12% – 83%). Both items 9 and 20 entailed conversion of units and multistep calculations, but item 9 required additional steps and concept integration which challenged students beyond what they were exposed to during regular classroom practice. Poor performance on item 9 was judged accurately, but not item 20, presumably because students were daunted by the challenges of item 9, but overconfident on item 20 where they made calculation errors that went unnoticed. Item 18 was deceptively simple; students understood the task, but lacked or failed to apply procedural knowledge regarding accepted notation for chemical reactions. Items 11 and 16 required interpretation of coefficients and subscripts in a balanced reaction expressed in schematic or symbolic form. Confusion between coefficients and subscripts is a misconception that is well documented (*e.g.* Huddle and Pillay, 1996). Misconceptions exist due to the inadequacy of mental models and have been suspected to give rise to overconfidence (Hasan *et al.*, 1999), presumably because they create the illusion of knowing (Rozenblit and Keil, 2002). An illusion of knowing will not be conducive to accurate metacognitive monitoring. A similar analysis was done on a second group of items where performance ranged between 47% and 58%. Students judged their performance well on two items with an unfamiliar design (Q6 and Q19), but less accurately on the other five items which required either multistep calculations or contained sub-microscopic representations which were designed to reveal the presence of misconceptions. We concluded that misconceptions and the lack of mathematical skills are likely to compromise accuracy of performance judgements. Our findings resonate with those of a previous study where inflated confidence levels in mechanics have also been shown to be associated with the presence of misconceptions and lack of problem solving skills (Potgieter *et al*., 2010).

Categorisation of students in terms of accuracy of performance evaluation

The students in our sample were categorised individually based on their demonstrated accuracy of performance evaluation both in the pretest and the posttest. The difference between the actual and perceived performance values was regarded as a measure of accuracy of judgement. It is to be expected that students would make an error in judgement at least in the pretest because they are poorly prepared for the topic of stoichiometry and some of the items require several steps of calculations or analytical thinking. Taking into consideration the difficulty of the topic, the level of preparedness of the students in our sample and the format of the test we set an "acceptable" margin of error beyond which we judged the consequences of poor self-calibration to be too serious in terms of risk of failing. We set this margin of error as the equivalent of three incorrect judgements out of 19 answers, which translates into a judgement error of 15.8%. For example, 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 acceptable error margin of 15.8%. Subjects whose average confidence scores exceeded their test scores by more than 15.8% were labelled as overconfident (OC). The realistic group (R) were students with a difference between actual and perceived performance between 15.8% and -15.8% (-15.8% and 15.8% included). Students whose test scores exceeded their average confidence scores by more than 15.8% were labelled as under-confident (UC). Using these criteria we were able to categorise BFYP students in terms of how accurately they assessed their performance in a stoichiometry test before and after instruction.

The "acceptable" margin of error that we have chosen for the purpose of categorisation is a heuristic decision made within the specific context of chemistry students in an academic development programme. It is widely recognised that metacognitive skills are poorly developed in weaker students (Kruger and Dunning, 1999; Dunning *et al.*, 2003, Carvalho, 2009). Our results indicated that students in the top performing quartile of the posttest still overestimated their performance by an average margin of 9% as was shown in Figure 3. The decision was further informed by the fact that stoichiometry was assessed in this study with multiple-choice test items where calibration is notoriously difficult (Carvalho, 2009), especially for students with low metacognitive ability. The results are shown in Table 2.

The number of students who were overconfident in their judgment remained fairly constant, *i.e.* from 69% in the pretest to 71% in the posttest. This means that approximately 70% of our sample 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. Students who were realistic in their judgement decreased marginally from 31% in the pretest to 26% in the posttest. Only 2% of the students became under-confident in the posttest. There are no clear gender trends

in the accuracy of self-evaluation, neither in the pre- or posttest.

Table 2 Categories of students in terms of accuracy of performance evaluation in the pre- and posttest

† *OC – Overconfident, R – Realist, UC – Under-confident*

Category†	PRETEST			Category†	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
Total	91‡	35	55	**Total**	91‡	35	55

‡ *One record with gender information omitted.*

Despite the fairly stable division of accuracy of judgment depicted in Table 2 there were shifts of students who were able to show an improvement in terms of accuracy in performance evaluation after instruction and those whose ability to do so deteriorated. Students were then categorised based on pre-post accuracy of performance evaluation. The results are reported as a two-way frequency table in Table 3.

Five subgroups were generated and labelled according to pre- and posttest categories as OC-OC (n = 50), OC-R (n = 13), R-R (n = 11), R-OC (n = 15) and the R-UC (n = 2). The majority of students showed no improvement in their accuracy of performance evaluation (OC-OC subgroup, 55%). The number of students who acquired the skill of reporting accurate self-evaluations of their performance (OC-R subgroup, 14%) was similar to the number of students who became overconfident of their performance upon exposure to teaching (R-OC subgroup, 16%).Eleven students were realistic in their judgments both before and after instruction (R-R subgroup, 12 %). Only two of the 28 students who were realistic in their pre-test performance evaluation became under-confident in the posttest (R-UC subgroup, 2%).

Table 3 Shifts in accuracy of performance evaluation 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**

¶OC – Overconfident, R – Realist, UC - Under-confident

Learning gain associated with shifts in accuracy of performance evaluation (Research question four)

Having separated the students into five subgroups based on the accuracy with which they evaluated their performance in the pre- and posttests, we wanted to investigate the relationship beween accuracy of performance evaluation and learning gain in stoichiometry (research question four). Such results would indicate whether accuracy of judgment as a metacognitive skill is a desired attribute for the learning of chemistry, specifically in stoichiometry. Research question four was investigated for four subgroups. The fifth subgroup, R-UC, was too small for meaningful inferences to be made. 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 prerequisite 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.

Learning gain is a variable which provides a measure of the extent of improvement in performance but it must be normalised against scope for improvement in order to compare individuals or groups with different levels of preknowledge (Hake, 1998). In our study normalised learning gain was calculated for individual students and the mean value was determined for each of the four subgroups. Normalising learning gain against room for improvement in our case yields the following equation:

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

where 19 minus the prescore represents the room for improvement on a test with 19 items. Table 4 is an overview of how the four subgroups compared in terms of performance, pass rates and the average normalised learning gain achieved, calculated as suggested by Hake (1998). Table 4 is divided into four quadrants. Each quadrant represents a performance evaluation subgroup. The results for the first year of implementation are discussed in detail below.

Statistical analyses were conducted to investigate whether the differences among the four subgroups shown in Table 4 were significant. The Kruskal-Wallis test was selected for this purpose based on the fact that the data set were skewed and the sizes of the subgroups were small. This test is the non-parametric analog to an ANOVA. The results indicated that there was no statistical difference among pretest performances of the subgroups at a 5% level of significance. Even though a significant difference was found for the performance of the subgroups in CMY 133 (p = 0.0435), from the post hoc tests, no significant differences were observed between any pairs of subgroups (the largest difference was observed between the OC-R and OC-OC subgroups, p = 0.055). We interpreted these results to mean that despite small differences, the four subgroups can be assumed to be comparable in terms of prior knowledge in stiochiometry based on pretest and CMY 133 performance, but the OC-R subgroup may have been

more able or better prepared than the OC-OC subgroup based on previous semester achievement.

It was further established that the students in the four subgroups differed significantly in terms of posttest performance ($p = 0.001$) and learning gain ($p < 0.001$). Statistically 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. This means that 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), nor the two subgroups that were overconfident in the posttest (OC-OC and R-OC). Post hoc tests between learning gains of subgroups showed that there were statistically significant differences between all the subgroups except between R-R and OC-OC, and between R-R and OC-R ($p = 0.0636$). Pretest performance averages were lower than the required passing mark of 50% for all the subgroups (33% to 45%) with the R-R obtaining the highest average score and OC-OC the lowest. The performance in the posttest was substantially higher than performance in the pretest, ranging from 45% to 68%. The OC-R subgroup achieved the highest average performance score in the posttest and the R-R subgroup managed to achieve an almost 100% pass rate. However, there was no improvement in the pass rate for the students in the R-OC subgroup, which corresponds with an insignificant increase in average performance from the pretest (41%) to the posttest (43%). According to Table 4 the best posttest performance and the most meaningful improvement were demonstrated by the OC-R and R-R subgroups. The proportion of students in these subgroups that passed the posttest was more than 50 percentage points higher than the proportion that passed the pretest. The learning gain of the OC-OC group was moderate and the R-OC subgroup did not achieve any learning gain at all. These findings confirmed that we were dealing with four discrete subgroups with different characteristics. To put the results for learning gain into perspective: The 49% gain demonstrated by the OC-R subgroup compares favourably with the best results achieved in mechanics where interactive teaching methods were used (Hake, 1998).

This study was repeated in the following year with the new intake of BFYP students ($N = 300$). Students were again categorised in terms of shifts in accuracy of performance evaluation and normalised learning gain was calculated for each student. The results are included in table 4. Three students were categorised as under-confident, two in the pretest and one in the posttest. The distribution of the other students was similar to that reported for Year 1: The majority of students belonged to the OC-OC subgroup (70%) with the remainder distributed fairly evenly between the other three subgroups. The average learning gain was higher than the previous year for all subgroups, but, most importantly, the intricate relationships between shifts in accuracy of performance evaluation and learning gain was confirmed. Students who were realistic in their performance judgement in the posttest achieved the highest learning gain, with 53% and 43% demonstrated by the OC-R and R-R subgroups, respectively. Students who remained or became overconfident in the posttest achieved considerably lower learning gains, with 32% and 15% demonstrated by the OC-OC and R-OC subgroups, respectively.

