Assessment for Learning Tasks and the Peer Assessment Process

Lorraine Lauf
The University of Queensland
<s4115031@uq.edu.au>

Shelley Dole
The University of Queensland
<s.dole@uq.edu.au>

A program of Assessment for Learning (AfL) was implemented with 107 Year 12 students as part of their preparation for a major external test. Students completed extended mathematics tasks and selected student responses were used for peer assessment purposes. This paper reports on two of the AfL elements, namely task selection and peer assessment as part of the AfL process. The importance of initial task selection in terms of supporting students in building awareness of quality of mathematical arguments is highlighted.

Assessment for Learning (AfL) is a relatively recently coined phrase that, according to Lee (2006) is "a way of shaping learning using evidence of pupils' understanding" (p. 43). It is not about checklists or criteria, but rather describes the ways in which teachers observe and try to understand student learning, and then use that information to further future learning (Drummond, 2003). Black, Harrison, Lee, Marshall and Wiliam (2004) have defined AfL as any assessment that has as its main priority, the promotion of student learning rather than ranking, or accountability, or of certifying competence. Assessment can be classified as AfL if it provides information that teachers and students can use to modify teaching and learning activities (Black et al., 2004).

AfL is "a body of theory and a range of classroom practices" (Brooks & Tough, 2006, p. 5). It came about because of the need to use assessment in a positive way to both raise standards and to empower students (Black & Wiliam, 1998). Brookhart, Moss and Long (2008) have claimed that AfL "contributes to student ownership of learning more than any other classroom-based practice" (p. 54). The effectiveness of AfL, according to Brookhart (2007) is due to its impact on students' cognitive and motivational factors; it shows students where they are in their understanding (the cognitive factor) and develops in them a feeling of control over their learning (the motivational factor).

The classroom practices of AfL include four main techniques of questions (or tasks), sharing criteria, self and peer assessment, and feedback. The first technique, questions, is the tasks that students undertake in the assessment process. The students' responses to the questions are the focus of the AfL process and therefore suitable questions are those that require students to demonstrate their thinking and justify their solutions. The questions must be sufficiently open to ensure a range of responses and solution pathways. In order to complete the questions appropriately, students need to be aware of the criteria for assessment. The second AfL technique is about sharing criteria. Sadler (1987) has suggested that sharing criteria should be done by modelling exercises where the criteria are applied to specific work. Time to apply criteria is an important aspect of AfL as providing students with criteria alone is unhelpful if students don't know how to apply them to their work. One way of doing this is by using self and peer assessment, and this is the third technique in the AfL process.

Self assessment occurs when students evaluate their own work and make a judgement about its quality. Peer assessment is the same process but students look at the work of others. Self and peer assessment make unique contributions to the progress of learning as, through this process, students come to understand what counts as quality through examples. Feedback is the fourth AfL technique. Feedback can be written, oral or by demonstration, and can be provided individually or to a group (Brookhart, 2007). Feedback

L. Sparrow, B. Kissane, & C. Hurst (Eds.), *Shaping the future of mathematics education: Proceedings of the 33rd annual conference of the Mathematics Education Research Group of Australasia.* Fremantle: MERGA.

can come from the teacher or from peer assessment. The most useful feedback contains information that a student can use, in that it focuses on the quality of the work and provides suggestions on what to do to improve. This is particularly helpful to lower achieving students as it shows that effort can help them to improve (Boston, 2002).

Implementing AfL is not a simple process. Marshall and Drummond (2006) have stated that implementing AfL effectively is actually difficult to achieve in practice, possibly because there is no simple recipe for it. AfL will look different in every classroom and teachers need to develop ways of incorporating its ideas into their own practice (Wiliam, 2005). Successful AfL is also predicated on teachers' beliefs about the nature of learning and assessment, and is dependent upon the degree to which the teacher values student autonomy and makes this an explicit aim of their teaching. AfL therefore, challenges notions about assessment and the students' role in the assessment process. Few studies provide prescriptive guidelines for AfL, and for mathematics in particular, there is even more limited guidance on how to implement AfL (Wiliam, 2008), compounded by a dearth of tested tasks (Black, et al., 2004). Even less is published literature relating to AfL in a senior (Year 12) Mathematics classroom.

