The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance

Sabina Kleitman *, Daniel S.J. Costa

School of Psychology, The University of Sydney, Australia

Abstract

There is increasing attention being paid to the role that metacognition, a key component of self-regulated learning, plays in learning, particularly adult learning. A novel formative assessment task, the Statistical Metacognitive Instrumentation Quizzes (Stats-mIQ) which provided extensive content and metacognitive feedback to students was developed for a senior undergraduate statistics course (N=214). The task consisted of multiple choice questions, each accompanied by a confidence rating in which participants were asked to rate how confident they were that their answer was correct. The feedback on their answers, their overall progress and metacognitive monitoring was provided within the quiz and after each quiz attempt. Path analysis was used to model the relationships amongst a range of psychological variables, for example quiz accuracy and confidence, and predicted actual exam performance. Demonstrating the effectiveness of the quizzes, the final exam mark was positively predicted by the total number of quiz attempts, a composite score comprising quiz accuracy and confidence, and students’ prediction of their exam mark. Total number of quiz attempts was positively predicted by a composite of conscientiousness and perceived time management skills, whilst quiz accuracy/confidence was positively predicted by the numeracy/statistics skills composite and reported need for feedback. Struggling students reported greater benefit from the confidence-allocation process than other students, highlighting the important role metacognitive feedback may play in a tertiary education setting.

Article info

Article history:
Received 18 June 2012
Received in revised form 7 November 2012
Accepted 11 December 2012

Keywords:
Computerised formative tools
E-Learning
Metacognition
Confidence
Statistics education
Self-regulated learning

1. Introduction

In educational research, metacognition and formative assessment tools have gained much attention, particularly through the self-regulated learning paradigm. Questions have been raised, however, about the lack of empirical evidence evaluating the effectiveness of formative assessment tools and the feedback they offer in real-life learning situations (Bennett, 2011). This paper examines the role of a novel computer-based teaching instrument—the Statistical Metacognitive Instrumentation Quizzes (Stats-mIQ)—used in a tertiary senior statistics course as a formative assessment tool. This novel e-learning teaching tool offered students both extensive content-based and metacognitive feedback. We will examine its effectiveness in improving the academic performance in a real-life learning situation and report on the implications to the theories of learning, individual differences and metacognition.

Statistics education is compulsory in Australia for students wishing to proceed in their specialised training in psychology (e.g., Masters, PhD programmes and/or having clinical supervision), making advanced statistical knowledge one of the key graduate attributes. Training in statistics entails the teaching of a range of complex and multi-faceted skills and knowledge. Furthermore, a variety of learning goals for students include understanding the purpose and logic of a range of statistical techniques, developing competency in their procedural details, understanding mathematical relationships and the nature of probability, and reading and appropriately interpreting statistical data (Gal & Garfield, 1997).

Formal course evaluations were conducted at the end of each teaching semester. Students enrolled in the advanced statistics course indicated in their comments that having more practical examples and receiving feedback on their performance would enhance their learning experience. These suggestions aligned with theories of learning advocating the key role that immediate feedback plays in acquiring and transferring knowledge (e.g., Kulik & Kulik, 1988). Thus, we developed a computer-based formative assessment tool in the form of several quizzes, each incorporating extensive immediate feedback on students’ performance. The feedback included content-based responses within each quiz item, different performance based indicators after each quiz, and their progress across all quiz attempts. This tool was available to students on a course-dedicated website.
Formative assessment refers to a particular type of assessment which is not ‘marked’ or judged by trainers/teachers (as summative assessment is). Instead, it offers insight for teachers and/or students in the form of feedback on performance/course (see Bennett, 2011 for a review). Accordingly, formative assessment tools often involve students undertaking a bank of structured items, typically accompanied by structured instructions and extensive feedback. This type of assessment is non-threatening: it provides students with additional practice which they can do at their own pace, thus reinforcing their interest in learning the subject matter and increasing their understanding of the material (e.g., Black & Wiliam, 1998, 2003; Gonzalez & Birch, 2000; Milheim & Martin, 1991).

Computer-assisted (or e-learning) formative tools have many additional benefits. They can provide immediate and effective response-specific feedback which students have found very useful (see Sosa, Berger, Saw, & Mary, 2011, for a review). This feedback can be corrective (indicate how accurate the user is), explanatory (offer in-depth reasoning as to why a particular answer is correct or incorrect), and/or support self-regulated learning via addressing misconceptions and biases the student may have about their performance and knowledge (Sosa et al., 2011). The Stats-mIQ provides all of the above. It consists of a series of multiple-choice quizzes accompanied by extensive feedback on each selected answer (explanatory and corrective), as well as overall performance feedback (corrective) and metacognitive feedback (addressing misconceptions and biases). These quizzes aimed to develop the level of understanding that is essential for the application and evaluation of basic research methods in psychology, including research design, multiple regression (MR) analysis and interpretation. The project was developed adhering to the principles of adult learning and was based in the theory of self-regulated learning. The application of a self-regulated learning approach is well suited to a statistics course. For example, Kesić, Baloglu, and Deniz (2011) demonstrated that students undertaking a statistics course who employed more self-regulated strategies reported less anxiety and more positive attitudes towards statistics.

1.1. Adult learning and self-regulated learning

A distinction between the features of adult learning and learning in younger individuals has been emphasised by many authors since the early 20th century (e.g., Thorndike, Bregman, Tilton, & Woodward, 1928). Knowles et al. outlined six assumptions underlying adult learning: (1) adults need to know why they need to learn; (2) adults have a need for self-direction; (3) adults have greater and different accumulated life experiences to younger learners; (4) adults apply learning to real-life situations; (5) adults take a problem- or task-centred approach to learning, as opposed to a subject-centred orientation; and (6) adults tend to be motivated by internal rather than external factors (Knowles, Holton, & Swanson, 2005). In summary, adult learning is characterised by a higher degree of self-direction, or self-regulation (Merriam, 2001).