Conclusions

The majority of the students in this study 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. A small number of students showed improved metacognitive monitoring after instruction but a similar number of students developed confidence in their performance that was unjustified. Surprisingly, our results suggest that academic overconfidence was not a crippling disposition, provided that exposure to subject content and learning opportunities resulted in an improvement in performance evaluation, as in the case of the OC-R subgroup.

An initial positive bias in performance evaluation may actually be beneficial to learning. Inaccuracy in self-evaluation in the pretest did not hamper learning for both the OC-OC and OC-R subgroups, but when overconfidence persisted despite teaching and learning (OC-OC) or developed upon exposure to subject content (R-OC) it had serious consequences. Students 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, the R-OC subgroup, were shown to be the most vulnerable based on their poor learning gain. Together, these two subgroups that were overconfident in the posttest represent 72% of our sample in year 1 and 82% of the sample in year 2.

In their normal practice BFYP teachers concentrated on the teaching and learning of stoichiometry and were not focussed on developing metacognitive monitoring skills as well. Our results suggest that students are slow to develop accurate metacognitive monitoring skills within a classroom environment that did not include instruction focused on the development of such skills. Students who improved their metacognitive monitoring also showed the highest mean learning gain, but simultaneous mastery of cognitive and metacognitive skills was achieved without an explicit intervention by a mere 11% or 14% of our sample, the OC-R subgroup. We conclude, therefore, that instructional design for under-prepared students should focus on development of both kinds of skills if risk of failure is to be averted. Instruction should focus on the teaching of specific monitoring and regulatory strategies that students can use in academic tasks such as preparation for summative assessment and test-taking. Our findings suggest that overconfidence may arise due to an illusion of knowing where knowing is compromised by the presence of misconceptions. Overconfidence may also arise where mathematical skills are inadequate and mistakes go unnoticed. Assessment practices as well as the quality and intervals of feedback provided by the educators could be improved with the aim of making students aware of what they know and do not know. Tests could consist of tasks that require higher cognitive demand and deeper engagement which may force students to critically and realistically judge their performance. Student generated submicro diagrams can also be used as a teaching tool

Table 4 Pre- and posttest performance data according to performance evaluation subgroup

	POST OC (*Overconfident in posttest*)			**POST R** (*Realistic in posttest*)		
PRE OC (*Overconfident in pretest*)		**Year 1**	**Year 2**		**Year 1**	**Year 2**
	Size of sample subset (% of sample)	***n* = 50 (55%)**	***n* = 207 (70%)**	Size of sample subset (% of sample)	***n* = 13 (14%)**	***n* = 33 (11%)**
	Average performance CMY 133	**50**	n/a	Average performance CMY 133	**61**	n/a
	Average Pretest performance (%)	**33**	28	Average Pretest performance (%)	**38**	35
	Average Posttest performance (%)	**45**	51	Average Posttest performance (%)	**68**	70
	% Pass Pretest	**10**	10	% Pass Pretest	**23**	27
	% Pass Posttest	**40**	58	% Pass Posttest	**77**	88
	Average Learning Gain (%)	**19**	32	Average Learning Gain (%)	**49**	53
PRE R (*Realistic in pretest*)	Size of sample subset (% of sample)	***n* = 15 (17%)**	***n* = 35 (12%)**	Size of sample subset (% of sample)	***n* = 11 (12%)**	***n* = 22 (7%)**
	Average performance CMY 133	**53**	n/a	Average performance CMY 133	**58**	n/a
	Average Pretest performance (%)	**41**	39	Average Pretest performance (%)	**45**	40
	Average Posttest performance (%)	**43**	48	Average Posttest performance (%)	**61**	66
	% Pass Pretest	**27**	20	% Pass Pretest	**36**	36
	% Pass Posttest	**27**	46	% Pass Posttest	**91**	100
	Average Learning Gain (%)	**-1**	15	Average Learning Gain (%)	**25**	43

to expose misconceptions and achieve mastery in stoichiometry (Davidowitz *et al*., 2010). These approaches may prevent the damage caused by failure and preserve the positive contribution of confidence, albeit excessively positive.

To conclude, we revisit our heuristic decision to allow an error in performance judgment equivalent to three questions in a test comprising of 19 items, i.e. 15.8%. This "acceptable" margin of error was chosen specific to our context in recognition of poor skills development of our sample, and the nature of subject content and the test instrument. However, students should become much better calibrated than this to avoid risk of failure in a challenging tertiary environment. Students in academic development programmes face numerous academic and personal challenges, but they also receive specialised support. Refining the art of accurate self-evaluation should be one of the objectives of such specialised support.

Acknowledgements

NRF funding
Graça Machel Scholarship for women
Jacqui Sommerville and Karien Adamski for statistical analysis.

Notes and references

[a] While judgments of confidence are commonly used in metacognition literature as an indication of perceived performance we acknowledge the potential ambiguity of this interpretation. The ambiguity about what exactly is measured by judgments of confidence warrants an in depth consideration by researchers in this field.

1. Bandura A., (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
2. Beyer S. & Bowden E. M., (1997), Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias, *Pers. Soc. Psychol. Rev.*, **23**, 157 – 180.
3. Bol L. and Hacker D., (2001), A comparison of the effects of practice tests and traditional review on performance and calibration., *J. Exp. Educ.*, **69**, 133 – 151.
4. Britner S. L. and Pajares F., (2006), Sources of science self-efficacy beliefs in middle school children, *J. Res. Sci. Teach.*, **43**, 485 – 499.
5. Campbell W. K., Goodie A. S., & Foster J. D., (2004), Narcissism, confidence, and risk attitude, *J. Behav. Decis. Making.*, **17**, 481-502.
6. Carter T. V. and Dunning D., (2008), Faulty self-assessment: why evaluating one's own competence is an intrinsically difficult task, *Soc. Pers. Psychol. Comp.*, **2**, 346 – 360.
7. Carvalho M. K. F. and Yuzawa M., (2001), The effects of social cues on confidence judgements mediated by knowledge and regulation of cognition, *J. Exp. Educ.*, **69**, 325 – 343.
8. Carvalho M. K. F., (2009), Confidence judgments in real classroom settings: monitoring performance in different types of tests, *Int. J. Psychol.*, **44**, 93 – 108.
9. Cooper M. M., Sandi-Urena S. and Stevens R., (2008), Reliable multi method assessment of metacognition use in chemistry problem solving, *Chem. Educ. Res. Pract.*, **9**, 18 – 24.
10. Davidowitz B., Chittleborough G. and Murray E., (2010). Student generated submicro diagrams: a useful tool for teaching and learning chemical equations and stoichiometry, *Chem Educ. Res. Pract.*, **11**, 154 – 164.
11. Dunning D., Johnson K., Ehrlinger J. and Kruger J., (2003), Why people fail to recognise their own incompetence, *Am. Psychol*, **12**, 83 – 87.
12. Dunning D., (2005), *Self-Insight: Roadblocks and detours on the path of knowing thyself*, New York: Psychology Press.
13. Dunlosky J., Serra M. J., Matvey G. and Rawson K. A., (2005), Second-Order Judgements About Judgements of Learning, *J. Gen Psychol.*, **132**, 335 – 346.
14. Ehrlinger J., (2008), Skill level, self-views and self-theories as sources of error in performance evaluation, *Soc. Pers. Psychol. Comp.*, **2**, 382 – 398.
15. Fernandez-Duque D. and Black S. E., (2007), Metacognitive judgment and denial of deficit: Evidence from frontotemporal dementia, *Judgm. Decis. Mak.*, **2**, 359 – 370.
16. Flavell J. H., (1979), Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry, *Am. Psychol.*, **34**, 906 – 911.
17. Gramzow R. H., Elliot A. J., Asher E. and McGregor H. A., (2003), Performance evaluation bias and academic performance: Some ways and some reasons why, *J. Res. Pers.*, **37**, 41 – 61.
18. Grimes P., (2002), The overconfident principles of economics student: An examination of a metacognitive skill, *J. Econ Educ.*, **33**, 15 – 30.
19. Hacker D.J., Bol L. and Bahbahani K., (2008), Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style, *Metacog. Learn.*, **3**, 101 – 121.
20. Hake R. R., (1998), Interactive-engagement *vs.* traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.*, **66**, 64 – 74.
21. Harrits G. S., (2011), More than method?: A discussion of paradigm differences within mixed methods research, *J. Mix. Method. Res.*, **5**, 150 – 166.
22. Hartman H. J., (2001), *Metacognition in Learning and Instruction: Theory Research and Practice*, Dordrecht: Kluwer Academic Publishers.
23. Hart J. T., (1965), Memory and the feeling-of-knowing experience, *J. Educ. Psychol.*, **56**, 208 – 216.
24. Hasan S., Bagayoko D. and Kelley E., (1999), Misconceptions and the certainty of response index (CRI), *Phys. Educ.*, **34,** 294–299.
25. Herron J. D., (1990), Research in Chemical Education: Results and Directions, in Gardner M. (ed.), in *Toward a scientific practice of science education*, Routledge.
26. Huddle P. A. and Pillay A. E., (1996), An In-Depth Study of Misconceptions in Stoichiometry and Chemical Equilibrium at a South African University, *J. Res. Sci. Teach.*, **33**, 65 – 77.
27. Johnstone A. H., (2010), You can't get there from here, *J. Chem. Educ.*, **87**, 22 – 29.
28. Johnstone A. H., (1991), Why is science difficult to learn? Things are seldom what they seem, *J. Comput. Assist. Lear.*, **7**, 75 – 83.
29. Kennedy E. J., Lawton L. and Plumlee L., (2002), Blissful ignorance: The problem of unrecognised incompetence and academic performance, *J. Marketing Educ.*, **24**, 243 – 252.
30. Koriat A. and Bjork R. A., (2005), Illusions of competence in monitoring one's knowledge during study, *J. Exp. Psychol. Learn.*, **31**, 187 – 194.
31. Kruger J. and Dunning D., (1999), Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated performance evaluations, *J. Pers. Soc. Psychol.*, **77**, 1121 – 1134.
32. Lazonby J., Morris J. and Waddington D., (1985), The mole: Questioning format can make a difference, *J. Chem. Educ.*, **62**, 60 – 61.
33. Leech N. L. and Onwuegbuzie A. J., (2011), Mixed research in counselling: Trends in the literature, *Meas. Eval. Couns. Dev.*, **44**, 169 – 180.
34. Lichtenstein S. and Fischhoff B., (1997), Do those who know more also know more about how much they know?, *Organ. Behav. Hum. Perf.*, **20**, 159 – 183.
35. Mabe P. A. III and West S. G., (1982), Validity of self-evaluation of ability: A review and meta-analysis, *J. Appl. Psychol.*, **67**, 280 – 296.
36. Maxwell J. A. and Loomis D. M., (2003), Mixed methods design: An alternative approach, in Tashakkori A. and Teddlie C. (eds.), in *Handbook of mixed methods in social and behavioural research*. India: Sage publications.
37. Nelson T. O. and Narens L., (1990), Metamemory: A theoretical framework and new findings, in Bower G. H. (ed.), in *The psychology of learning and motivation: advances in research and theory*. San Diego, California: Academic Press, Inc.
38. Nowell C. and Alston M. R., (2007), I thought I Got an A! Overconfidence Across the Economics Curriculum, *J. Econ. Educ.*, **38**, 131 – 142.
39. Ochse C., (2003), Are positive self-perceptions and expectancies really beneficial in an academic context?, *South Afr. J. High. Educ.*, **17**, 6 – 73.
40. Pajares F., (1996), Self-efficacy beliefs in academic settings, *Rev. Ed. Res.*, ***66***, 543–578.
41. Pallier G., Wilkinson R., Danthiir V., Kleitman S., Knezevic G., Stankov L. and Roberts R. D., (2002), The role of individual differences in the accuracy of confidence judgments, *J. Gen. Psychol.*, ***129***, 257 – 299.
42. Potgieter M., Davidowitz B. and 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, *Proceedings of the 38th annual conference of the Australasian Science Education Research Association (ASERA)*, Fremantle, WA, 2007.

