Paper presented at American Educational Research Association (AERA)
Annual Meeting, San Francisco, California, USA, 7-11 April, 2006

The Effects of Using Performance Assessment Tasks on Singapore Students' Learning of Mathematics

ZHU Yan & FAN Lianghuo
National Institute of Education, Singapore

Background of the Study

The inadequacies of traditional assessment, which is typically based on paper-and-pencil tests, conducted within classrooms over a limited amount of time (e.g., one or two hours) have been increasingly recognized by educational researchers over the last decades. There is no doubt that because of the dynamic nature of students' learning, it is basically impossible to use a single assessment mode to assess students for the full range of educational goals and instructional objectives. In particular, while traditional assessment is powerful in assessing students' factual knowledge, it is relatively weak in connecting knowledge being taught in school with students' experience in their daily life as well as in engaging students in critical thinking (Brown & Shavelson, 1994). To overcome the limitations of the traditional assessment is one of the important reasons for many researchers to call for the use of alternative assessment (e.g., see Fan, 2006).

A two-year project focusing on integrating new assessment strategies in mathematics classrooms has been conducted within Singapore educational settings since December 2003. The project was funded by the Ministry of Education (MOE) through the Centre for Research in Pedagogy and Practice (CRPP), National Institute of Education, Nanyang Technological University. The so-called new assessment strategies under investigation include communication tasks, performance assessment, project work, and self-assessment. Each new assessment mode was implemented in two primary and two secondary local schools for about one and half school-year. The effects of using the new assessment strategies on students' learning of mathematics have been investigated in both affective and cognitive domains.

As part of the project, the study presented herein focuses on the effects of using performance assessment tasks on students' learning of mathematics in one participating secondary school. Three research questions are addressed in the study:

(1) What are the effects of using performance assessment tasks in mathematics classrooms on students' attitudes toward mathematics and mathematics learning?
(2) What are the effects of using performance assessment tasks in mathematics classrooms on students' mathematics performance as measured in solving conventional problems?
(3) What are the effects of using performance assessment tasks in mathematics classrooms on students' mathematics performance as measured in solving unconventional problems?

It is hoped that the study can provide research-based evidence on the potential influences of using performance assessment tasks on students' mathematics learning so as to help school

teachers better align assessment practice with the desired educational goals and hence improve the quality of teaching and learning.

Theoretical Framework and Perspectives

'Performance assessment' is not a new term in education. Nevertheless, there is no consensus on its definition. According to Buechler (1992), the mergence of performance assessment movement was due to the fairly widespread dissatisfaction with high-stakes multiple-choice tests. It is believed that performance assessment has more pedagogical value and could more accurately reflect students' achievement than traditional multiple-choice tests (Kane, Khattri, Reeve, & Adamson, 1997). Gripps (1994) claimed that, in the United States, performance assessment was often regarded as any type of evaluation which was not multiple-choice or standardized testing. However, such a definition is rather broad and it almost covers all types of alternative assessment (e.g., project work).

In the Third International Mathematics and Science Study (TIMSS), performance assessment was included as one important component in the international comparison; in particular, performance assessment was referred to integrated and practical tasks targeting on students' content and procedural knowledge as well as their ability in using knowledge for reasoning and problem solving (Harmon et al., 1997). The Wisconsin Education Association Council (1996), at the root of the meaning of the word 'performance', set their definition as the one requiring students to demonstrate skills and competencies by performing or producing something. The central idea in Stenmark's (1991) definition about performance assessment is to assess what students actually know and can do. It is clear that while all the researchers tend to differentiate performance assessment from traditional assessment, they have different concerns and focuses. It also somehow reveals that there are diverse aspects involved in performance assessment.

To be more applicable to Singapore school education context, this study particularly defined performance assessment tasks as those having two distinguishing characteristics: authentic in context and open-ended in approaches and answers. In fact, these two aspects are to a large degree lacking in traditional assessment tasks, which consequently often received criticism over the last decades (e.g., see Howe & Jone, 1998; Wu, 1994).

The authenticity of a problem, according to the National Council of Teachers of Mathematics (NCTM) is the degree to which tasks are faithful, comprehensive, and complex, which can be found in important, real-life performances of adults that are non-routine yet meaningful and engaging for students (NCTM, 1995). It is believed that tasks with this feature could engage students in applying knowledge and skills they have learned in classroom to real-world challenges, and help them appreciate the usefulness of mathematics.

The open-endedness of a problem includes two aspects: (1) multiple venues of access or ways of solutions, and (2) multiple acceptable answers to the problem. It is believed that solving open-ended problems is more challenging than close-ended ones as students usually encounter in their school work, and normally requires higher-order thinking.