The Study

This study was undertaken with Year 12 students, to explore the potential of AfL in preparing them to undertake a major external exam. In the state in which this study was conducted (Queensland), there are no subject-based external examinations. Instead, all Year 12 students undertake external assessment in the form of a core skills test that examines knowledge and skills. This test, the Queensland Core Skills (QCS) test, is used to compare the achievement of students doing different subjects at different schools. QCS test results are used to scale school results in order to determine overall Year 12 positions of students. The QCS test is high-stakes, and many Queensland high schools spend much time preparing students for it. The test does not assess knowledge of specific Year 12 subjects but rather a set of 49 generic skills, called the Common Curriculum Elements (CCEs). Examples of these are: interrelating ideas/themes/issues; hypothesising; judging/evaluating; criticising; justifying; reaching a conclusion which is consistent with a given set of assumptions; translating from one form to another (QSA, 2009). Although generic, some CCEs are more specific to mathematics, as indicated in the following list: graphing; calculating; estimating numerical magnitude; approximating a numerical value; substituting in formulae; structuring/organising a mathematical argument.

The Queensland Studies Authority (QSA), as the administering body of the QCS test, states that Year 10 knowledge of mathematical operations is assumed. Mathematical (referred to as numeracy) items on the QCS are extended tasks that require higher order mathematical thinking, high-level analysis and problem solving skills. Statistics on school performance on the QCS test are available from the Queensland Studies Authority (QSA). For the school where the research was undertaken, results indicated that there was potential for improvement in students' scores on items that assess the common curriculum elements related to numeracy. The focus of this study was on the specific CCE of structuring/organising a mathematical argument.

The participants comprised all the Year 12 students (n = 107) enrolled at a girls' high school in Queensland, Australia. The majority of the girls were 16 or 17 years old. Data sources in this project included four extended mathematics tasks, a survey and interviews. The four tasks used in this study (Barbie, Pegs in Holes, Greek Flag, Pi), were selected on

the basis of open-endedness and requirement of a mathematical argument for solution. A brief description of each of these tasks follows.

The *Barbie* task provides students with a set of measurements of a Barbie doll and an 'average' teenage girl: height, leg length, waist, hips, feet, neck length and neck circumference. Students are required to use some of the measurements to develop a convincing mathematical argument showing why it is unrealistic for girls to aim to look like Barbie. The *Pegs in Holes* task asks students to explain which fits better: a square peg in a round hole or a round peg in a square hole? No other information is given and students are expected to set up their own sketch and dimensions of squares and circles. This could be done using formulae for areas of squares and circles, or using specific numbers for the dimensions. Students would need to define 'fit' and to calculate the proportion or percentage fit according to their definition. The *Greek Flag* task was taken from a 2008 QCS test. In this task, a picture of the flag is given, together with a brief description of the sizes of the stripes and the ratio of height to length. The question asks students to calculate the exact fraction of blue on the flag and explain their reasoning. The *Pi* task was also taken from a previous QCS test. This task requires the calculation of an incorrect value of π, given a description in words of how to perform the calculation. The task asks students to translate from words into algebra and requires students to show all steps.

The survey was printed on a double-sided A4 sheet of paper and consisted of four sections seeking information on: demographic details, and students' opinions on QCS in general, QCS preparation, and the AfL process. Apart from the background information items, all other items followed the same format, providing a statement and asking the students to indicate the degree to which they agreed or disagreed (using a 5-point Likert scale) with each statement. The survey was designed to take approximately 5-10 minutes to complete.

The interview consisted of four structured questions: How did you feel about the AfL process? Did the peer assessment component help? How did you find the AfL tasks? Do you feel nervous/apprehensive about the forthcoming QCS test? Please elaborate. The interview took approximately 10-15 minutes.

Procedure

The AfL process can be loosely described as a series of steps, with students presented with tasks, the criteria for task completion discussed and analysed, and students then undertaking the tasks, with student responses used for peer and self-assessment purposes. Students thus receive feedback on the quality of responses in relation to the criteria. To support students' learning, the process is repeated to enable students to apply their knowledge about criteria to new tasks.

In this study, this was the process followed. But, without clear guidelines, the study proceeded in a somewhat tentative and exploratory manner. Five QCS preparation lessons of approximately 40 minutes each were used for AfL in this study over a course of five weeks. In the first lesson, discussion of quality in relation to a mathematical argument was the focus. Criteria for assessment of extended mathematical tasks were shared with the students. Students were then presented with the *Barbie* task and were given time to discuss possible approaches to the task. Students began the task in the lesson then completed it for homework. Their completed responses were collected by the teacher/researcher who analysed all responses, selecting a variety to share with students in the following lesson. In the second lesson, students analysed the selected responses against the criteria. They were then given the second task, which was completed for homework. This cycle was repeated

for the third lesson and fourth lesson. In the final lesson, students completed the fourth task under test conditions.