43. Potgieter M., Malatje E., Gaigher E. and Venter E., (2010), Evaluation versus performance as indicator of the presence of alternative conceptions and inadequate problem solving skills in mchanics, *Int. J. Sci. Educ*., **32**, 1407 – 1429.
44. Ridley D. S., Schutz P. A., Glanz R. S. and Weinstein C. E., (1992), Self-regulated learning: the interactive influence of metacognitive awareness and goal-setting, *J. Exp. Educ.*, **60**, 293 – 306.
45. Rozenblit L, and Keil F., (2002), The misunderstood limits of folk science: an illusion of explanatory depth, *Cognitive Sci.*, **26**, 521 – 562.
46. Rosenthal D. M., (2000), Consciousness, Content, and Metacognitive Judgments, *Conscious Cogn.*, **9**, 203 – 214.
47. Sandi-Urena S., Cooper M. M. and Gatlin T. A., (2011), Graduate teaching assistants' epistomological and metacognitive development, *Chem. Educ. Res. Pract*., **12**, 92-100.
48. Sandi-Urena S., Cooper M. M. and Stevens R. H., (2011), Enhancement of metacognition use and awareness by means of a collaborative intervention, *Int. J. Sci. Educ.*, **33**, 323-340.
49. Schaefer P. S., Williams C. C., Goodie A. S. and Campbell W. K., (2004), Overconfidence and the Big Five, *J. Res. Pers.*, **38**, 473 – 480.
50. Schwartz B. L. and Perfect T. J., (2002), Introduction: toward an applied metacognition, In Schwartz B. L. and Perfect T. J., (Eds.). *Applied Metacognition*. Cambridge University Press.
51. Schraw G., (2009), A conceptual analysis of five measures of metacognitive monitoring, *Metacog. Learn.*, **4**, 33 – 45.
52. Schraw G., Crippen K. J. and Hartley K., (2006), Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning, *Res. Sci. Ed.*, **36**, 111 – 139.
53. Zimmerman B. J., (2000), Self-efficacy: An essential motive to learn, *Contemp. Educ. Psychol*., **25**, 82 – 91.
54. Zohar A. and Dori Y.J., (2012), Metacognition in science education, Trends in current research in Ziedler, D. (ed.), in *Contemporary trends and issues in science education*, Springer. ISBN 978-94-007-2132-6, pp. 1 – 19.A.

Chemistry Education Research and Practice

Cite this: DOI: 10.1039/c0xx00000x

FULL PAPER

www.rsc.org/xxxxxx

Metacognitive monitoring and learning gain in foundation chemistry

Kgadi C. Mathabathe,*[a] and Marietjie Potgieter[b]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

The ability to make realistic judgements of one's performance is a demonstration of the possession of strong metacognitive skills. In this study we investigate the relationship between accuracy of self-evaluation as an expression of metacognitive skill, and learning gain in stoichiometry. The context is an academic development programme at a South African University, offered for under-prepared students enrolled for science and engineering. These students generally exhibit unrealistically high levels of confidence in performance and this could potentially place them at risk by negatively affecting decisions regarding time management and self-regulation. We investigated whether overconfidence before instruction is corrected upon exposure to teaching. A three-tier stoichiometry test was used to collect qualitative and quantitative data before and after instruction. Findings indicate that the majority of the students were overconfident in the evaluation of their performance in both the pre- and posttests. Overconfidence was not a debilitating disposition when demonstrated in the pretest provided that it was corrected during teaching and learning. The most vulnerable students were those that judged their performance or lack thereof realistically in the pretest but became overconfident during the teaching and learning of stoichiometry. Our results suggest that under-prepared students are slow to develop accurate metacognitive monitoring skills within a classroom environment that did not include instruction focused on the development of such skills. We recommend a proactive and constructive response by educators which may reduce the incidence of failure and preserve the positive contribution of confidence, albeit excessively positive.

Keywords: learning gain, foundation chemistry, metacognition, stoichiometry

[a] *University of Pretoria, Department of Science, Mathematics and Technology Education, Pretoria, South Africa. Fax: +27865128641; Tel: +27 12 420 2758; E-mail: kgadi.mathabathe@up.ac.za*
[b] *University of Pretoria, Department of Chemistry, Pretoria, South Africa. Fax: +27 12 420 5441; Tel: +27 12 420 6472; E-mail:marietjie.potgieter@up.ac.za*

† Electronic Supplementary Information (ESI) available: The Stoichiometry test instrument. See DOI: 10.1039/b000000x/

Introduction

The need for the teaching of metacognitive skills is one of the main implications of research on the teaching and learning of science that has emerged during the past three decades, according to the editors of a recent monograph on metacognition in science education (Zohar and Dori, 2012). A small number of studies have been reported in this journal over the past five years focussing on the development of metacognition in problem solving (Cooper *et al*., 2008) and in the chemistry laboratory (Sandi-Urena *et al*., 2011). Metacognition is generally accepted to consist of both knowledge of cognition and regulation of cognition (Flavell, 1979), where knowledge of cognition refers to the ability to monitor learning or evaluate performance. In this study we explore the relationship between metacognitive monitoring and learning gain in stoichiometry.

Accurate performance evaluation is critical in decisions on the time required to study for a specific course, what study methods to employ as well as what topics to give the most attention to (Grimes, 2002; Nowell and Alston, 2007). Assessment in the form of test-taking plays a vital role in the development of important metacognitive skills such as accurate performance evaluation. Test-taking is a challenging academic requirement but it provides a valuable opportunity for students to learn how to regulate their own learning in a certain domain, to better monitor their performance in that domain and to make valid attributions of their failures and successes (Carvalho, 2009).

Previous studies have reported the general occurrence of bias in performance evaluation in the form of overconfidence as well as the potentially negative consequences that it may have on academic success (Ochse, 2003; Potgieter *et al*., 2007). In the current study, overconfidence is defined as inflated levels of confidence that a student displays with regard to the accuracy of answers in a test. The context of this study is tertiary chemistry in an academic development programme at a South African university, specifically learning gains achieved in stoichiometry, a core content topic in the first-year syllabus. The majority of students in this programme come from disadvantaged backgrounds with inadequate preparation for tertiary science. Many of these students have been found to exhibit exaggerated levels of confidence in their mastery of foundational concepts and skills in chemistry (Potgieter *et al*., 2007). This study is an investigation of how accurately students in this programme evaluate their performance during test-taking and the influence of teaching on accuracy of performance evaluation.