While the authenticity is more about task context, the open-endedness can be more used to assess students' ability in problem solving. As a result, all the performance assessment tasks used in this study are contextualized in a real-world scenario/story; they can be approached in various ways, from trial and error to more systematic methods, and ended with different answers (not just in different representation forms). Below is a sample performance assessment task.

> The sum of both a mother and her daughter's age was 30 in one particular year. Given that when the sum increased to 40, the mother's age was greater than three times of her daughter's, find all possible age of the mother when she gave birth to her daughter. (Hint: Let x be the mother's age when she gave birth to her daughter, hence when the daughter is y years old, the mother's age is $x + y$ years old)

This task is about age difference between a mother and her daughter, which is about the topic of inequality. The task shows students a fact that although the difference between two persons' ages is fixed, the ratio between the two persons' ages changes every year. It is a common sense but probably ignored by many people. To solve this task, one can use "trial and error" to find some possible ages for the mother without much difficulty. However, solely depending on this method, one would have much difficulty in getting all the answers. In fact, there are a total of seven different answers. More systematic methods can be employed, such as making a systematic list or establishing appropriate inequalities. Using the more effective methods, one can solve the task with much less difficulty. Due to the fact that the task can be solved via different methods, students can approach the task at their own levels and solve the task accordingly.

As a matter of fact, the Singapore mathematics syllabus also emphasized the importance of students' applying mathematics in solving real-life problems and being engaged in open-ended investigations in their learning of mathematics (MOE, 2002). However, an analysis of two widely-used Singapore secondary mathematics textbooks revealed that fewer than 2% of textbook tasks were authentic and about 2% were open-ended (Fan & Zhu, 2000). Moreover, according to Schoenfeld (1992), the beliefs that mathematics learning has little or no relation to the real world and that any mathematics task has one and only one answer are actually very common among students. In this connection, the study has both theoretical and practical significance.

Research Methods

As mentioned earlier, this study focuses on one participating secondary school, which is identified as a high-performing school, as it was randomly selected from the 50 best performing secondary schools according to year 1999 to year 2002 GCE "O" Level Examination results released by the MOE. The following sub-sections provide the detailed information about the participants, including both students and their mathematics teachers, the instruments used in this study, as well as the procedures of data collection and data analysis.

Participants

Thirty-eight Secondary One students from one class of high ability in the high-performing school were selected in this study to receive chapter-based interventions on performance assessment tasks during regular mathematics lessons for about three school-terms starting from early 2004. A parallel class of 40 students was chosen as an intact comparison group, which was taught as

usual during this period of time. No significant difference was found between the two classes in terms of students' Primary School Leaving Examination (PSLE) overall scores (t [76] = .81, p = .42) and mathematics grades (U [38, 40] = 747.00, p = .84). Table 1 provides the profiles of the participating students and their mathematics teachers from the two classes.

Table 1
A Profile of Participating Students and Their Mathematics Teachers

	Experimental Class	Comparison Class[1]
No. of students	38	40
Boys	24	25
Girls	14	15
Mathematics teachers		
Gender	Male	Male
Length of teaching experience	3 month	9 month
Qualification	M.Eng, PGDE[2]	MSc

Note. [1] In year 2005, the teacher teaching the experimental class also took over the comparison class; [2] PGDE stands for Postgraduate Diploma in Education.

The table shows that the two classes were taught by different teachers with basically equivalent professional background during the year 2004. However, due to some unforeseen reasons, starting from January 2005, the teacher teaching the experimental class had to take over the comparison class as well. Given the change, the teacher was advised not to use the intervention tasks in the comparison class so as to keep the teaching practices unchanged in both the classes in terms of interventions. In addition, the teacher from the experimental class received training and guidance on how to use performance assessment tasks in teaching before and during the intervention from the researchers.

Instruments

Three main instruments were designed for this study: questionnaires, performance assessment task tests, and intervention task worksheets.

1. Questionnaires. Two questionnaires were designed, one for the pre-intervention survey and the other for the post-intervention survey, to find out students' attitude toward mathematics and mathematics learning as well as their experience with performance assessment tasks. Both questionnaires contain two parts with all items in Likert-type scale format. The items in the first part (22 items) are the same in the both questionnaires focusing on students' perceptions about mathematics and their learning of mathematics, which targeted on four specific aspects: general view towards mathematics and mathematics learning, anxiety level in mathematics learning, perceptions of own performance in mathematics, and beliefs about the usefulness of mathematics. A nine-point scale ranging from "disagree totally" to "agree totally" is employed in this part.