Self-regulated learning incorporates the individual's ability to plan, manage, monitor, reflect on and regulate their own learning with personal autonomy and goal setting (e.g., Caffarella, 1993; Greene & Azevedo, 2007; Pintrich, 2000; Veenman, Van Hout-Wolters, & Afflerbach, 2006; Zimmerman, 2000). Metacognition is one of the three essential components of self-regulated learning, along with cognition and motivation (Schraw, Crippen, & Hartley, 2006). In this study, we will examine the role of a certain type of metacognitive feedback available via quizzes on academic outcomes, in addition to the role of content-related training.

1.1.1. Metacognition

Metacognition refers to the executive processes of reflecting on and regulating one's own thinking: that is, “thinking about thinking” (Flavell, 1979). These processes are central to learning (Azevedo, 2005; Efklides, 2006; Hacker, Dunlosky, & Graesser, 1998; Schraw et al., 2006), decision-making processes (Bruine de Bruin, Parker, & Fischhoff, 2007), and are a key ingredient of self-regulated learning (Azevedo, 2005; Schraw et al., 2006). Research in this area is providing insight into how real-world learning can be improved. This paper will focus on the metacognitive experience of “feeling of confidence” estimates, which have shown a clear and direct relationship with learning and decision-making in children and adults (e.g., Efklides, 2006; Efklides & Tsiora, 2002; Kleitman & Gibson, 2011; Kleitman & Moscrop, 2010; Kleitman, Stankov, Allwood, Young, & Mak, 2012; Stankov & Lee, 2008; Stankov, Lee, Luo, & Hogan, 2012; Stankov, Lee, & Paek, 2009).

There are different methods of assessing metacognitive processes, including self-report measures and more objective means which have a long history in psychology (see Lichtenstein, Fischhoff, & Phillips, 1982, for a review). One such method involves respondents being asked (immediately after responding to a question) to rate on a percentage scale how confident they are that their answer is correct. When multiple-choice items are employed, the minimum possible rating reflects the number of response alternatives offered. That is, for questions with four alternatives (as those employed in the quizzes), the lowest rating point is 25%, reflecting a chance to guess the answer correctly by randomly choosing one of the four provided alternatives.

There is empirical evidence that these metacognitive experiences of feeling of confidence are important to many decision-making and learning processes (e.g., Bruine de Bruin et al., 2007; Efklides, 2006; Efklides & Tsiora, 2002; Jackson and Kleitman, submitted for publication; Kleitman & Gibson, 2011; Kleitman & Moscrop, 2010; Kleitman et al., 2012; Stankov & Lee, 2008; Stankov et al., 2009; Stankov et al., 2012). They are especially pertinent to test-taking behaviour, as assigning confidence levels to one's answers helps to guide a student's choices and ongoing adjustment of strategies. Thus, it allows learners to reflect on actions and outcomes in learning contexts, helping them to detect and correct their errors. For instance, appropriate confidence that a given answer is correct allows the student to allocate mental resources to considering their answers to the items in which they have lower confidence. In addition, combined with immediate feedback on performance, these confidence ratings may provide students with the opportunity for self-reflection. For instance, a student may think that s/he answered a question correctly, thus stating a high confidence in that item. However, if the feedback indicates that this answer is incorrect, the student needs to consider why it is incorrect. S/he may realise that there is a need to learn more on this topic, as such feedback would indicate that s/he does not know as much about this matter as s/he thought. This type of self-evaluative, reflective thinking improves students' self-regulated learning and motivation (Lee, Paek, & Stankov, in press; Nicol & Macfarlane-Dick, 2006).

These confidence ratings can also be used both to assess levels of confidence across a range of tasks (confidence), and also to assess the match between confidence and accuracy (metacognitive self-monitoring). Thus, the objectivity of these confidence ratings can be verified by comparing them with actual performance (‘confidence’ minus ‘accuracy’), offering immediate and reliable on-task bias scores (Stankov, 1999; Stankov & Kleitman, 2008). Using this feedback, students are immediately confronted by the biases and misconceptions that they may have about their performance, promoting their self-regulative skills and learning. Feedback which addresses these misconceptions rather than only dealing with a lack of knowledge has been found to be a powerful stimulator of goal-oriented learning (Hattie & Timperley, 2007). These authors suggested that feedback is most effective when students are aware of their goal. In the context of the quizzes, this means that the goal is optimal performance (objective and subjective) and feedback is related to their attainment — whether students' subjective perception of their performance reflected by confidence judgments matches their objective
accuracy. Students exhibit an overconfidence bias if their confidence level exceeds their performance accuracy, and an underconfidence bias if that level is lower than is warranted by their performance.

Another metacognitive measure we will consider in this research is the prediction of final exam mark. Although little real-life research is available, many teachers of statistics and mathematics would attest to seeing highly anxious students convinced that they will perform poorly in exams who ultimately do well. Many hours and tissues are devoted to consoling these students prior to exams. On the other hand, instructors around the world tell stories of students who habitually challenge their assignment marks as not reflecting their assumed performance levels or their abilities. Thus, in this study, we will examine two bias scores — on-quiz bias in confidence ratings (on-task confidence levels minus quiz performance, averaged across all first attempts on quizzes) and exam prediction bias (exam mark prediction minus actual exam performance).