Literature review

When we are prompted to make judgements on how we perceive our ability, how well we know something or how well we have performed a particular task, the judgements we report are called metacognitive judgements (Rosenthal, 2000; Dunlosky *et al*., 2005; Koriat and Bjork, 2005; Fernandez-Duque and Black, 2007). Metacognitive judgements have been extensively investigated partially due to the fact that mastery of the skill of accurately making them may result in the effective management of self-regulated study, which is necessary in a tertiary environment where an autonomous approach to studying is required (Dunlosky *et al.*, 2005).

Metacognition refers to the knowledge and experiences we have about our own cognitive processes (Flavell, 1979). Metacognition consists of metacognitive knowledge and metacognitive experiences. 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) that an individual performs to control cognitive activities to ensure that cognitive goals are met. 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. All the facets of metacognitive knowledge are necessary for one to self-regulate one's thinking and learning effectively (Hartman, 2001). 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. Making accurate judgements about one's performance and competence is a metacognitive process that people can use to regulate their behaviour towards successful learning (Hacker *et al.,* 2008). Students who can accurately assess the effectiveness of their learning strategies and their understanding of concepts in a particular subject area should be able to make informed decisions as to whether and when to intensify or redirect their studying for a test (Hacker *et al.,* 2008). Metacognitive skills should therefore differentiate a novice learner from an expert learner. An expert learner knows how to learn and also knows which strategies work best (Grimes, 2002; Nowel and Alston, 2007).

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). Although the two paths have remained separate, contemporary research was introduced to the construct of metacognition through the publication of Nelson and Narens (1990)'s 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). Nelson and Narens (1990) identified several types of metacognitive judgements namely ease-of-learning judgements, judgements of knowing or judgements of learning, feeling-of-knowing judgements and confidence judgements. The theoretical framework of Nelson and Narens (1990) describes three stages, namely the acquisition, retention and retrieval stages during which metacognitive judgements are made. 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. In our study we were particularly interested in the judgements made during test taking. In addition to exaggerated confidence judgements made based on feelings or incorrect information, several factors associated with bias in performance evaluation or monitoring, particularly overconfidence, have been identified and categorised by Carvalho (2009) as personal, task-related and environmental factors. Task related factors include the lack of skill (Kruger and Dunning, 1999; Ehrlinger. 2008); properties of the task (Lichtenstein and Fischhoff, 1997); format selected for evaluation (Carvalho, 2009) and the quality of feedback received (Carter and Dunning, 2008). Personal factors include factors such as the tendency to rely on chronic self-views to evaluate performance (Ehrlinger 2008); the need for self-protection and self-enhancement (Gramzow *et al*., 2003); theories of intelligence that respondents adhere to (Ehrlinger, 2008); personality traits (Campbell, Goodie and Foster, 2004) and gender (Beyer and Bowden, 1997). Some of these factors were investigated in this study but will be reported in a separate paper.

Metacognition is but one component of self-regulated learning (Schraw *et al., 2006*). Self-efficacy, which is a subcomponent of the motivation component of self-regulated learning, is described as the extent to which an individual is confident that he or she can perform a specific task (Bandura, 1997). Self-efficacy is an important aspect of self-regulated learning as it affects the extent to which learners engage and persist at challenging tasks (Schraw *et al.*, 2006). Bandura (1997) argued that self-efficacy judgements that are slightly higher than actual accomplishments serve to increase individual's effort and persistence. However, self-efficacy should not be confused with confidence statements made after a task has been completed. Self-efficacy is an expression of confidence about a task that must be performed. Confidence judgements that are made *ex post facto* are an expression of metacognitive monitoring. While self-efficacy has been found to be a strong predictor of academic performance (Britner and Pajares, 2006), overly confident evaluation of performance is not.

Accuracy of metacognitive judgements has been studied extensively in cognitive psychology and educational psychology over the past three decades. In general, the perceptions people hold of either their overall ability or specific performance tend to be correlated only modestly with their performance, with better accuracy being correlated with better performance (Bol & Hacker, 2001). Mabe and West (1982) surveyed 55 self-evaluation studies with a combined population of 14,811 subjects across a variety of domains and found the average correlation between self-estimates and actual performance to be only 0.29. While performance estimates have been shown to be generally unrealistic, overestimation of performance was particularly prevalent (for a recent review, see Dunning, 2005). Despite much poorer performance, weaker students have been found to be particularly overly optimistic about the correctness of their answers in tests (Carvalho, 2009; Carvalho & Yuzawa, 2001; Kruger & Dunning, 1999; Kennedy, Lawton & Plumlee, 2002, Potgieter *et al.*, 2007). These students display poor judgement in the sense that many of the answers which they expected to be correct are indeed wrong.

In studies on accuracy in self-evaluation a number of different methods have been reported to quantify the extent of inaccuracy. These methods can be described as either direct or indirect probes to elicit an expression of perceived performance, where direct methods refer to studies where respondents were asked to predict their total score before or after completing the task. The predicted score is then compared with actual performance on the test to obtain a calibration score (*e.g.* Dunning *et al*., 2003; Hacker *et al*., 2008). In indirect methods respondents are asked to report their confidence in their performance on each test item immediately after completion of the task. Confidence judgments are reported on a Likert scale and are interpreted as expressions of the likelihood of getting the answer correct. The average of confidence ratings over all test items is interpreted as perceived performance.[a] The statistic used in many studies on overconfidence in psychology is called the bias score which is the difference between average confidence ratings over all test items and the percentage of correct responses on the test (Pallier *et al*., 2002; Schraw, 2009). Confidence ratings are typically grouped into discrete categories on a Likert scale, but may also be unstructured. Schaefer *et al*. (2004) studied overconfidence using two-option fixed-choice questions on general knowledge followed by a seven category rating of confidence in the answer provided starting from 50% (50 – 52%, 53 – 60%, 61 – 70%, 71 – 80%, 81 – 90%, 91 – 97%, 98 – 100%). In his study, Carvalho (2009) asked psychology majors (N = 129) to indicate their confidence in the accuracy of their responses on a Likert scale ranging from 0 to 100 per cent. He reported that these students were less accurate in performance evaluation on multiple-choice test items than on short answer questions. We have found that direct methods for eliciting perceived performance were too insensitive a probe in our context and opted for requesting a confidence rating after every item on an eleven point Likert scale with 10% intervals between 0 and 100%.

Ochse (2003) conducted a study at a South African university where accuracy of self-evaluation was used to categorise third-year psychology students as overestimators, realists or underestimators in order to investigate the academic success of the different groups. Before their final examination, students were asked to indicate the 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 examination, actual scores were compared with expected scores. A difference of less than nine percentage points between actual and expected scores was considered a realistic estimate. Students who overestimated their actual score by nine or more percentage points were categorised as overestimators and those who underestimated their actual score by nine or more percentage points were categorised as underestimators. Overestimators, on the whole, expected higher marks than realists and underestimators, were significantly more confident about the accuracy of their expected scores, and perceived themselves to have higher ability, but however obtained the lowest scores of the three groups. On average the overestimators failed the course.

In studies with psychology students, Dunning and coworkers

(Kruger and Dunning, 1999; Dunning *et al*., 2003) have consistently found that the mismatch between perceived and actual performance was the largest for poor performing students, that this mismatch was smaller for students with better performance and that the best performing students were not only best calibrated but often slightly underestimated their performance. They have called this finding the dual burden of incompetence, to be unskilled and unaware of it (Kruger and Dunning, 1999), and argued that one of the reasons for overconfidence is the fact that the same knowledge and skills that are required for good performance are also required for accurate self-evaluation of performance (Dunning *et al.*, 2003). Making informed and accurate judgements of performance in a test requires the learner to have a good knowledge and understanding of the content and metacognitive knowledge of task and strategy variables with regard to that content or subject matter. How else can one judge if the problem was solved successfully when one does not even recognise which approach is required to solve the problem in the first place? For this study we chose a chemistry topic, stoichiometry, for which prior knowledge was known to be limited, so that there would be a large scope for improvement, in order to see whether growth in knowledge and understanding would be accompanied by improvement in the accuracy of metacognitive judgments. Success in solving stoichiometry problems requires representational competence, formal reasoning and being able to correctly apply multistep mathematical operations (Johnstone, 1991; Huddle and Pillay, 1996). Previous research has shown that students are able to solve one-step problems such as finding the molar mass of a compound or calculating the number of moles corresponding to a given mass of the compound, but they are unsuccessful when a problem requires the stringing together of such steps to solve a more complex task (Lazonby *et al*., 1985; Huddle and Pillay, 1996). Herron (1990) suggests that failure on the more complex tasks may be due to poorly developed metacognitive skills that govern the organisation of work, sequencing of tasks, and checking of results. For example, when a learner is asked to solve a stoichiometry problem in chemistry, the learner must be able to:

a. recognise and understand what is being asked or what is expected of him or her;
b. 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);
c. perceive and acknowledge when he or she cannot solve the problem;
d. reflect on the reasons why he or she cannot solve the problem and make changes to his or her strategies in order to improve.

The problem arises when there is a mismatch between students' perception of mastery and actual performance. Overconfidence could potentially hamper learning by negatively affecting decisions regarding the regulation and control of cognition (Schraw *et al*., 2006). However, overconfidence could also motivate students to take on challenges that would otherwise be too imposing (Pajares, 1996; Zimmerman, 2000; Ehrlinger, 2008). This dichotomy has prompted us to ask the question whether overconfidence is necessarily undesirable within the context of first year chemistry in an academic development programme.