The second part in the pre-intervention questionnaire (6 items) was intended to measure students' experience with various alternative assessment tasks (3 items relevant to performance assessment tasks) in their mathematics learning before intervention with a six-point scale on frequency, while the corresponding part in the post-intervention one (16 items) focused on

students' feeling about using performance assessment tasks in mathematics learning using the same scale as the first part.

A pilot test of the pre-intervention questionnaire was conducted in January 2004, involving 56 secondary one students from two other schools. The questionnaire was improved based on the results of the pilot tests. In particular, two items, one on general view and one on belief were finally removed from the pilot version in order to enhance the reliability level of the two sub-scales (general view: from .91 to .92; belief: from .78 to .80), with an average reliability being .85.

2. Performance assessment task tests. Similar to the questionnaires, two sets of parallel performance task tests, a pre-test and a post-test, were designed. The use of the pre-test enables researchers to have a better understanding about students' entry levels in mathematics problem solving, while the use of the post-test enables researchers to detect possible changes of students' ability in problem solving after three school terms with or without being exposed to performance assessment tasks in mathematics learning. Both tests contain three open-ended tasks, with one being also authentic.

A pilot study of the pre-test was conducted in February 2004 with 35 secondary one students from one school. Based on the students' feedback, necessary modifications on test tasks were made. The modified tasks were again piloted by a group of 36 secondary one students from another school in March 2004. As a result, while about 60% of the students felt that the tasks were challenging to them, all the students had no difficulty in understanding the tasks. Some minor modifications were further made in finalizing the pre-test items.

3. Intervention task worksheets. As the study emphasized the integration of performance assessment tasks into classroom teaching and learning. The design of the intervention tasks strictly followed the stipulated school scheme of work. Moreover, all the intervention tasks meet both the criteria as described earlier: authentic as well as open-ended. One or two performance assessment task worksheets were designed for each chapter by the researchers and necessary modifications were made by the participating teacher so as to better fit the students and teaching scheme. The researchers observed most interventions to monitor how the performance assessment tasks were carried out the classroom. The observations were also useful for the researchers to improve the design of future performance assessment tasks.

Data collection

The pre-intervention questionnaire survey was conducted in February 2004 for both the classes with a response rate being 100% and the post-intervention questionnaire survey was in May 2005 with a response rate being 81.6%.

The pre-test on performance assessment tasks was conducted in March/April 2004 with a response rate being 97.4% and the post-test was in May 2005 with a response rate being 82.4%. As performance assessment tasks assess more about students' ability in solving unconventional tasks, students' performance in solving conventional tasks were assessed via normal school exam scores. With the participating teachers' assistance, we were able to collect all the 78 students'

PSLE overall scores and mathematics grades (Exam A), year 2004 school mid-year mathematics exam scores (Exam B), year 2004 school final-year mathematics exam scores (Exam C), as well as year 2005 school first mathematics common test scores (Exam D).

During the three school terms, the experimental class managed to carry out a total of 12 interventions. In most cases, the interventions were recorded with field notes, or audio/video taping. Students' work was collected by the classroom teacher and then handed to the researchers for evaluation. After grading, a copy of students' work with researchers' comments was returned back to individual students for their revision.

Data process and analysis

The data from the two questionnaires were analyzed using quantitative methods. Descriptive statistics (e.g., frequency and percentage) was applied to describe students' overall perceptions about mathematics and mathematics learning. Mann-Whitney U tests were used to examine the possible differences between the two classes of students in each survey so as to enable researchers to detect the impact of using performance assessment tasks on the experimental students' attitudes.

Students' work in the two performance assessment task tests were graded based on task-specific rubrics by two independent researchers. The inter-rater reliability was calculated by the Intraclass Correlation Coefficient (ICC) on absolute agreement. As a result, the reliability on three performance criteria (i.e., Approaches, Solutions, and Presentation) over the three tasks for the two tests ranged from .98 to 1.00, with an average being .99. Similar to the analysis for the questionnaire data, the rubric-based grades from the performance assessment task tests were analyzed by descriptive statistics to investigate students' overall performance at class levels before and after intervention period. Mann-Whitney U tests were employed to identify possible differences between the experimental and comparison classes in each test. Wilcoxon Signed-Ranks tests were used to detect the change of students' grades from the pre- to post-tests. Moreover, possible differences on the changes between the two classes were examined by Mann-Whitney U tests to identify the potential relationship to the intervention program.

Students' PSLE overall scores and mathematics grades were compared by t-tests and Mann Whitney U tests respectively to ensure the equivalence of the experimental and comparison classes in terms of students' academic performance. The followed three exam scores were analyzed by 2 × 2 ANOVA with time (Exam B vs. Exam D; Exam B vs. Exam C; Exam C vs. Exam D) as a within-subject factor and treatment (experimental vs. comparison) as a between-subjects factor to investigate potential effects of using performance assessment tasks on students from the experimental class.