1.1.2. A novel measure of motivation

Another key component of self-regulated learning is motivation (e.g., Schraw et al., 2006). Although motivation is a very important construct to capture, general measures of motivation in the context of this course may produce limited variance. As mentioned above, the quizzes were used during the advanced senior level course, which is a prerequisite for Honours entry (a compulsory 4th year of training in Australia if one wishes to proceed to further training, including becoming a registered psychologist). The level of competition is very high, as only approximately 30% of students enrolled in the 3rd year course will be admitted (based on their marks only) into Honours. This naturally ensures that most students are highly motivated to succeed in this course.

Given that the quizzes provided response-specific metacognitive feedback pertaining to this learning environment, instead of measuring general motivational tendencies, we focus instead on the newly proposed construct of Need for Metacognition. This construct captures individual differences in tendencies to enjoy and to engage in thinking about thinking (a separate paper devoted to the development of this scale is currently in preparation). Aligned with the well-established Need for Cognition construct of Cacioppo and Petty (1982), who captured individual tendencies in enjoyment and to engagement in thinking, we have postulated a similar construct, but in relation to a metacognitive domain: that is, the motivation to be engaged in thinking about one's own abilities, knowledge and learning, as well as a need to understand how one thinks.

The construction of this novel scale was based in the general theory of metacognition and the need for cognition (Cacioppo & Petty, 1982). To create an item pool we used the original Schraw and Dennison (1994) Metacognitive Awareness Inventory (MAI), adjusting the items to capture tendencies to enjoy the engagement in acquiring knowledge relevant to self-awareness. After three iterations of the scale, the results of the confirmatory factor analysis suggested that this pool of items captured a hierarchical construct with a single broad second-order factor (Need for Metacognition) and four-first order factors — Need for Planning, Profiting from Assessing Learning Strategies, Profiting from monitoring/evaluation and Need for Feedback (Kleitman, in preparation).

1.1.3. Individual learning profile

Students have different attitudes and beliefs regarding statistics that are known to be influential in their academic performance (Gal, Ginsburg, & Schau, 1997). In this study, we aimed to capture individual differences in students’ learning competencies. We used the Individual Learning Profile (ILP, Pulford & Sohal, 2006) to assess self-perceptions of abilities relevant in the real-life learning environment — perceived competence/confidence in speaking and listening, reading and researching, writing, time management, numeracy skills and information technology (IT) skills. The questionnaire was first developed to identify students with low confidence in their abilities, at risk of failure, to enhance (through personal tutoring) students’ confidence and motivation (Pulford & Sohal, 2006). The results indicated that self-perceptions of time management predicted greater overall GPA (Level 1), explaining 4.9% of the variance in the GPA over all eight course modules. Self-perception of numeracy skills predicted performance on a module assessing statistical skills (Pulford & Sohal, 2006). Thus, we utilised this measure to capture self-perceptions of the different learning competencies of interest to learning experiences during a statistical course.

1.1.4. Personality factors

There are certain individual differences constructs that are implicated in learning outcomes. The typical suspects are personality dimensions, especially Conscientiousness (particularly the aspects of dutifulness, organisational and time-management) which has been shown to be associated with higher academic achievement (see Chamorro-Premuzic & Furnham, 2005; Paunonen & Ashton, 2001 for reviews). Thus, in this study, we utilised the Big 6 personality measure (Saucier, 2009), a personality measure that captures agreeableness, conscientiousness, extraversion, honesty/propriety, originality/intellect and resiliency dimensions.

1.2. Aims and hypotheses

The overarching aim of the project is to examine the effectiveness of the novel teaching tool, Stats-mIQ, on a real-life academic outcome (exam mark) and its expectation (prediction of the exam mark) in a senior level compulsory statistics course. In pursuit of this aim, we examined how the number of quiz attempts and the accuracy/confidence of performance predict these two learning outcomes. We also examined how useful students found the confidence-allocation process and metacognitive feedback, and their relationship with two biases (difference between on-task confidence and accuracy, and difference between prediction of exam mark and actual mark). The last two pursuits were exploratory (thus no hypotheses are offered), as limited literature exists in relation to both bias scores outside of the psychology lab and this is the first investigation explicitly asking students to evaluate their experience with the confidence procedure.

The relevant hypotheses are:

Hypothesis 1. Total number of quiz attempts, accuracy and confidence levels each positively predicts the exam mark and its prediction.

Hypothesis 2. Expected exam mark positively predicts the actual exam mark (although we expect that there will be individual differences in prediction bias which will be investigated separately).

Hypothesis 3. Conscientiousness and self-reported time management positively predict the number of quiz attempts, accuracy of quiz performance, the exam mark, and its prediction.

Hypothesis 4. Reported numeracy and statistical competence positively predict accuracy and confidence of quiz performance, the exam mark, and its prediction.

Hypothesis 5. Need for Metacognition (or its facets) positively predicts the accuracy and confidence of quiz performance, the exam mark, and its prediction.

We employed full questionnaires to capture the constructs in their entirety. However, no relationships between personality and learning competence variables (other than those for which we had specific hypotheses) were expected in relation to the real-life performance on the senior tertiary statistics course. This was confirmed by the results of this study. Thus, their discussion is omitted from this manuscript.
1.3. Statistical analyses

To examine Hypotheses 1–5, path analysis was conducted using MPlus (Muthén & Muthén, 2007). In this model, the independent variables were psychological constructs, which were posited to influence quiz-taking behaviour (the number of times students chose to complete them and their accuracy and confidence levels). The dependent variables were prediction of exam mark (2 weeks prior to the exam) and the actual exam mark. We also modelled the relationship between the students’ prediction of their exam mark and the actual mark. Path analysis enabled the investigation of all of the above-mentioned relationships simultaneously and incrementally to each other. The words ‘effect’ and ‘influence’ are used only for the sake of simplicity, and refer only to the predictive nature of the relationships between the different constructs. All other analyses were conducted in SPSS v19.