The context of this study

This study was conducted in partnership with educational psychology to investigate the relationship between accuracy of self-evaluation as an indication of metacognitive skill, and learning gain in stoichiometry, and whether overconfidence before instruction is reduced upon exposure to teaching of the topic. In addition we wanted to investigate the underlying reasoning informing metacognitive judgments on content mastery. The study was therefore guided by the following questions:

1. How accurately do BSc Four-year programme (BFYP) students evaluate their performance in a stoichiometry test?
2. How does accuracy of performance evaluation change upon teaching of the topic?
3. How does accuracy of performance evaluation vary as a function of item type?
4. How is learning gain associated with shifts in accuracy of performance evaluation?

Quantitative data was required to measure performance and learning gain, and to determine accuracy of performance evaluation. Qualitative data, on the other hand, was collected to enrich our understanding of the metacognitive factors underlying judgements of performance (Nelson and Narens, 1990). The current article reports only the analysis of quantitative data whereas the qualitative work will be reported later.

Methods

An embedded experimental mixed methods approach was followed to answer the above-mentioned research questions (Maxwell and Loomis, 2003; Harrits, 2011; Leech & Onwuegbuzie, 2011).

Participants

The sample consisted of 91 participants who voluntarily participated in the study and for whom complete records were obtained, *i.e.* 35 male students (38%) and 55 female students (60%), and one record with gender information omitted. The ages of participants ranged between 17 and 25 years (M = 19). These students shared commonalities in that they were all taught stoichiometry by the same lecturer in large group lectures and they completed compulsory computerised quizzes on the topic. In addition, they attended three fifty minute long small group tutorial sessions per week over a period of three weeks, where they had plenty of problem-solving and feedback opportunities. Even though they had different lecturers with different teaching styles in the small group sessions, the lecturers worked collaboratively in terms of the material used in these sessions as well the quantity, format and content of tests and tasks given to the students. Stoichiometry is taught extensively at high school

level, specifically at grades 10 and 11. By the time these students were in their first year undergraduate courses after completing grade 12, it had been a year since they received instruction on the topic. Participants were duly informed of the objectives of the study and were promised complete anonymity and that the results obtained in the study would not affect their grades in any way. The study was repeated the following year with a sample of 300 students from the new cohort of students enrolled in the BFYP.

Research setting

The study took place in the second semester of the university's academic development programme, the BSc Four-year programme (BFYP). The minimum duration of BSc programmes at the particular South African university is three years but in this programme an extra year is added at the foundational level to address under-preparedness of incoming first year students. The study received ethical clearance from the institution where it was conducted.

Data collection instrument

Pretest data were collected before a 3-week period of instruction on stoichiometry and posttest data after formal instruction was completed and ample opportunity was provided for guided and unguided problem-solving. The posttest was used for summative purposes about which students were informed beforehand. Pretest and posttest data were collected through a 20-item test instrument. Each item in the instrument consisted of three tiers to enable the simultaneous collection of both quantitative and qualitative data. The first tier consisted of a multiple choice question on stoichiometry followed by the second tier, a Likert scale from 0% to 100% with 10% intervals, on which participants were asked to indicate their confidence in the accuracy of the chosen response in the first question. Lastly, in the third tier participants were asked to explain their choice of confidence indicators. The common structure of test items is demonstrated in Figure 1.

1.1 Given the equation 3A + B → 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?

Fig. 1 Example of a three-tier test question

The test instrument was piloted and refined. The face and content validity were checked by tertiary educators and Grade 12 teachers and small changes were made as recommended by them. Cronbach alpha coefficients were calculated with respect to content items to determine the reliability of the test items. Item 2 was excluded *post hoc* based on validation feedback and poor item performance in the pretest. Posttest Cronbach's alpha coefficients of 0.69 and 0.70 were obtained in successive years of implementation. The test instrument is included as Appendix I.

The test instrument comprised nineteen items of a range of difficulties chosen and adapted from current literature and first year chemistry textbooks. Eight items measured procedural knowledge, formal reasoning and numeric problem-solving skills using multistep mathematical operations. Two items assessed declarative knowledge. To probe for conceptual understanding, eight items incorporated sub-microscopic representations or particulate drawings of atoms and molecules. One item assessed students' ability to match the symbolic representation of a balanced chemical equation with a graphical representation of the reaction.

Results and discussion

The sum of all correctly answered items on the stoichiometry test was used as a measure of performance with correct answers scored one (1) and incorrect answers zero (0). The confidence judgement ratings reported on a scale of 0% to 100% per item were used to determine the average of confidence judgements for each individual student and for the sample as a whole. Table 1 below presents a summary of the results in terms of performance and confidence scores obtained in the pre- and posttests for the first year of the study. The mean performance improved from 37% in the pretest to 51% in the posttest as would be expected, while the mean of average confidence scores increased by a similar margin from 63% in the pretest to 76% in the posttest. *P-values* ($p < 0.05$) confirmed that there was a statistically significant difference between the pre- and posttest performance and between pre- and posttest average confidence scores.

Table 1 Summary of student performance and average confidence scores in the pre- and posttest

	Performance (test scores, %), $N = 91$		Average confidence scores (%), $N = 91$	
	Pretest (19 items)	Posttest (19 items)	Pretest	Posttest
Mean score	7.0 (37%)	9.6 (51%)	63.0	75.7
Standard deviation	2.9 (15%)	3.4 (18%)	17.4	13.5
Minimum	2 (11%)	3 (16%)	16.3	40.5
Maximum	15 (79%)	18 (95%)	94.7	99.5

Test scores out of a total mark of 19 were converted to a percentage to represent *actual performance*. The average confidence scores represent the *perceived performance* of the students. The comparison between the actual and perceived performance of students are presented in Figures 2 and 3 for the pretest and posttest, respectively, in terms of the mean values for confidence ratings and test scores for students per performance quartile. This method of analysis and presentation corresponds with that of Dunning *et al.* (2003) and enables direct comparison

of findings. The results presented in Figures 2 and 3 show that students in our sample overestimated their performance before and after instruction. Our results corroborated with those of Dunning *et al.* (2003) in that the students in the bottom quartile misjudged their performance by the biggest margin. One would have expected students to judge their performance better after instruction, but this happened only to a limited extent for the third and top performance quartiles. The mismatch between actual and perceived performance increased marginally for the bottom quartile in the posttest. Dunning *et al.* (2003) reported that the gap between actual and perceived performance narrowed as performance improved and a small underestimation of performance was demonstrated by the top quartile. This did not happen in our case where there was still a 9% overestimation of performance for the top quartile in the posttest.

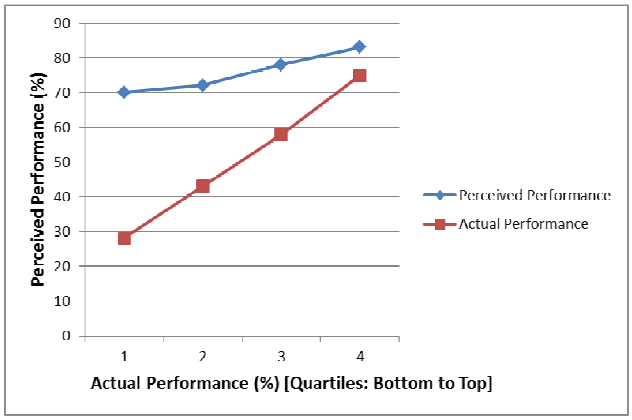

Fig. 2 Perceived *versus* actual pretest performance of BFYP students in the four performance quartiles

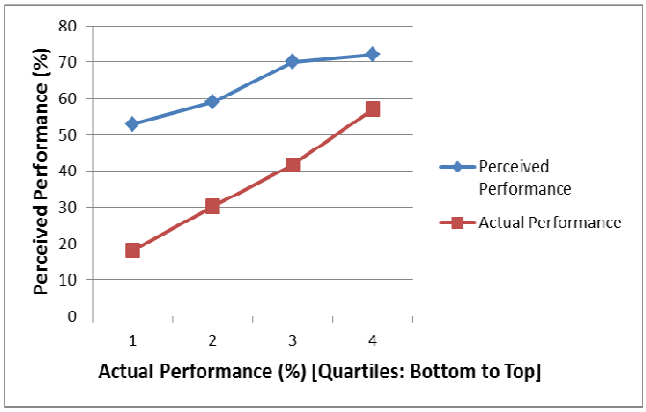

Fig. 3 Perceived *versus* actual posttest performance of BFYP students in the four performance quartiles

The accuracy of judgements made in the pre- and posttest (Research questions one and two)

It is clear from the results presented in Table 1 and Figures 2 and 3 that the students were unrealistic in their judgement of performance both in the pretest and the posttest. Despite getting on average only 37% of answers correct in the pretest, they were 63% certain that the answers were correct. Similarly, despite achieving a mere 51% in the posttest they were on average 76% certain of a correct answer. These results provide a general answer to the first two research questions. On average, the performance improved after teaching of stoichiometry but the accuracy of performance evaluation did not. However, a wealth of information is hidden behind mean values which requires further unpacking.