Limitations of the study

To investigate the effects of using performance assessment tasks on students' learning of mathematics, this study involved one experimental class and one parallel comparison class. Ideally, the comparison class should not be exposed to performance assessment tasks during the intervention period, whereby the experimental class do. However, in year 2005, the mathematics

teacher from the experimental class took over the comparison class due to unforeseen reasons. Although the teacher was explicitly asked not to try out intervention tasks in the comparison class, the experience of the teacher working with the experimental class could still influence his teaching in the comparison class one way or another, which can, to a more or less extent, affect the results of the study.

According to the research design, the experimental students should be exposed to the performance assessment tasks in a systematic and scheduled way. Nevertheless, due to some unexpected school activities, it was often very difficult for the teacher to do so in delivering the tasks to the students. As a result, in the first semester, the class only managed to carry out one intervention task.

In addition, while this study introduced performance assessment tasks to mathematics teaching in the experimental class, at the school level those students are still assessed based on the traditional assessment practice for their school performance grading and reporting. Such an inconsistency in the two domains could also have some negative influences on the results of the study (see more discussions in the next section).

Results and Discussions

The main findings of the study were reported below, based on the three research questions mentioned earlier.

Effects of using performance assessment tasks on students' attitudes toward mathematics and mathematics learning

The first part of the questionnaires was targeted on students' perceptions about mathematics and mathematics learning, including four specific aspects. The first aspect was about students' general views about mathematics and their learning of mathematics and six items were designed for this aspect. The data revealed that the two classes of students overall provided positive responses to these items in both the surveys. However, it was also found that the students in both the classes became more negative in the post- than the pre-survey in terms of average rating.

In general, there were no significant differences between the two classes of students in terms of their general views toward mathematics and mathematics learning, but the differences were in favor of the experimental class in both the surveys. However, it was also found that the difference between the two classes became smaller in the post- than pre-survey. In particular, while the experimental students appeared significantly more willing to spend time in studying math than the comparison students in the pre-survey (Q16: U [38, 39] = 552.50, $p < .05$, $r = .25$[1]), the responses between the two classes had no significant difference in the post-survey (Q16: U [33, 29] = 363.50, $p = .101$). A possible reason for this change is that the students from the experimental class had more opportunities to work on performance assessment tasks, which cost them much more efforts and longer time than the school mathematics tasks they usually encountered. Moreover, due to the challenging nature, more efforts and longer time actually do

[1] Effect size r is calculated when significant difference is detected. According to Cohen (1988, 1992), an r value over .5 is considered to be 'large', around .1 to be 'weak', and around .3 to be 'medium'.

not guarantee the students to get the tasks solved. All these factors may have some negative impact on students' perceptions about spending time on mathematics.

The second aspect, reflected in another six items designed, was about students' anxiety level in the learning of mathematics. While the students from the experimental class gave overall positive responses to all the relevant items in the two surveys, those from the comparison class provided negative responses to some items in the post-survey but all positive in the pre-survey. In particular, the comparison students in the post-survey expressed that they were somehow under terrible strain in mathematics lessons (Q2) and unconfident when came to mathematics (Q20).

While the students from both the classes became more anxious about their mathematics learning from the pre- to post-survey, the experimental students were consistently less anxious than the comparison students. Moreover, some differences between the two classes of students reached a significant level in the post-survey but none was found in the pre-survey (see Table 2). In particular, the experimental students were significantly less stressful (Q2), less afraid of (Q6), less nervous (Q17), and more confident about mathematics (Q20) than their counterparts and the effect sizes on the four items ranged from .33 to .42 with an average being .38.

Table 2
Comparison between the Experimental and Comparison Classes on Anxiety Level Items

	Pre-Survey	Post-Survey
Q2	654.50	293.50**
Q6	675.00	297.00**
Q10	677.50	357.00
Q14	600.00	403.50
Q17	570.00	322.00*
Q20	645.00	278.00**

Note. * $p < .05$, ** $p < .01$; The values in the tables are obtained by Mann-Whitney U-test, which examines the differences in the ranked positions of ratings between the experimental and comparison class.

Being exposed to performance assessment tasks appears to be one reason for such a result. As the performance assessment tasks are generally more challenging than normal school mathematics tasks, the experimental students then had more opportunities to be engaged in higher-order thinking via working on those tasks. Therefore, these students became less anxious about mathematics for both challenging tasks as well as normal school tasks.