2. Method

2.1. Participants

214 (50 males, Mean age = 19 years, range: 17–50 years) psychology students participated in the study, all enrolled in a third-year psychological statistics course at the University of Sydney. These included seven students who dropped out early from the course without penalties and eight students who received an absent fail grade as they failed seven students who dropped out early from the course without penalties. Out of 199 remaining students, 160 students took at least one quiz (80.40%), and 149 (74.87%) and 92 (46.23%) students took part in Surveys 1 and 2 respectively. Ethics approval for the surveys was granted by the University of Sydney Human Research Ethics Committee (12159).

2.2. Materials

2.2.1. Online quizzes: formative assessment

A bank of 45 questions was developed and became available to students in the form of nine quizzes, in addition to the teaching tools already available in the course. Each quiz comprised a set of five questions, available on the course accompanied website, with extensive immediate feedback in both statistics and metacognitive self-monitoring (corrective, explanatory and addressing performance-related misconceptions and biases). Items covered content related to a typical senior level multiple linear regression course. For each question, students were asked to: 1) answer the question; and 2) rate their confidence in their response by entering a number between 25% and 100% in a text box. After students selected the answer and allocated their level of confidence, they were provided with extensive corrective and explanatory content-related feedback about why this was a correct or an incorrect response (see Figs. 1 and 2). Students were able to attempt the quizzes as many times as they wanted.

After each quiz attempt, the students were presented with a screen summarising their overall progress within this attempt and for all their attempts across all quizzes. For each quiz attempt, students were provided with relevant mean performance scores: the accuracy (% correct), the average confidence bias, as well as confidence scores for their correct and incorrect answers (see Fig. 3).

Advantages of the course compared to other courses were provided with extensive corrective and explanatory content-related feedback about why this was a correct or an incorrect response (see Figs. 1 and 2). Students were able to attempt the quizzes as many times as they wanted.

2.2.2. Examination performance

The overall MR exam mark (50% of the total exam mark) was used to index academic performance. It consisted of ten multiple-choice questions and three short-answer questions.

2.2.3. Expectation of performance

Students were asked to predict their exam mark for the MR part of the course (out of 100%) during the final teaching week.

2.2.4. Usefulness of Metacognitive Feedback

Students were asked to indicate how useful the confidence procedure was for helping them to understand their level of progress with the material, using a 5-point scale: not at all useful; of limited use; of some use; useful; and very useful.

2.2.5. Individual Learning Profile (ILP, Pulford & Sohal, 2006)

The ILP is a 43-item inventory assessing perceived abilities in speaking and listening, reading and researching, writing, time management, numeracy skills and information technology skills. Respondents were invited to report to what extent each question described their academic competence, on a scale of 1 to 5 (‘Never’, ‘Sometimes’, ‘Mostly’, and ‘Always’). Pearson correlations between measures collected one year apart on each of the six domains ranged from .49 to .75, and the Cronbach’s alpha coefficients of these six factors ranged from .74 to .93 (Pulford & Sohal, 2006).

However, prior to forming the relevant composites for the ILP measure, Exploratory Factor Analysis (EFA, Principle Axis Factoring with Promax rotation) was performed. This was undertaken to determine whether the stability of the facets of the scale originally reported using 1st year students enrolled in a psychology degree (Pulford & Sohal, 2006) were replicated using a sample of the students enrolled in the senior level statistics course. This was necessary as there were several questions originally included in the Numerical, “Hard” and “Easy” IT skills factors that asked the respondents to rate their confidence in a range of statistical skills (e.g., Are you confident using computers for statistics?: … for spreadsheets?: Are you confident about working with statistics?: … with graphs?: … with charts?). Given the nature of the course, it was expected that such items would merge into a single factor – Statistical Skills. The results revealed that this indeed was the case — a factor comprised of those five items was obtained. The other items which originally defined “Hard” and “Easy” IT skills merged into the broad IT Skills factors. The overall structure of the other factors remained as reported by the authors: Numeracy; Reading and Writing; Time Management; and Speaking Skills. All forthcoming analyses are based on the identified six latent factors and factor scores resulting from this EFA. Promoting awareness of performance-related misconceptions, students were given oral and written explanations about what over- and under-confidence meant, and informed that they should aim to have a bias score close to zero. It was also explained to students that they should aim to have higher levels of confidence in correct answers.

For all data analyses, only results from students’ first attempt of each quiz were used.

2.2.6. Big 6 personality measure (Saucier, 2009)

The Big 6 personality measure comprises 25 items covering six domains of personality: conscientiousness (e.g., “I like order”); agreeableness (e.g., “I am usually a patient person”); extraversion (e.g., “I usually

1 The Multiple Regression part of the Advanced Statistics course (taught for 2 weeks) was comprised of two weekly one-hour lectures, one two-hour tutorial, one major written summative assignment and discussion boards. The rest of the course (another 6 weeks) covered Analysis of Variance (or ANOVA) content which is outside of the scope of this paper.

2 The rest of the exam comprised questions covering the ANOVA content.

3 Further discussion of these results is outside of the scope of this paper, the results are available from the first author of this paper.
enjoy being with people”); honesty/propriety (e.g., “I sometimes misrepresent the facts”); originality/intellect (e.g., “I am an extraordinary person”); and resiliency (e.g., “I rarely worry”). Respondents used a five-point scale (‘Strongly Disagree’ to ‘Strongly Agree’) to report the extent to which each statement describes them. Except for Honesty/Propriety which included 5 items, each dimension was measured with 4 items. Trait scores were computed as the mean of trait-relevant items with higher scores reflecting greater self-perception of that trait. Internal alpha reliability estimates for each trait have been demonstrated to be reasonable for research purposes, ranging from .65 to .76 (Saucier, 2009).