Accuracy of performance evaluation as a function of item type (Research question three)

Test items were classified according to the problem solving skills they required. Shifts in accuracy of judgement per item after instruction were also investigated. Items were categorised into four types, namely type A requiring multi-step mathematical operations, type B requiring declarative knowledge, type C requiring representational competence, and type D requiring interpretation of graphs. Type A was further divided into two subcategories, types A1 and A2. Type A1 represented all the items that required *simple* multistep mathematical operations to solve while Type A2 represented all questions that required the use of *complex* multistep mathematical operations. Accuracy of judgement as represented by the difference between *perceived* and *actual* performance was plotted for each item against item difficulty as reflected by average performance on that item. The results are shown in Figures 4 and 5 for the pre- and posttests, respectively. Different symbols are used to indicate item types.

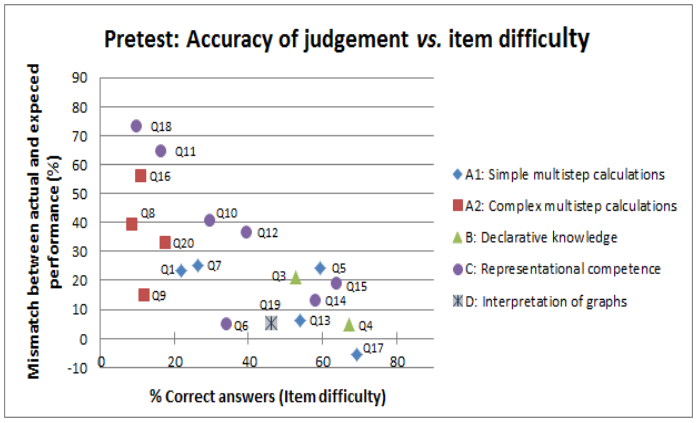

Fig.4 Accuracy of judgement *versus* item difficulty for the pretest

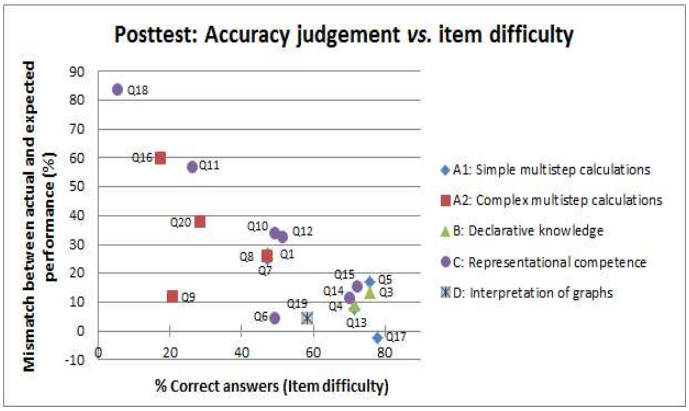

Fig.5 Accuracy of judgement *versus* item difficulty for the posttest

The first observation that can be made from Figures 4 and 5 is that accuracy of judgement does not seem to depend on item type. All five item types are represented in the band where the mismatch between actual and perceived performance is no more than 15 - 20%, *i.e.* where the judgement is fairly accurate. Similarly, three of the five item types are present in the band where judgement is poor or very poor, *i.e.* where the mismatch between actual and perceived performance is more than 20%. There were not enough items of types B and D in the instrument to be able to make general inferences about tasks requiring declarative knowledge or interpretation of graphs. The second observation is that good performance is accompanied by accurate judgement, but the reverse is not true. Students seem to have been aware of task properties associated with items 6, 9 and 19 in both tests and realised that they were unlikely to have answered them correctly. The third observation is that performance increased from the pretest to the posttest on all items except Q18, but accuracy of judgement did not. The shift in the position of items on the graphs from Figure 4 to Figure 5 is horizontal and not downwards to the right as one would have wished. The error in judgement in the posttest is therefore not reflecting lack of exposure or practice, but rather factors inherent in the task required for each item.

We now turn to the relationship between the nature of task properties and accuracy of performance judgment. For this purpose we compare item 9 with items 11, 16, 18 and 20. The difficulty of these items was comparable, with performance varying between 5% and 29%, but accuracy of judgment differed greatly (12% – 83%). Both items 9 and 20 entailed conversion of units and multistep calculations, but item 9 required additional steps and concept integration which challenged students beyond what they were exposed to during regular classroom practice. Poor performance on item 9 was judged accurately, but not item 20, presumably because students were daunted by the challenges of item 9, but overconfident on item 20 where they made calculation errors that went unnoticed. Item 18 was deceptively simple; students understood the task, but lacked or failed to apply procedural knowledge regarding accepted notation for chemical reactions. Items 11 and 16 required interpretation of coefficients and subscripts in a balanced reaction expressed in schematic or symbolic form. Confusion between coefficients and subscripts is a misconception that is well documented (*e.g.* Huddle and Pillay, 1996). Misconceptions exist due to the inadequacy of mental models and have been suspected to give rise to overconfidence (Hasan *et al.*, 1999), presumably because they create the illusion of knowing (Rozenblit and Keil, 2002). An illusion of knowing will not be conducive to accurate metacognitive monitoring. A similar analysis was done on a second group of items where performance ranged between 47% and 58%. Students judged their performance well on two items with an unfamiliar design (Q6 and Q19), but less accurately on the other five items which required either multistep calculations or contained sub-microscopic representations which were designed to reveal the presence of misconceptions. We concluded that misconceptions and the lack of mathematical skills are likely to compromise accuracy of performance judgements. Our findings resonate with those of a previous study where inflated confidence levels in mechanics have also been shown to be associated with the presence of misconceptions and lack of problem solving skills (Potgieter *et al*., 2010).

Categorisation of students in terms of accuracy of performance evaluation

The students in our sample were categorised individually based on their demonstrated accuracy of performance evaluation both in the pretest and the posttest. The difference between the actual and perceived performance values was regarded as a measure of accuracy of judgement. It is to be expected that students would make an error in judgement at least in the pretest because they are poorly prepared for the topic of stoichiometry and some of the items require several steps of calculations or analytical thinking. Taking into consideration the difficulty of the topic, the level of preparedness of the students in our sample and the format of the test we set an "acceptable" margin of error beyond which we judged the consequences of poor self-calibration to be too serious in terms of risk of failing. We set this margin of error as the equivalent of three incorrect judgements out of 19 answers, which translates into a judgement error of 15.8%. For example, 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 acceptable error margin of 15.8%. Subjects whose average confidence scores exceeded their test scores by more than 15.8% were labelled as overconfident (OC). The realistic group (R) were students with a difference between actual and perceived performance between 15.8% and -15.8% (-15.8% and 15.8% included). Students whose test scores exceeded their average confidence scores by more than 15.8% were labelled as under-confident (UC). Using these criteria we were able to categorise BFYP students in terms of how accurately they assessed their performance in a stoichiometry test before and after instruction.

The "acceptable" margin of error that we have chosen for the purpose of categorisation is a heuristic decision made within the specific context of chemistry students in an academic development programme. It is widely recognised that metacognitive skills are poorly developed in weaker students (Kruger and Dunning, 1999; Dunning *et al.*, 2003, Carvalho, 2009). Our results indicated that students in the top performing quartile of the posttest still overestimated their performance by an average margin of 9% as was shown in Figure 3. The decision was further informed by the fact that stoichiometry was assessed in this study with multiple-choice test items where calibration is notoriously difficult (Carvalho, 2009), especially for students with low metacognitive ability. The results are shown in Table 2.

The number of students who were overconfident in their judgment remained fairly constant, *i.e.* from 69% in the pretest to 71% in the posttest. This means that approximately 70% of our sample 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. Students who were realistic in their judgement decreased marginally from 31% in the pretest to 26% in the posttest. Only 2% of the students became under-confident in the posttest. There are no clear gender trends

in the accuracy of self-evaluation, neither in the pre- or posttest.

Table 2 Categories of students in terms of accuracy of performance evaluation in the pre- and posttest

† *OC – Overconfident, R – Realist, UC – Under-confident*

Category†	PRETEST			Category†	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
Total	91‡	35	55	**Total**	91‡	35	55

‡ *One record with gender information omitted.*

Despite the fairly stable division of accuracy of judgment depicted in Table 2 there were shifts of students who were able to show an improvement in terms of accuracy in performance evaluation after instruction and those whose ability to do so deteriorated. Students were then categorised based on pre-post accuracy of performance evaluation. The results are reported as a two-way frequency table in Table 3.

Five subgroups were generated and labelled according to pre- and posttest categories as OC-OC (n = 50), OC-R (n = 13), R-R (n = 11), R-OC (n = 15) and the R-UC (n = 2). The majority of students showed no improvement in their accuracy of performance evaluation (OC-OC subgroup, 55%). The number of students who acquired the skill of reporting accurate self-evaluations of their performance (OC-R subgroup, 14%) was similar to the number of students who became overconfident of their performance upon exposure to teaching (R-OC subgroup, 16%).Eleven students were realistic in their judgments both before and after instruction (R-R subgroup, 12 %). Only two of the 28 students who were realistic in their pre-test performance evaluation became under-confident in the posttest (R-UC subgroup, 2%).

Table 3 Shifts in accuracy of performance evaluation 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**

¶OC – Overconfident, R – Realist, UC - Under-confident

Learning gain associated with shifts in accuracy of performance evaluation (Research question four)

Having separated the students into five subgroups based on the accuracy with which they evaluated their performance in the pre- and posttests, we wanted to investigate the relationship beween accuracy of performance evaluation and learning gain in stoichiometry (research question four). Such results would indicate whether accuracy of judgment as a metacognitive skill is a desired attribute for the learning of chemistry, specifically in stoichiometry. Research question four was investigated for four subgroups. The fifth subgroup, R-UC, was too small for meaningful inferences to be made. 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 prerequisite 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.