The third aspect measured in the questionnaires was about students' perceptions about their own performance in mathematics. It also consisted of six items. The data from the two surveys revealed that the students were happy with their own performance except that the comparison students indicated that they did not like solving challenging mathematics problems in the post-survey (Q21). A comparison of the two classes of students' responses to this item actually showed that the experimental students were significantly more willing to attempt challenging mathematics tasks than the comparison students in the post-survey (U [33, 29] = 333.00, $p < .05$, $r = .30$) but no difference in the pre-survey (U [38, 40] = 671.00, $p = .369$). It appears that the

experience with performance assessment tasks helped the experimental students to establish confidence in working on challenging tasks.

However, we also found that while the experimental students had significantly stronger belief that they could do well in mathematics (Q15) than the comparison students in the pre-survey (U [38, 39] = 546.50, $p < .05$, $r = .26$), such a difference did not show again in the post-survey (U [33, 29] = 398.50, $p = .252$). It could be due to the fact that the performance assessment tasks were generally not easily solvable so that the relevant experience may lead the experimental students believe that they were somehow still weak in mathematics, especially challenging tasks, whereas the responses from the comparison students may only referred to the normal school mathematics tasks. Therefore, more clarification of students' interpretation of the term "mathematics" on this item is needed. On the other hand, the data revealed that the experimental students were consistently more positive toward their own performance in mathematics than the comparison students in the two surveys, although the students from both the classes gave more negative responses in the post- than pre-survey.

One main feature of the performance assessment tasks used in this study was the authenticity in the task context. Four items were designed to examine students' beliefs about the usefulness of mathematics. The results showed that the students from the two classes provided overall positive responses on all the relevant items in both the surveys. However, compared to the responses in the pre-survey, those in the post-survey were more negative. Such a change may be related to the fact that when students moved to higher grades, mathematics became more abstract and appeared further away from students' daily life. Moreover, we noticed that the negative changes by the experimental students were greater than those by the comparison students. In particular, the comparison students appeared more agreeable with the importance of knowing mathematics nowadays (Q8) and the meaningfulness of studying mathematics (Q12). Such a result may be related to the fact that while the experimental students were given many opportunities to work on performance assessment tasks which involved real-life application of mathematics knowledge, the skills they learned from the new assessment strategy, however, seldom were assessed in their school examinations. The inconsistent practice may bring students to downgrade the value of studying mathematics and in turn believe that studying mathematics (which was never been tested) was somehow wasting time. On the other hand, the conversations with the experimental students and their mathematics teacher actually revealed their appreciation of the real-life task setting, and some students did believe that some performance assessment tasks they have tried out were irrelevant to their normal school mathematics. The results reminded that teaching and assessment should be aligned consistently.

The second part of the questionnaires was targeted on students' experience with new assessment strategies, including authentic tasks as well as open-ended tasks. In particular, for the experimental students, they were asked about their perceptions about performance assessment tasks in the post-survey. The data from the pre-survey showed that both the classes of students had overall similar experience in doing the tasks with the aforementioned features. Basically, they worked on the tasks with the relevant features either in a monthly base or a weekly base. The comparison students' responses to the same items in the post-survey were not significantly different from those in the pre-survey, which indicates that the teaching practice in the

comparison class remained unchanged in terms of intervention and it is consistent with the research design.

Regarding their new experience with performance assessment tasks, the experimental students in the post-survey expressed their general acceptance of the specific features of the tasks, including multiple approaches of the tasks (Q26) and the authenticity in task contexts (Q30 & Q31). Moreover, the students believed that doing performance assessment tasks helped them to be more creative (Q27) and systematic (Q32). However, it seems that the students were still uncomfortable with the open-endedness in final answers. It is understandable that with previous school experience, students were often merely requested to provide one and only one correct answer to each task and they had already been used to such tradition and felt comfortable with it. The open-endedness in the final answers, in contrast, brought the students not only challenges but also somehow confusions, as commented by one experimental student.

The experimental students generally felt that doing performance assessment tasks were very challenging. In particular, more than 65% of the students claimed that they had to think harder in doing the tasks (Q28), 63% believed that it was time-consuming (Q35), about 30% felt lost in doing the tasks (Q29), and 47% asked for hints' help. In terms of the usefulness of doing performance assessment tasks, the majority of experimental students believed that such experience helped them in learning mathematics (Q25) and it was not wasting of their time (Q38). However, about 49% of the students did not think that doing performance assessment tasks could help them learn mathematics better (Q36) and only 38% indicated that they hoped to have more of such tasks in their future learning. Such a result could be related to the fact that performance assessment tasks have not been included in normal school assessment system. Therefore, the experimental students were unable to "see" the immediate benefit of doing the tasks and at least it does not seem to help them to get higher marks in the conventional school tests. Consequently, the students devalued the usefulness of doing performance assessment tasks in their mathematics learning and some became unwilling to have more in future study.