Results

This paper focuses on results associated with the process of task selection and peer and self-assessment. Students' responses to Tasks 1 (*Barbie*), 2 (*Pegs in Holes*) and 3 (*Greek Flag*) showed great variety, and served as an excellent resource for self and peer assessment and feedback. Analysis and selection of tasks for the AfL process was for the purpose of stimulating discussion and providing students with opportunities to deepen their understanding of what constitutes quality in mathematical arguments. Therefore, students' performances on these tasks were not collated to determine the number of students who scored at particular levels, but rather were analysed for their potential for classroom use.

In the first lesson, to focus their attention on the concept of quality, students were given a chocolate. While they were eating the chocolate they were asked to find words to describe quality chocolate. This laid the groundwork to ask students to consider criteria for determining quality in relation to a mathematical argument. Standard criteria used in Queensland for assessing extended mathematics tasks were then displayed, and students recognised their own suggestions (but in different words) in these statements. In sharing criteria, it was hoped that students would develop greater awareness of the features of high quality responses to mathematics tasks: correct interpretation of the situation; use of effective and appropriate strategies to solve the problem; language and mathematics used accurately and appropriately; procedures justified; logical reasoning used to develop convincing arguments to support a conclusion or result. The *Barbie* task was then presented and students were given time to discuss possible solution pathways they might take. This discussion time was animated and sustained for approximately 20 minutes. Students completed the task at home and returned their responses the next day. Of the selected responses, the teacher/researcher typed the responses to ensure author anonymity.

Barbie responses 1, 2 and 3 were selected because they provided some impressive use of mathematical calculations and/or use of mathematical procedures, but only at a superficial level. *Barbie 1* provided many percentages and the dialogue was presented with authority. The following is an excerpt:

> With the careful use of ratios and percentages it was found that *Barbie's* leg length in real life would be 77% longer than an average teenage girl's…Many more convincing comparisons could be made if needed.

These words attempt to communicate mathematically but are too vague and do not explain the underlying calculations. Furthermore, the 77% is incorrect.

Barbie 2 took a much more structured approach, providing a data table. All the numbers were presented without explanation, but were correct. A ratio approach was used for the calculations. The response, however, did not use the calculated numbers effectively and did not produce a mathematical argument to answer the question. It seemed to expect the reader to come to conclusions without thorough explanation and communication.

Barbie 3 included some complicated ratios that were used to compare *Barbie* measurements with the teenage girl, as follows:

> A normal teenage girl has a bust-waist-hip measurement of 88: 72: 96. However when Barbie is enlarged to the average height, her body measures 81.5: 48.3: 75.4

The response did not communicate the basis or method of calculation. The response made other claims about *Barbie's* proportions but the numbers were not believable because they were not explained or justified. This response did not present a convincing argument.

When these first three responses were shown (consecutively) to the students and they were asked to comment on their quality, there was a general agreement that these responses were quite convincing, although they were quite weak in their mathematical argument. The students were impressed with the use of mathematical calculations and the authoritative presentation, referring to criteria statements to support their evaluations. The teacher/researcher needed to intervene to challenge students to critically interrogate the criteria as they analysed the selected responses. Through this process, students came to realise that these first three responses were not very convincing at all. When shown *Barbie 4* response, the students stated that it contained many unsubstantiated numbers and claims; that the numbers were not believable because there was no communication of how they were calculated.

Barbie 5 was deemed to be one of the best responses submitted. It explained clearly why a teenager's height is six times that of *Barbie*, as follows:

> When measuring your product it stands at 29cm, a teenage girl's height is approximately 175cm. That means a young girl is six times taller than the Barbie doll.

The response then used this factor of six to scale down the teenager's other measurements:

> If the proportions and measurements of the other body parts of the Barbie doll are to be realistic you would divide all the humans' measurements by six.

The response then concluded (correctly) that *Barbie's* waist was the dimension that was most significantly out of proportion.

> The most staggering discovery was the difference in the waist measurements. The current Barbie has a waist of 8cm, if this was to be realistic for the height of the Barbie you would divide the average waist of a human by six and find that it is 12cm.

When *Barbie 5* was presented, students commented that it was a simple, well-communicated response. They noted that it used a few well-chosen dimensions, explained the basis for calculations, and communicated findings clearly. For most students, this response highlighted the inadequacies of the previous four responses. The teacher/researcher had very little involvement in this discussion, as students were able to work this out for themselves. This was reassuring for the teacher/researcher as it seemed that the process of peer assessment was having an impact in terms of training students to recognise a quality mathematical argument. Students were then presented with *Pegs in Holes* and after discussing possible approaches, were required to complete it for homework.