2.2.7. Need for Metacognition (NFM) scale

The NFM scale is a newly-developed scale comprising 31 items. This is a preliminary version of the novel scale developed by the first author of this paper, and a separate manuscript devoted to this scale is forthcoming. The development of this scale is based on the theory of metacognition (Schraw & Dennison, 1994; Schraw and Moshman, 1995) and motivation, and it captures the perceptions of whether or not a respondent enjoys and seeks self-awareness with respect to their metacognitive knowledge and skills. The scale consists of four related facets: 1. Need for Planning (7 items; e.g., “I need to estimate how long a task should take before I start it” and “Projects are much easier for me if I set up a schedule to follow”); 2. Profiting from Assessing Learning Strategies (7 items; e.g., “I enjoy considering a variety of strategies to solve problems”, and “It is useful for me to analyse how I learn”); 3. Profiting from monitoring/evaluation (10 items; e.g., “It is helpful for me to evaluate how well I understand new material” and “I like to monitor whether my knowledge is improving”), and 4; Need for Feedback (7 items; e.g., “I look forward to getting feedback on my assignments” and “Even negative feedback helps my learning”). Participants used a five-point scale (‘Strongly Disagree’ to ‘Strongly Agree’) to report the extent to which each statement describes them.

2.3. Procedure

Students had access to the nine quizzes (released gradually, based on the material covered in lectures and tutorials) throughout the semester. Students completed these at their own pace and were
able to make multiple attempts. The demographic (e.g., gender and degree) and self-report measures (ILP, Big 6 measure, NFM scale and additional questions of interest) were split into two surveys which were made available to students to be answered once only at different times of semester. During administration of the second survey (anytime in the last week of the semester, approximately ten days prior to the final exam), students were also asked to predict their final examination mark for the unit of study and indicate how useful they found the confidence rating procedure.

### 3. Results

#### 3.1. Preliminary analysis

##### 3.1.1. Missing values analysis

About 19.6% of students enrolled in the course did not attempt any quizzes. Moreover, as Surveys 1 and 2 were completed voluntarily, the dataset contained some non-random missing values. In all analyses, all missing values were treated in a pairwise fashion. The relevant sample sizes are reported in the tables and figures below.

##### 3.1.2. Descriptive statistics and reliabilities

##### 3.1.2.1. Quiz performance and academic achievement

Descriptive statistics for quiz performance and academic achievement variables are summarised in Table 1, with Fig. 4 providing the overall summary of the number of quizzes students engaged in and the relevant accuracy, confidence and bias scores.

<table>
<thead>
<tr>
<th>Quiz #</th>
<th>Student</th>
<th>Attempt</th>
<th>Date</th>
<th>Accuracy (%)</th>
<th>Average Confidence (%)</th>
<th>Bias (%)</th>
<th>Confidence for Correct Answers</th>
<th>Confidence for Incorrect Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01</td>
<td>1</td>
<td>2009-10-18 09:42</td>
<td>60</td>
<td>62</td>
<td>2</td>
<td>75.0</td>
<td>42.5</td>
</tr>
<tr>
<td>2</td>
<td>02</td>
<td>1</td>
<td>2009-10-18 09:52</td>
<td>60</td>
<td>57</td>
<td>-3</td>
<td>61.7</td>
<td>50.0</td>
</tr>
<tr>
<td>3</td>
<td>03</td>
<td>-</td>
<td>2009-10-18 09:59</td>
<td>100</td>
<td>84</td>
<td>-16</td>
<td>84.0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>04</td>
<td>1</td>
<td>2009-10-18 10:08</td>
<td>60</td>
<td>67</td>
<td>7</td>
<td>86.7</td>
<td>37.5</td>
</tr>
<tr>
<td>5</td>
<td>05</td>
<td>1</td>
<td>2009-10-18 10:23</td>
<td>80</td>
<td>50</td>
<td>-30</td>
<td>56.2</td>
<td>25.0</td>
</tr>
<tr>
<td>6</td>
<td>06</td>
<td>1</td>
<td>2009-10-26 14:00</td>
<td>80</td>
<td>60</td>
<td>-20</td>
<td>68.8</td>
<td>25.0</td>
</tr>
<tr>
<td>7</td>
<td>07</td>
<td>1</td>
<td>2009-10-26 14:11</td>
<td>60</td>
<td>50</td>
<td>-10</td>
<td>58.3</td>
<td>37.5</td>
</tr>
<tr>
<td>8</td>
<td>08</td>
<td>1</td>
<td>2009-10-26 14:48</td>
<td>40</td>
<td>40</td>
<td>0</td>
<td>25.0</td>
<td>50.0</td>
</tr>
<tr>
<td>9</td>
<td>09</td>
<td>1</td>
<td>2009-10-27 10:36</td>
<td>100</td>
<td>85</td>
<td>-15</td>
<td>85.0</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>09</td>
<td>2</td>
<td>2011-11-09 18:50</td>
<td>80</td>
<td>85</td>
<td>5</td>
<td>93.8</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Fig. 3. Example of quiz attempts' feedback including a summary of all up to date attempts.

Overall, the level of engagement with the formative assessment (for all nine quizzes) varied between never doing the quizzes (22.6% of the students) and having 58 attempts (.5%), with 39.3% of students taking every quiz at least once throughout the semester. As evident in Fig. 4, students were most engaged in the first four quizzes. These four quizzes were released prior to their first major assessment in the MR part of the course, with the number of attempts declining as the pressures of the semester increased. Nevertheless, 48.1% of students completed more than 9 overall attempts across all quizzes.