Effects of using mathematics performance assessment tasks on students' mathematics performance in solving conventional problems

Concerning students' performance in solving conventional mathematics problems, we used students' PSLE mathematics grades and end-of-semester assessment scores throughout the intervention period. As reported earlier, there was no significant difference in the students' PSLE overall scores as well as mathematics grades (Exam A) between the experimental and comparison students, which indicated the equivalence between the two classes. In Exam B, the equivalence still remained (t [75] = 0.015, p = .988). In fact, till year 2004 mid-year school examination, the experimental class only managed to implement one intervention. Therefore, no great change for the experimental class was expected. More interventions were carried out later on, as shown in Figure 1.

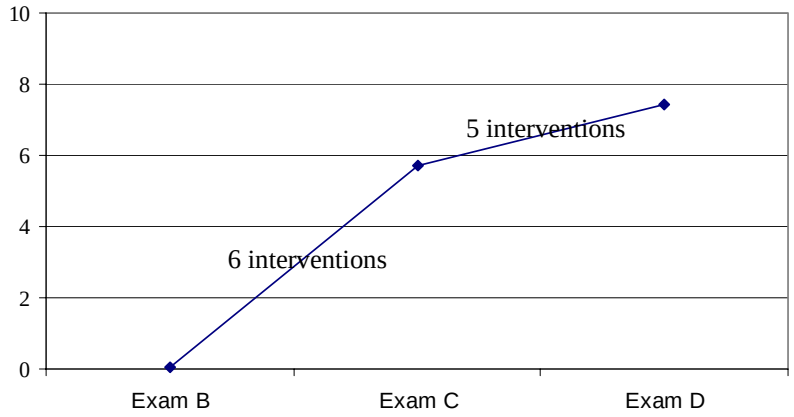

Figure 1. Difference between the experimental and comparison class on school exams

Note. The difference was calculated by the experimental class' mean score minus away the comparison class'.

From the figure, one can easily find that the differences between the two classes increase rapidly in the last two school exams. In particular, the two differences reached a considerable level (Exam C: t [74] = 1.935, p = .057, r = .22; Exam D: t [74] = 1.961, p = .054, r = .22). A repeated measurement analysis of various between Exam B and Exam D revealed that there was a significant interaction between time and treatment effects (F [1, 74] = 8.392, p < .005, r = .32) and the effect size is about medium, which is in favor of the experimental class. A further analysis revealed the significant interaction actually occurred between the period from Exam B and Exam C (F [1, 73] = 6.682, p < .005, r = .29) and in the next period (from Exam C to Exam D), the experimental class maintained the superiority. It appears clear that the students from the experimental class had an advantage of using performance assessment tasks when doing their conventional school exam tasks. It is also interesting to investigate how long the positive influence would maintain, which is, however, beyond the scope of the study.

Effects of using performance assessment tasks on students' mathematics performance in working on unconventional problems

Compared to the school exams, the performance assessment task test items are similar to the intervention tasks. All the test tasks are open-ended in approaches and final answers and one from each test is contextualized in a real-life scenario. Moreover, the tasks between the pre- and post-tests were designed in a parallel way so as to enable the researchers to identify possible relationship between the changes in students' performance and their experience with performance assessment tasks. In terms of the overall scores, the Wilcoxon Signed Ranks tests showed that the students from both the classes made significant improvement from the pre- to post-test (Experimental: Z = 3.598, p < .001, r = .62; Comparison: Z = 3.882, p < .001, r = .75). Moreover, no significant difference between the two classes was detected in either the test.

As all the tasks are open-ended in nature, it is more meaningful to examine students' performance in terms of their usage of effective strategies, the number of answers obtained, as well as the representation of solutions than their overall scores. As a matter of fact, these aspects

are the three performance rubrics designed in the tests. It is believed that the analysis on the sub-domains can provide more in-depth information on how students approach and solve such challenging unconventional mathematics problems, especially those from the experimental class. A brief description of the three performance rubrics is listed in Table 3.