For the *Pegs in Holes* task, *Pegs 1* response was selected because it was a fairly common incorrect answer. *Pegs 1* showed a circle with dimensions that fitted into a square with correct calculations of the areas for each. The respondent then tried to fit that particular square into the same circle, without changing the dimensions of either. Of course, the square would not fit and it was impossible to come to a conclusion about which was the better fit. The dimensions of one of the shapes would be required to change for there to be a fit. *Pegs 2* response was chosen because it was a well-presented, elegant solution. Numbers were selected to make the circle fit into the square, and areas were calculated. The method then kept the same square but increased the size of the circle to make the square fit into it. In order to compare the fit of both, the response compared the percentage of wasted space. The lower the percentage of wasted space, the better the fit

would be. *Pegs 3* response took a more algebraic approach because it did not use specific numbers for any of the dimensions – variables were used. It looked at the fraction of the total space used up by the inside shape. The more space used up, the better the fit. This approach demonstrated the concept for all squares and circles that fit into each other, providing a convincing mathematical argument using algebra. The third response concluded as follows:

> The circle in the square fits better as the circle takes up 78% of the squares area. A square in a circle the square only takes up 63% of the circles area. Therefore the better fit is the circle within the square.

By the time these three responses were presented, the students seemed to have become more discerning and were more critical when assessing their peers' work. In the third lesson of this AfL sequence, the students were presented with the *Greek Flag* task, and after initial discussion time, were required to complete it for homework.

The third task involved calculating the exact fraction of blue in the *Greek Flag*. Although this question appeared to be relatively straightforward, it was by far the most difficult task to complete correctly. Four responses were selected. *Flag 1* response gave the answer: one-half. This was a very superficial answer and it can be seen clearly from the drawing of the flag that this answer was wrong. *Flag 2* was a slightly better response because it attempted to explain that $^5/_9$ was blue. This is correct for one of the sections of the flag. *Flag 3* was selected because it was well set out and provided a clear explanation. It divided the flag into three sections, all clearly labelled (top left, top right and bottom), and calculated the amount of blue for each. It then calculated the whole area and explained how it worked out the fraction of blue:

> First I measured all the blue sections. Then I calculated the areas of the rectangles and squares by using the formula: L x W. I then measured the length and height of the whole flag and calculated its area. Fraction of blue = $\frac{\text{Total blue area}}{\text{Total area}}$

Unfortunately, despite the detailed explanation, it cannot yield the correct answer. It does not use the given ratio of height to length to determine the dimensions of the various blue sections. Instead, dimensions were measured using a ruler, with the associated measurement error. As a result, it was not possible to calculate the exact fraction of blue and the response did not answer the question, which was to calculate the exact fraction of blue. *Flag 3* demonstrated that, even though mathematical communication is very important, it needed to communicate correct mathematics that answered the question. *Flag 4* was selected because it was a correct answer with a clear explanation of the method. It showed clearly how it scaled up the given ratios of the dimensions of the flag to arrive at useful dimensions for the whole flag as shown:

Given	Height: Length
=	2 : 3
=	1 : 1.5
=	9 : 13.5

So the flag is 9 units high and 13.5 units long.

Total area = 9 x 13.5
= 121.5

The response divided the flag into three regions and correctly calculated the area of blue in each. It then proceeded in a similar way to *Flag 3* by adding up all of the blue areas, and working this out as a fraction of the total: $^{137}/_{243}$.

When shown the response by *Flag 1*, the students readily identified the limitation of the response, and similarly for *Flag 2* although the students rated *Flag 2* higher than *Flag 1*. For *Flag 3* however, they were more hesitant to identify the limitations of the response, as they were impressed with the clear presentation of the mathematical argument. The teacher then reminded students of the question and asked them if the exact, rather than estimated area had been identified in the responses. When students were presented with the response by *Flag 4*, they could see the depth of response but also realised the effort required to achieve such a response. When students realised that the task required an actual answer of $^{137}/_{243}$, there was a strong murmur of unrest throughout the room as the majority of students knew that they had not achieved this result. As a previous QCS task, the teacher was able to provide a second level of feedback to the students on each of the four selected responses using QCS criteria. Scoring of this QCS task was via a 7-point scale, with the criteria summarised as follows:

A = correct answer explained
B = one mechanical error
C = correct wording leading to a rounded answer
D = two aspects correct
E = one aspect correct
N = unintelligible response
O = no response

Only *Flag 4* would score an 'A' on a QCS test, with *Flag 3* scoring a 'C'. In the previous lesson, students were seen to readily attempt the *Greek Flag* task, but the majority of responses were a version of the *Flag 3* response – an estimate rather than an exact solution. Receiving this feedback was visibly unsettling at this point in the AfL program.