Learning gain is a variable which provides a measure of the extent of improvement in performance but it must be normalised against scope for improvement in order to compare individuals or groups with different levels of preknowledge (Hake, 1998). In our study normalised learning gain was calculated for individual students and the mean value was determined for each of the four subgroups. Normalising learning gain against room for improvement in our case yields the following equation:

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

where 19 minus the prescore represents the room for improvement on a test with 19 items. Table 4 is an overview of how the four subgroups compared in terms of performance, pass rates and the average normalised learning gain achieved, calculated as suggested by Hake (1998). Table 4 is divided into four quadrants. Each quadrant represents a performance evaluation subgroup. The results for the first year of implementation are discussed in detail below.

Statistical analyses were conducted to investigate whether the differences among the four subgroups shown in Table 4 were significant. The Kruskal-Wallis test was selected for this purpose based on the fact that the data set were skewed and the sizes of the subgroups were small. This test is the non-parametric analog to an ANOVA. The results indicated that there was no statistical difference among pretest performances of the subgroups at a 5% level of significance. Even though a significant difference was found for the performance of the subgroups in CMY 133 (p = 0.0435), from the post hoc tests, no significant differences were observed between any pairs of subgroups (the largest difference was observed between the OC-R and OC-OC subgroups, p = 0.055). We interpreted these results to mean that despite small differences, the four subgroups can be assumed to be comparable in terms of prior knowledge in stiochiometry based on pretest and CMY 133 performance, but the OC-R subgroup may have been

more able or better prepared than the OC-OC subgroup based on previous semester achievement.

It was further established that the students in the four subgroups differed significantly in terms of posttest performance ($p = 0.001$) and learning gain ($p < 0.001$). Statistically 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. This means that 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), nor the two subgroups that were overconfident in the posttest (OC-OC and R-OC). Post hoc tests between learning gains of subgroups showed that there were statistically significant differences between all the subgroups except between R-R and OC-OC, and between R-R and OC-R ($p = 0.0636$). Pretest performance averages were lower than the required passing mark of 50% for all the subgroups (33% to 45%) with the R-R obtaining the highest average score and OC-OC the lowest. The performance in the posttest was substantially higher than performance in the pretest, ranging from 45% to 68%. The OC-R subgroup achieved the highest average performance score in the posttest and the R-R subgroup managed to achieve an almost 100% pass rate. However, there was no improvement in the pass rate for the students in the R-OC subgroup, which corresponds with an insignificant increase in average performance from the pretest (41%) to the posttest (43%). According to Table 4 the best posttest performance and the most meaningful improvement were demonstrated by the OC-R and R-R subgroups. The proportion of students in these subgroups that passed the posttest was more than 50 percentage points higher than the proportion that passed the pretest. The learning gain of the OC-OC group was moderate and the R-OC subgroup did not achieve any learning gain at all. These findings confirmed that we were dealing with four discrete subgroups with different characteristics. To put the results for learning gain into perspective: The 49% gain demonstrated by the OC-R subgroup compares favourably with the best results achieved in mechanics where interactive teaching methods were used (Hake, 1998).

This study was repeated in the following year with the new intake of BFYP students ($N = 300$). Students were again categorised in terms of shifts in accuracy of performance evaluation and normalised learning gain was calculated for each student. The results are included in table 4. Three students were categorised as under-confident, two in the pretest and one in the posttest. The distribution of the other students was similar to that reported for Year 1: The majority of students belonged to the OC-OC subgroup (70%) with the remainder distributed fairly evenly between the other three subgroups. The average learning gain was higher than the previous year for all subgroups, but, most importantly, the intricate relationships between shifts in accuracy of performance evaluation and learning gain was confirmed. Students who were realistic in their performance judgement in the posttest achieved the highest learning gain, with 53% and 43% demonstrated by the OC-R and R-R subgroups, respectively. Students who remained or became overconfident in the posttest achieved considerably lower learning gains, with 32% and 15% demonstrated by the OC-OC and R-OC subgroups, respectively.

Conclusions

The majority of the students in this study 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. A small number of students showed improved metacognitive monitoring after instruction but a similar number of students developed confidence in their performance that was unjustified. Surprisingly, our results suggest that academic overconfidence was not a crippling disposition, provided that exposure to subject content and learning opportunities resulted in an improvement in performance evaluation, as in the case of the OC-R subgroup.

An initial positive bias in performance evaluation may actually be beneficial to learning. Inaccuracy in self-evaluation in the pretest did not hamper learning for both the OC-OC and OC-R subgroups, but when overconfidence persisted despite teaching and learning (OC-OC) or developed upon exposure to subject content (R-OC) it had serious consequences. Students 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, the R-OC subgroup, were shown to be the most vulnerable based on their poor learning gain. Together, these two subgroups that were overconfident in the posttest represent 72% of our sample in year 1 and 82% of the sample in year 2.

In their normal practice BFYP teachers concentrated on the teaching and learning of stoichiometry and were not focussed on developing metacognitive monitoring skills as well. Our results suggest that students are slow to develop accurate metacognitive monitoring skills within a classroom environment that did not include instruction focused on the development of such skills. Students who improved their metacognitive monitoring also showed the highest mean learning gain, but simultaneous mastery of cognitive and metacognitive skills was achieved without an explicit intervention by a mere 11% or 14% of our sample, the OC-R subgroup. We conclude, therefore, that instructional design for under-prepared students should focus on development of both kinds of skills if risk of failure is to be averted. Instruction should focus on the teaching of specific monitoring and regulatory strategies that students can use in academic tasks such as preparation for summative assessment and test-taking. Our findings suggest that overconfidence may arise due to an illusion of knowing where knowing is compromised by the presence of misconceptions. Overconfidence may also arise where mathematical skills are inadequate and mistakes go unnoticed. Assessment practices as well as the quality and intervals of feedback provided by the educators could be improved with the aim of making students aware of what they know and do not know. Tests could consist of tasks that require higher cognitive demand and deeper engagement which may force students to critically and realistically judge their performance. Student generated submicro diagrams can also be used as a teaching tool

Table 4 Pre- and posttest performance data according to performance evaluation subgroup

	POST OC (*Overconfident in posttest*)			**POST R** (*Realistic in posttest*)		
PRE OC (*Overconfident in pretest*)		**Year 1**	**Year 2**		**Year 1**	**Year 2**
	Size of sample subset (% of sample)	***n* = 50 (55%)**	***n* = 207 (70%)**	Size of sample subset (% of sample)	***n* = 13 (14%)**	***n* = 33 (11%)**
	Average performance CMY 133	**50**	n/a	Average performance CMY 133	**61**	n/a
	Average Pretest performance (%)	**33**	28	Average Pretest performance (%)	**38**	35
	Average Posttest performance (%)	**45**	51	Average Posttest performance (%)	**68**	70
	% Pass Pretest	**10**	10	% Pass Pretest	**23**	27
	% Pass Posttest	**40**	58	% Pass Posttest	**77**	88
	Average Learning Gain (%)	**19**	32	Average Learning Gain (%)	**49**	53
PRE R (*Realistic in pretest*)	Size of sample subset (% of sample)	***n* = 15 (17%)**	***n* = 35 (12%)**	Size of sample subset (% of sample)	***n* = 11 (12%)**	***n* = 22 (7%)**
	Average performance CMY 133	**53**	n/a	Average performance CMY 133	**58**	n/a
	Average Pretest performance (%)	**41**	39	Average Pretest performance (%)	**45**	40
	Average Posttest performance (%)	**43**	48	Average Posttest performance (%)	**61**	66
	% Pass Pretest	**27**	20	% Pass Pretest	**36**	36
	% Pass Posttest	**27**	46	% Pass Posttest	**91**	100
	Average Learning Gain (%)	**-1**	15	Average Learning Gain (%)	**25**	43

to expose misconceptions and achieve mastery in stoichiometry (Davidowitz *et al*., 2010). These approaches may prevent the damage caused by failure and preserve the positive contribution of confidence, albeit excessively positive.

To conclude, we revisit our heuristic decision to allow an error in performance judgment equivalent to three questions in a test comprising of 19 items, i.e. 15.8%. This "acceptable" margin of error was chosen specific to our context in recognition of poor skills development of our sample, and the nature of subject content and the test instrument. However, students should become much better calibrated than this to avoid risk of failure in a challenging tertiary environment. Students in academic development programmes face numerous academic and personal challenges, but they also receive specialised support. Refining the art of accurate self-evaluation should be one of the objectives of such specialised support.

Acknowledgements

NRF funding
Gra□a Machel Scholarship for women
Jacqui Sommerville and Karien Adamski for statistical analysis.

Notes and references

[a] While judgments of confidence are commonly used in metacognition literature as an indication of perceived performance we acknowledge the potential ambiguity of this interpretation. The ambiguity about what exactly is measured by judgments of confidence warrants an in depth consideration by researchers in this field.