Table 3
A Brief Description of General Rubrics by Approaches, Solutions, and Representation

	Level 0	Level 1	Level 2	Level 3	Level 4
Approaches (decision/strategy about approaching tasks)	No attempt or No evidence of a strategy	Strategy is ineffective and could not lead to any correct answer	Strategy could lead to correct answer but not systematic (e.g., guess and check)	Strategy shows partial systematic pattern	Strategy is effective that would lead to a complete set of answers
Solutions (no. of answers obtained)	No correct answer obtained	Only one correct answer obtained	More than one correct answer obtained	At least 50% of the full answers obtained	A complete set of answers obtained
Representation (documentation of problem solving procedures)	No attempt or Working is irrelevant	Working is not clear and hard to read	Working is not organized so that the approach is not observable	Working is organized and approach is partially observable	Working is well organized and approach is fully observable

Regarding the approaches employed by the students, the data revealed that in most cases, the students were able to use more systematic/effective methods in the post- than pre-test. That is, more students received a mean score over 2 on this performance scale in the post- than pre-test (Experimental: 79.4% vs. 42.1%; Comparison: 82.8% vs. 23.7%). The improvements in both the classes reached a significant level (Experimental: $Z = 3.929$, $p < .001$, $r = .67$; Comparison: $Z = 4.115$, $p < .001$, $r = .79$) but no significant difference was found between the classes in terms of the improvements.

As indicated earlier, all the tasks in the tests contain more than one correct answer, as listed below:

	Pre-Test	Post-Test
Task 1	7	10
Task 2	56	25
Task 3	2	2

Since that the last task in each test only had two answers, a task-specific rubric on solutions was set for the two tasks, shown as follows:

Task 3	Level 0	Level 1	Level 2	Level 3	Level 4
Solutions (no. of answers obtained)	• No correct answer obtained	• Only partial correct answer obtained, i.e., getting correct central number(s) Or • Answers obtained just by switching surrounding numbers without changing the central numbers	• One complete answer with different central number obtained	• Two correct central numbers with one complete answer obtained	• Two complete answers with different central numbers obtained

The analysis revealed that compared to the pre-test, the percentages of students who stopped at obtaining one correct answer (i.e., average score ≤ 1.33) were much smaller in the post-test for both the classes. In fact, the two classes of students made significant improvement in getting multiple correct answers from the pre- to post-test (Experimental: $Z = 3.39$, $p < .001$, $r = .58$; Special/Comparison: $Z = 3.787$, $p < .001$, $r = .73$). However, a between-class comparison did not display any significant difference regarding the improvements as well as students' performance on the particular performance rubrics in either the test.

It is believed that presentation is also an important skill in problem solving. Therefore, although it is not a focus of the intervention program, how students represent their solutions in the performance assessment task tests was also examined. The results revealed that the students generally did not have significant changes in their representation from the pre- to post-test. Moreover, consistent with the results on the other two performance rubrics, the two classes of students did not have significantly different performance on the aspect of representation in either the test.

Overall, it was found that both the experimental and comparison students made great improvement from the pre- to post-test, especially in the aspects of using effective strategies and getting multiple correct answers. On the other hand, one may also notice that while there was no significant between-class difference detected in both the tests, the advantage was on the side of the experimental class in the pre-test but the comparison class in the post-tests. The seemingly unexpected result may be related to the fact that the students from the experimental class knew well about the study and they were clear that all their grades on performance assessment tasks would not be counted into their school records. Correspondingly, it is possible that these students did not treat the post-test as seriously as their peers from the comparison class who were just given the tests without further information. However, the fact that the experimental students did not show significant different performance from their counterparts in the post-test, yet they still made significant progress, indicated that using performance assessment tasks were of no harm to students' learning of mathematics.

Conclusions and Implications

As part of two-year mathematics assessment project, this study was intended to investigate the effects of integrating performance assessment tasks into regular school mathematics teaching and learning. In the study, performance assessment tasks were defined as those contextualized in the real-world settings, approachable by various methods, and with multiple acceptable answers. Thirty-eight Secondary One students from one randomly selected high-performing Singapore local school received three-school-term intervention, with the other forty students as an intact comparison group. By using questionnaire surveys, performance assessment task tests, as well as students' school exam scores, the researchers investigated possible impact of using performance assessment tasks on the experimental students' mathematics learning in both academic and affective aspects.

Regarding students' academic achievement, the study looked into students' performance in both the conventional assessment (school exams) and unconventional assessment (i.e., performance assessment task tests). The results showed that the changes in students' performance across three

continuous school semester tests were significantly preferable to the experimental classes. Moreover, the favorite changes actually occurred after the intervention program had been implemented about one school year, where the experimental class completed 7 interventions, and maintained till the intervention ended. Although it is hard to attribute the positive result solely to students' experience with performance assessment tasks, it appears clear that the students from the experimental classes did benefit from being exposed to performance assessment tasks.