The overall level of performance on the first attempt of the quizzes was reasonably high (67.34%), and consistent with the mean confidence estimate (68.99%), resulting in a low estimate of the overall on-task mean bias (1.65%). The overall levels of accuracy and

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Descriptive statistics for the quiz performance, examination performance and mark prediction variables.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Quizzes (N=164)</td>
<td>Number of attempts</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td>On-task bias</td>
</tr>
<tr>
<td>Overall course mark (N=199)</td>
<td>Overall course mark</td>
</tr>
<tr>
<td>MR exam mark (N=199)</td>
<td>MR exam mark</td>
</tr>
<tr>
<td>Examination mark prediction (N=101)</td>
<td>Examination mark prediction</td>
</tr>
<tr>
<td>Examination mark bias (N=101)</td>
<td>Examination mark bias</td>
</tr>
</tbody>
</table>

Note: Accuracy and confidence scores are calculated on the basis of the first attempt only.
confidences scores on the first attempt remained stable throughout the nine quizzes. This is not surprising given that each quiz covered distinct, yet related course material. Alphas for accuracy of performance and confidence across the nine quizzes were .77 and .94. These estimates are typically calculated to reflect the stability of performance within the test. However, when estimated across nine quizzes, they highlight similar psychometric information. That is, on-quiz bias scores ranged between 32.67% and 36%, and similarly the exam prediction bias ranged between 12.9% for the students who did not attempt any quizzes. Whenever students attempted the quizzes throughout the semester. These estimates can also serve as a proxy for the time stability estimates.

The mean overall course mark for the Advanced Statistics course was 66.31% (a Credit) and there was nothing unusual about the overall distribution of the marks. The overall exam performance on the MR part of the course was of a similar level (66.06%). For those students who undertook Survey 2, the mean prediction of their exam performance was remarkably close to their overall actual performance, resulting in a very small discrepancy between them (−.08%). The low estimates for both types of biases (on-quiz and exam mark prediction) suggest that overall the students were well-calibrated in their assessment of their own performance for the MR course content.

However, the spread of both bias scores suggests that there were students who were grossly under- and over-confident in their performance. That is, on-quiz bias scores ranged between −32.67% and 35%, and similarly the exam prediction bias ranged between −36% and 32%. Indeed, the overall distribution of the bias scores confirmed that overall low levels of both bias scores were simply a result of almost symmetrical distribution of under- and over-confidence scores for both biases. The histograms are presented in Fig. 5.

3.1.2.2. Quiz participation and exam expectation. To investigate the effectiveness of quizzes in improving the exam mark prediction, a separate analysis was performed and the pictorial presentation of the mean is summarised in Fig. 6. Students were split into six groups based on the number of times they attempted all nine quizzes during the semester: no participation (0 attempts; 19.6% of the overall cohort); few trials (between 1 and 4 attempts); limited participation (between 5 and 9 attempts, roughly 1 or less attempts on each quiz); moderate participation (between 10 and 18 attempts, roughly 1 to 2 attempts on each quiz); strong participation (between 19 and 27 attempts, about more than 2 to 3 attempts on each quiz); and extreme participation (more than 37 attempts, roughly more than 4 attempts per quiz).

As evident from Fig. 6, a discrepancy between the exam mark prediction and the actual mark is the highest (reflecting overly optimistic predictions, bias of 12.9%) for the students who did not attempt any quizzes. The bias is of a similar size and clusters within ±6% for all other groups. The size of the prediction bias for the group who never did the quizzes was statistically different to all other groups overall (F = 12.16, p < .01). Both linear and quadratic trends were also significant with the exam prediction bias declining steadily for students who had up to 18 attempts on the quizzes, before rising up somewhat afterwards but still staying close to zero (F_{linear} = 8.98, p < .01; F_{quadratic} = 5.50; p < .05). This suggests that even a small amount of participation in the quizzes was beneficial to foster realism of students’ prediction of their exam mark.

3.1.2.3. Self-report measures. Descriptive statistics and reliability estimates coefficients (Cronbach’s alpha) for self-report measure variables are summarised in Table 2.

Minor modifications in the calculations of the sub-scales of the ILP measure (see Method) resulted in very little change in the descriptive statistics (means ranged from 2.69 to 3.48) and reliability estimates (ranging between .65 and .93) for the original subscales of Numeracy, Reading and Writing, Speaking and Time Management Skills (Pulford & Sohal, 2006). Descriptive statistics for the personality variables were comparable with those reported in previous studies (Saucier, 2009). However, reliability estimates (alphas) were somewhat low (ranging from .56 to .64).
This version of the NFM scale was further refined in the forthcoming studies which are outside of the scope of this paper. Nevertheless, reliability estimates (Cronbach’s alphas) for this version of the scale are moderate (.77) to high (.82), indicating good internal stability of the responses for each facet of the scale. Respondents had similar levels of motivation (means ranged from 3.50 to 3.73) to gain and profit from metacognitive awareness for each facet, excepting the Need for Feedback facet of the scale, which had a notably higher mean (4.12). This is consistent with students’ requests from student evaluation questionnaires in previous years to have as much feedback on their learning as possible.

3.1.3. Correlations

Table 3 summarises the correlation coefficients between variables in this study.