1. Bandura A., (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
2. Beyer S. & Bowden E. M., (1997), Gender differences in self-perceptions: convergent evidence from three measures of accuracy and bias, *Pers. Soc. Psychol. Rev.*, **23**, 157 – 180.
3. Bol L. and Hacker D., (2001), A comparison of the effects of practice tests and traditional review on performance and calibration., *J. Exp. Educ.*, **69**, 133 – 151.
4. Britner S. L. and Pajares F., (2006), Sources of science self-efficacy beliefs in middle school children, *J. Res. Sci. Teach.*, **43**, 485 – 499.
5. Campbell W. K., Goodie A. S., & Foster J. D., (2004), Narcissism, confidence, and risk attitude, *J. Behav. Decis. Making.*, **17**, 481-502.
6. Carter T. V. and Dunning D., (2008), Faulty self-assessment: why evaluating one's own competence is an intrinsically difficult task, *Soc. Pers. Psychol. Comp.*, **2**, 346 – 360.
7. Carvalho M. K. F. and Yuzawa M., (2001), The effects of social cues on confidence judgements mediated by knowledge and regulation of cognition, *J. Exp. Educ.*, **69**, 325 – 343.
8. Carvalho M. K. F., (2009), Confidence judgments in real classroom settings: monitoring performance in different types of tests, *Int. J. Psychol.*, **44**, 93 – 108.
9. Cooper M. M., Sandi-Urena S. and Stevens R., (2008), Reliable multi method assessment of metacognition use in chemistry problem solving, *Chem. Educ. Res. Pract.*, **9**, 18 – 24.
10. Davidowitz B., Chittleborough G. and Murray E., (2010). Student generated submicro diagrams: a useful tool for teaching and learning chemical equations and stoichiometry, *Chem Educ. Res. Pract.*, **11**, 154 – 164.
11. Dunning D., Johnson K., Ehrlinger J. and Kruger J., (2003), Why people fail to recognise their own incompetence, *Am. Psychol*, **12**, 83 – 87.
12. Dunning D., (2005), *Self-Insight: Roadblocks and detours on the path of knowing thyself*, New York: Psychology Press.
13. Dunlosky J., Serra M. J., Matvey G. and Rawson K. A., (2005), Second-Order Judgements About Judgements of Learning, *J. Gen Psychol.*, **132**, 335 – 346.
14. Ehrlinger J., (2008), Skill level, self-views and self-theories as sources of error in performance evaluation, *Soc. Pers. Psychol. Comp.*, **2**, 382 – 398.
15. Fernandez-Duque D. and Black S. E., (2007), Metacognitive judgment and denial of deficit: Evidence from frontotemporal dementia, *Judgm. Decis. Mak.,* **2**, 359 – 370.
16. Flavell J. H., (1979), Metacognition and cognitive monitoring: A new era of cognitive-developmental inquiry, *Am. Psychol.*, **34**, 906 – 911.
17. Gramzow R. H., Elliot A. J., Asher E. and McGregor H. A., (2003), Performance evaluation bias and academic performance: Some ways and some reasons why, *J. Res. Pers.*, **37**, 41 – 61.
18. Grimes P., (2002), The overconfident principles of economics student: An examination of a metacognitive skill, *J. Econ Educ.*, **33**, 15 – 30.
19. Hacker D.J., Bol L. and Bahbahani K., (2008), Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style, *Metacog. Learn.*, **3**, 101 – 121.
20. Hake R. R., (1998), Interactive-engagement *vs.* traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.*, **66**, 64 – 74.
21. Harrits G. S., (2011), More than method?: A discussion of paradigm differences within mixed methods research, *J. Mix. Method. Res.*, **5**, 150 – 166.
22. Hartman H. J., (2001), *Metacognition in Learning and Instruction: Theory Research and Practice*, Dordrecht: Kluwer Academic Publishers.
23. Hart J. T., (1965), Memory and the feeling-of-knowing experience, *J. Educ. Psychol.*, **56**, 208 – 216.
24. Hasan S., Bagayoko D. and Kelley E., (1999), Misconceptions and the certainty of response index (CRI), *Phys. Educ.,* **34,** 294–299.
25. Herron J. D., (1990), Research in Chemical Education: Results and Directions, in Gardner M. (ed.), in *Toward a scientific practice of science education*, Routledge.
26. Huddle P. A. and Pillay A. E., (1996), An In-Depth Study of Misconceptions in Stoichiometry and Chemical Equilibrium at a South African University, *J. Res. Sci. Teach.*, **33**, 65 – 77.
27. Johnstone A. H., (2010), You can't get there from here, *J. Chem. Educ.*, **87**, 22 – 29.
28. Johnstone A. H., (1991), Why is science difficult to learn? Things are seldom what they seem, *J. Comput. Assist. Lear.*, **7**, 75 – 83.
29. Kennedy E. J., Lawton L. and Plumlee L., (2002), Blissful ignorance: The problem of unrecognised incompetence and academic performance, *J. Marketing Educ.*, **24**, 243 – 252.
30. Koriat A. and Bjork R. A., (2005), Illusions of competence in monitoring one's knowledge during study, *J. Exp. Psychol. Learn.*, **31**, 187 – 194.
31. Kruger J. and Dunning D., (1999), Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated performance evaluations, *J. Pers. Soc. Psychol.*, **77**, 1121 – 1134.
32. Lazonby J., Morris J. and Waddington D., (1985), The mole: Questioning format can make a difference, *J. Chem. Educ.*, **62**, 60 – 61.
33. Leech N. L. and Onwuegbuzie A. J., (2011), Mixed research in counselling: Trends in the literature, *Meas. Eval. Couns. Dev.*, **44**, 169 – 180.
34. Lichtenstein S. and Fischhoff B., (1997), Do those who know more also know more about how much they know?, *Organ. Behav. Hum. Perf.*, **20**, 159 – 183.
35. Mabe P. A. III and West S. G., (1982), Validity of self-evaluation of ability: A review and meta-analysis, *J. Appl. Psychol.*, **67**, 280 – 296.
36. Maxwell J. A. and Loomis D. M., (2003), Mixed methods design: An alternative approach, in Tashakkori A. and Teddlie C. (eds.), in *Handbook of mixed methods in social and behavioural research*. India: Sage publications.
37. Nelson T. O. and Narens L., (1990), Metamemory: A theoretical framework and new findings, in Bower G. H. (ed.), in *The psychology of learning and motivation: advances in research and theory*. San Diego, California: Academic Press, Inc.
38. Nowell C. and Alston M. R., (2007), I thought I Got an A! Overconfidence Across the Economics Curriculum, *J. Econ. Educ.*, **38**, 131 – 142.
39. Ochse C., (2003), Are positive self-perceptions and expectancies really beneficial in an academic context?, *South Afr. J. High. Educ.*, **17**, 6 – 73.
40. Pajares F., (1996), Self-efficacy beliefs in academic settings, *Rev. Ed. Res., **66***, 543–578.
41. Pallier G., Wilkinson R., Danthiir V., Kleitman S., Knezevic G., Stankov L. and Roberts R. D., (2002), The role of individual differences in the accuracy of confidence judgments, *J. Gen. Psychol.*, ***129***, 257 – 299.
42. Potgieter M., Davidowitz B. and 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, *Proceedings of the 38th annual conference of the Australasian Science Education Research Association (ASERA),* Fremantle, WA, 2007.
43. Potgieter M., Malatje E., Gaigher E. and Venter E., (2010), Evaluation versus performance as indicator of the presence of

alternative conceptions and inadequate problem solving skills in mchanics, *Int. J. Sci. Educ*., **32**, 1407 – 1429.
44. Ridley D. S., Schutz P. A., Glanz R. S. and Weinstein C. E., (1992), Self-regulated learning: the interactive influence of metacognitive awareness and goal-setting, *J. Exp. Educ.,* **60**, 293 – 306.
45. Rozenblit L, and Keil F., (2002), The misunderstood limits of folk science: an illusion of explanatory depth, *Cognitive Sci.*, **26**, 521 – 562.
46. Rosenthal D. M., (2000), Consciousness, Content, and Metacognitive Judgments, *Conscious Cogn.*, **9**, 203 – 214.
47. Sandi-Urena S., Cooper M. M. and Gatlin T. A., (2011), Graduate teaching assistants' epistomological and metacognitive development, *Chem. Educ. Res. Pract*., **12**, 92-100.
48. Sandi-Urena S., Cooper M. M. and Stevens R. H., (2011), Enhancement of metacognition use and awareness by means of a collaborative intervention, *Int. J. Sci. Educ.*, **33**, 323-340.
49. Schaefer P. S., Williams C. C., Goodie A. S. and Campbell W. K., (2004), Overconfidence and the Big Five, *J. Res. Pers.*, **38**, 473 – 480.
50. Schwartz B. L. and Perfect T. J., (2002), Introduction: toward an applied metacognition, In Schwartz B. L. and Perfect T. J., (Eds.). *Applied Metacognition*. Cambridge University Press.
51. Schraw G., (2009), A conceptual analysis of five measures of metacognitive monitoring, *Metacog. Learn.*, **4**, 33 – 45.
52. Schraw G., Crippen K. J. and Hartley K., (2006), Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning, *Res. Sci. Ed.*, **36**, 111 – 139.
53. Zimmerman B. J., (2000), Self-efficacy: An essential motive to learn, *Contemp. Educ. Psychol*., **25**, 82 – 91.
54. Zohar A. and Dori Y.J., (2012), Metacognition in science education, Trends in current research in Ziedler, D. (ed.), in *Contemporary trends and issues in science education*, Springer. ISBN 978-94-007-2132-6, pp. 1 – 19.A.