In the unconventional tests, the students from both the experimental and comparison classes performed significantly better in the post- than pre-tests, not only in terms of their overall scores but also in specific performance domains, including use of effective strategies and obtaining multiple answers. However, no significant difference was detected between the two classes regarding their progresses. It seemed that working on performance assessment tasks did not give the experimental students the advantage in unconventional tasks. As suggested earlier, one possible reason for such a result is that the students from the experimental class well knew that their performance in such tests would not be counted into their school records, which might have affected their performance in the tests. The result might also imply that developing students' ability to a higher level in solving challenging performance tasks could take a longer time than we have expected. In this regard, further evidence is needed before we can make a definite conclusion, which is beyond the scope of the study. Nevertheless, it is clear that no negative effect of using the new assessment strategy were found on students' performance.

Compared to the cognitive aspect, it is believed that the impact of using performance assessment tasks on students' attitude toward mathematics and mathematics learning is a more gradual procedure. Consistent with many other researchers' findings (e.g., Macnab & Payne, 2003; Wong, Lam, Wong, Leung, & Mok, 2001), the students in this study in general become more negative toward mathematics and mathematics learning in the post- than pre-survey in all the four sub-domains. On the other hand, the study also revealed that the changes were generally in favor of the students from the experimental class, especially the anxiety level about mathematics. However, it appears not expected that the changes on the view about the usefulness of mathematics were preferable to the students from the comparison class, while contextualization in real life was one important characteristic of performance assessment tasks. On the other hand, the experimental students expressed their appreciation of the specific features of performance assessment tasks in the second part of the post-survey. One possible reason for the seemingly contradictive results is that the students may find that what they experienced in the performance assessment tasks seldom appeared in their regular school mathematics learning. To them, normal school mathematics was more important, as it would really be tested. Correspondingly, the experimental students may have even stronger feelings that the mathematics they encountered in the regular school learning was farther from their daily life.

The results from the study suggested that teachers and students were capable of handling with performance assessment tasks. Although the effects of using the new strategy in some cases were not obvious, it is clear that no negative impact was observed. On the other hand, the study also observed some positive effects on students' academic achievement and anxiety about mathematics learning. We believe that it is pedagogically helpful to provide opportunity for students to work on problems contextualized in real life and open-ended investigations.

Finally, we would like to point out that this study was an initial step to explore the possible effects of using performance assessment tasks on both teachers' teaching and students' learning of mathematics. More research on the impact of using such new strategies on teaching and learning in various aspects is needed. Moreover, how to effectively and efficiently use the strategy for instruction also needs further investigation.

References

Brown, J., & Shavelson, R. (1994). Does your testing match your teaching style? *Instructor*, *103* (7), 86-89.

Buechler, M. (1992). Performance assessment. *Policy Bulletin*, No. PB-B13.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Fan, Lianghuo (2006). Making alternative assessment an integral part of instructional practice. In P. Y. Lee (Ed.), *Teaching secondary school mathematics: A resource book* (pp.343-354). Singapore: McGraw Hill.

Fan, L., & Zhu, Y. (2000). Problem solving in Singaporean secondary mathematics textbooks. *The Mathematics Educator*, *5*(1/2), 117-141.

Gripps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., et al. (1997). *Performance assessment in IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Howe, A. C., & Jones, L., (1998). *Engaging children in science* (2nd ed.). Upper Saddle River, NJ: Merrill.

Kane, M. B., Khattri, N., Reeve, A. L., & Adamson, R. J. (1997). *Assessment of student performance*. Washington, DC: Studies of Education Reform, Office of Education Research and Improvement, U.S. Department of Education.

Macnab, D. S., & Payne, F. (2003). Beliefs, attitudes and practices in mathematics teaching: Perceptions of Scottish primary school student teachers. *Journal of Education for Teaching*, *29*(1), 55-68.

Ministry of Education. (2002). *Mathematics syllabus* (*lower secondary*). Singapore: Curriculum Planning Division.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics*. Reston, VA: Author.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. A. Grouws (Ed.), *Handbook on mathematics teaching and learning* (pp. 334-370). New York: Macmillan.

Stenmark, J. (1991). *Mathematics assessment: Myths, models, good questions, and practical suggestions.* Reston, VA: National Council of Teachers of Mathematics.

Wisconsin Education Association Council. (1996). *Performance assessment*. Education Issues Series.

Wong, N.-Y., Lam, C.-C., Wong, K.-M. P., Leung, F. K.-S., & Mok, I. A.-C. (2001). Students' views of mathematics learning: A cross-sectional survey in Hong Kong. *Education Journal*, *29*(2), 37-59.

Wu, H. (1994). The role of open-ended problems in mathematics education. *Journal of Mathematics Behavior*, *13*(1), 115-128.