Discussion

The purpose of AfL is to promote students' learning, yet there are few specific guidelines, and none for Year 12 students preparing for a high-stakes test. In this study, the first challenge was finding appropriate tasks. According to Marshall and Drummond (2006), teachers implementing AfL are more successful if they select and sequence questions that demand high quality dialogue, deep thinking, and require clarification and refinement of answers. Such questions require challenging mathematical activity and advanced mathematical communication skills (Black et al., 2004). For this study, the literature provided little practical guidance. The teacher/researcher developed and modified tasks that were very different from each other, but appeared to serve as good AfL questions at two levels: first, their solutions were not immediately obvious and required students to devise a plan of attack, and second, they required extended responses and justification of the solution strategy. Appropriate task selection is the first and critical step in the AfL process as, according to Marshall and Drummond (2006), this affects all subsequent discussion and communication. When student responses to the *Barbie* and *Pegs in Holes* tasks were presented, students grappled with issues of quality in mathematical arguments but were seen to become more discerning in identifying quality as they were presented with more responses. The *Greek Flag* (Task 3), however, was a slightly different story. As previously stated, students had little problem interpreting the task and completing a response. The majority of students' responses to this task, however, were not of a high quality, with many students providing an estimated rather than precise solution. When students realised that few responses to the *Greek Flag* would result in anything higher than a 'C' rating, their confidence in discerning quality was visibly shaken. Knowing this was a previous QCS task, this task in the AfL at this time was quite detrimental. The impact of this particular task highlights the difficulty and consequences of task selection.

Through task analysis, students engaged in peer assessment, reviewing selected responses of their peers against given criteria. Black et al., (2004) have stated that it is through examples that students work out what counts as quality, and this study concurs with this finding. Peer assessment is a vital element in the AfL process. This study highlighted the developmental process of understanding criteria. Initially, students were not sufficiently critical of the mathematical arguments offered in their peers' responses, and students were seen to be convinced by weak arguments in some cases. A further issue that is not overtly addressed in the AfL literature is the public use of students' responses for peer and self-assessment. Although not a major issue in this study, the teacher/researcher was acutely aware of the potential negative impact to a student's self-esteem when her work sample is publicly presented to peers, and so took time to type the selected responses so that students' responses could not be identified by their handwriting. The students, however, appreciated being provided with work samples from their peers as they could compare their own responses to those on display.

Conclusion

This study investigated implementing AfL practices in a numeracy/mathematics context in Year 12. Trialling this in an authentic school situation has shown that this approach can readily be incorporated into a QCS preparation program. The difficulty was finding tasks and compiling a bank of responses, and dealing with students' simplistic solutions to complex tasks. In this study, students needed a lot of guidance to determine quality but, in the short time available, they demonstrated a noticeable increase in their awareness of quality in mathematical arguments. Further research into the effects upon numeracy outcomes for Year 12 students is warranted. This study is a tentative first step in an alternative approach to preparation for high-stakes tests, and the results from it are promising as they have direct practical implications for school mathematics programs.

References

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam (2004). Working inside the black box: Assessment for Learning in the classroom. *Phi Delta Kappan, 86*(1), 9-21.

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation, 8*(9), 1-6.

Brookhart, S. (2007). Feedback that fits. *Educational Leadership, 65*(4), 54-59.

Brookhart, S., Moss, C. & Long, B. (2008). Formative assessment that empowers. *Educational Leadership, 66*(3), 52-57.

Brooks, R. & Tough, S. (2006). *Assessment and Testing: Making space for teaching and learning*. London: Institute for Public Policy Research.

Drummond, M. J. (2003). *Assessing children's learning.* London: David Fulton.

Lee, C. (2006). *Language for Learning Mathematics: Assessment for Learning in Practice.* Berkshire: OUP.

Marshall, B. & Drummond, M. (2006). How teachers engage with Assessment for Learning: Lessons from the classroom, *Research Papers in Education, 21*(2), 133-149.

Queensland Studies Authority (2009). *About the QCS Test*. Retrieved from www.qsa.qld.edu.au/assessment**.**

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13* (2), 191-209.

Wiliam, D. (2005). Keeping learning on track: Formative assessment and the regulation of learning. In M. Coupland, J. Anderson & T. Spencer (Eds.), *Making mathematics vital*: (Proceedings of the 20th biennial conference of the Australian Association of Mathematics Teachers, pp. 20-34). Adelaide: AAMT.

Wiliam, D. (2008). *More about formative assessment*. Retrieved from www.kcl.ac.uk/ssp/education/research/iccamsfa.html