There was an interesting, and at times strong pattern of correlations between the variables in this study, which is best summarised by the forthcoming path analysis. Several variables were combined into relevant composites to avoid collinearity problems. Only focal relationships between the variables are described here. In particular, the overall quiz accuracy and confidence scores were highly correlated at .62 (p < .01). This might be due to the nature of the quizzes and learning situation (their on-going nature) and the fact that the students may have utilised metacognitive feedback. Thus, the scores were combined into the accuracy/confidence composite for use in path analysis.

Similarly, Statistics Skills and Numeracy shared a strong correlation with each other (r = .63; p < .01), and consequently were combined to form the Numeracy/Stats composite which was used in the path analysis. The Conscientiousness facet of the personality scale shared a strong positive correlation with Time Management (r = .50, p < .01).
This relationship is not surprising as time management is an important component of conscientiousness (Saucier, 2009). Subsequently, to avoid a collinearity problem, the Time Management facet of ILP and the Conscientiousness facet of the personality scale were combined into a composite indexing time-managing tendencies. The use of these three composites was also warranted by important theoretical links between these constructs.

Although different facets of the NFM scale correlated positively with each other (between .22 and .46, p < .01), none of the facets showed any consistent pattern of relationships with other variables in this study, with the exception of Need for Feedback, which correlated positively with quiz and exam performance variables and the Numeracy, Statistical Skills and IT Skills facets of the ILP scale (r’s range between .26 and .38, p < .01). Consequently, only the Need for Feedback facet of the NFM scale was used in all further analyses.

### 3.2. Path analysis

To clarify relationships between exam mark and the other constructs under study, a path model was postulated. Our theoretical model takes into account variables that can be classiﬁed into three broad clusters: (1) psychological constructs (perceptions of Stats/Numerical learning competencies; Time Management/Conscientiousness personality characteristics; and Need for Feedback); (2) quiz-related variables (performance; quiz accuracy/conﬁdence and participation; total number of quiz attempts); and (3) different aspects of academic achievement (prediction of the exam mark and the actual exam mark). In this model, the dependent variables are exam mark and its prediction. The independent variables are Conscientiousness/Time management, Numeracy/stats competency composites and Need for Feedback.

We examined predictive relations between quiz-related variables (participation and performance) and academic outcomes (expectation and actual exam mark), whilst also examining the influence that psychological factors hold on these two clusters of variables. We also considered the relationships between the outcome measures.

### 3.3. Subjective appreciation of conﬁdence procedure and metacognitive feedback

There was about a 50–50 split in students’ attitudes with respect to the usefulness of the conﬁdence ratings procedure and metacognitive feedback for their learning outcomes. That is, 50% of students found the procedure to be very useful (33%), useful (24%) or of some use (23%), whilst another 50% reported it to be of limited use (37%) or not at all useful (13%). To investigate whether there were any consistent trends in relationship between these attitudes and levels/expectations of performance, we plotted the means of academic achievement/expectation variables as a function of ﬁve resulting groups (see Fig. 8).

As can be seen in Fig. 8, students with lowest levels of expectation and achievement found the conﬁdence rating procedure to be very useful for their learning process, compared to higher performing students who found the procedure to be of ‘no’ or ‘limited use’. Trend analyses revealed that the relationships between the performance/expectation variables and students’ attitudes to the usefulness of the confidence procedure were linear. That is, there were statistically signiﬁcant linear trends for the overall course mark (F(1, 82) = 7.65, p < .01), for the MR exam mark (F(1, 82) = 5.82, p < .05), and its prediction (F(1, 81) = 11.93, p < .001). This suggests that the students who tend (and expect) to perform more poorly tend to assess the conﬁdence procedure and metacognitive assessment teaching tool, Stats-mIQ, on a real-life academic

Thus, the number of quiz attempts and the quiz composite were each posited to predict the examination mark and its prediction, and the expectation of examination mark was posited to predict the examination mark.

As a first step, all possible paths between exogenous and endogenous variables were built into the model, making it redundant to test any alternative models. To simplify the model, the non-signiﬁcant paths were dropped. The path model is presented in Fig. 7.

The model had an excellent fit: \( \chi^2=12.29, p=.422 \), Comparative Fit Index (CFI) and Tucker–Lewis Index (TLI) = .99, Root Mean Square Error of Approximation (RMSEA) = .013 (.00 < RMSEA < .05), Standardised Root Mean Residual (SRMR) = .053 (see Byrne, 2010; Kline, 2005).

Consistent with Hypothesis 1, the number of quiz attempts positively predicted exam marks, but not its prediction. The Accuracy/confidence composite positively predicted both exam mark and its prediction. That is, confirming the efficacy of the formatative assessment tool, quiz-related variables were positively associated with exam mark and its prediction. Consistent with Hypothesis 2, expectation of the exam mark positively predicted the actual exam mark. Partly supporting Hypotheses 3, Conscientiousness/Time Management composite positively predicted the total number of quiz attempts, but not the exam mark. Similarly, Numeracy/Stats composite positively predicted both exam mark prediction and Accuracy/Conﬁdence composite, but not the exam mark (Hypothesis 4). Partly supporting Hypothesis 5, the Need for Feedback facet of the NFM positively predicted quiz Accuracy/Conﬁdence composite, but not the exam mark or its prediction. The model accounted for a notable amount of variance in the outcome variables: 62% of the variance in the number of quiz attempts; 25.7% of the variance in Accuracy/Conﬁdence composite; 41.2% of the variance in mark prediction; and 28.3% of the variance in the exam mark.

### 4. Discussion

In this paper, we examined the effectiveness of the novel formativ
The results of this study indicate that the degree of engagement with this novel computerised formative assessment and performance on the quizzes was associated with higher exam performance. Both of the quiz-performance variables—number of quiz attempts and quiz accuracy/confidence levels—positively predicted exam mark, and quiz accuracy/confidence scores predicted the exam mark expectation, which in turn positively predicted the actual exam mark. With respect to the students’ prediction of their exam marks, the results indicate that even a small degree of engagement with the quizzes was associated with higher accuracy of the exam mark prediction (minimising mark-prediction bias). Students’ solicited and unsolicited comments included that the quizzes were ‘an excellent resource’, ‘very helpful’, ‘often clarified points that I was confused about’, ‘feedback for each answer was really useful’, ‘love, love, love your revision quizzes. They really are very helpful’. Thus, the results confirm the efficacy of this formative assessment tool and its feedback to students’ objective learning outcomes, as well as their subjective learning experiences.

Importantly, students who performed poorly on the course reported the most benefit from the confidence-allocation procedure. These results are in agreement with several studies that showed that learners with low knowledge, who may not have strong self-regulatory learning skills, require more scaffolding (Azevedo, 2005; Chen, Fan, & Macredie, 2006; Moos & Azevedo, 2008). Particularly interesting were some unsolicited comments provided with respect to the confidence feedback: “I’m really interested to see the confidence feedback. Despite the anxiety that takes over under exam conditions, I have found trying to focus on the fact that I had low confidence (but achieved better than I thought) in the online quizzes helps me to remain a little more in check when it comes to understanding statistics. I realise most of the mistakes I’ve made in the course are either not reading the question properly, time management issues or making silly mistakes. So despite my reservations to the idea, I’ve actually found it very useful, thank you.” We believe that this student summarised well many of the benefits that this type of feedback offers: it helps to reduce the exam-related anxiety to students with low confidence and provides them with an anchor to re-consider and adjust their learning strategies.

Psychological factors that played a role in this real-life learning process were time conscientiousness/time-management, perceived numerical/statistical competence and the need for feedback facet of the Need for Metacognition scale: with students with better time management skills completing more quizzes; students with better numerical/stats skills having higher accuracy/confidence scores on quizzes, as well as higher mark prediction; and students with stronger need for feedback having higher accuracy/confidence levels.

Although encountering some important limitations (limited sample size, exploratory nature of some predictions and non-experimental design), this study carries with it a number of important implications for educational theory and practice, with respect to the effectiveness of computer-assisted formative tools (see Sosa et al., 2011, for a review). Supporting educational theory, these results suggest that computerised formative assessment tools with immediate response-specific content (both corrective and explanatory) and metacognitive feedback (addressing misconception and biases students had in their performance) may indeed be beneficial for overall academic performance and the quality of student experiences.

Theoretically, this study has also created an avenue for new directions into metacognition research, by examining individual differences in the novel motivational construct, Need for Metacognition: mapping a novel connection between the need for positive or negative feedback and higher levels of performance and confidence. Thus, with important caveats, the results of this study support the claim that metacognition is an important factor in self-regulated

| Table 3: Correlation coefficients for between quiz accuracy and confidence, examination performance, mark prediction variables and self-report measures (ILP, Need for Metacognition and Personality). |
|-----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Quiz accuracy 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Quiz confidence .62** | 1.00 | .62** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Overall exam mark .43** | .37** | 1.00 | .43** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Mark prediction .30** | .45** | .39** | 1.00 | .30** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| ILP: Numeracy .29** | .38** | .25** | .44** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| ILP: Writing .16* | .21* | .10 | .09 | .32** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Conscientiousness .03 | .01 | .09 | .28** | .20* | .05 | .34** | .23* | .19* | .31** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Agreeableness −.12 | .05 | .13 | .08 | .17 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 |
| Resiliency .02 | .11 | .09 | .17 | .03 | .10 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 | .15 | .13 |
| NFM: PFME −.12 | −.11 | −.10 | −.13 | −.14 | −.16 | .20* | .27* | .45** | .35** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| NFM: PALS −.15 | −.05 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 | −.01 |
| NFM: NFF .33** | .38** | .15 | .25** | .27** | .06 | .26* | .35** | .16 | .04 | .27** | .45** | .35** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Note: NFM = Need for Metacognition scale; NFP = Need for Planning; PALS = Profiling from Assessing Learning Strategies facet; PFME = Profiling from monitoring-evaluation. NFP = Need for Feedback, p < 0.05, **p < 0.01.
learning (Azevedo, 2005; Schraw et al., 2006). This is reflected by the result that enjoyment of receiving feedback acted as an important positive predictor of quiz performance, which in turn predicted academic achievement. Particularly important findings were indications that students who struggled with the course material reported to benefit more from the confidence-allocation process than other students. Considering the disadvantages and stress this group of students endures during compulsory statistical training, this additional feedback improved the quality of their overall learning experience within the course.

Although these results require replication and extension (using a much larger sample, a longitudinal design with follow ups and a larger selection of psychological variables), the study emphasises the important role the formative computer-assisted teaching tool (which included immediate response-specific content and metacognitive feedback) played in the tertiary learning environment.

Acknowledgements

This research was partially supported by two Small Teaching Improvement and Equipment Scheme Grants and a Research Infrastructure Block Grant given to the first author from the University of Sydney. Special thanks to the School of Psychology, University of Sydney IT team, especially Yohans Bastian who was invaluable with computerisation matters involved in the quizzes. Special thanks are also due to a team of tutors involved in creating a pool of items, especially Lisa Karlov (PhD) and Alex Russell. Special thanks to Simon Jackson, Suzie Moroney (PhD) and Sarah Young for their suggestions regarding the manuscript.

References


Jackson, A. S., & Kleitman, S. (submitted for publication). To cure, to kill, or to test? Decision-making tendencies in a medical paradigm: The role of individual differences in feelings of confidence.


