Assessment of Basic Competencies

Amy K.M. Cheung, Guanzhong Luo and Gregory Chan
Hong Kong Examinations and Assessment Authority

Abstract

In its 2000 report entitled *Learning for Life, Learning Through Life*, the Hong Kong Education Commission set out detailed proposals for assessing students' basic competencies in Chinese language, English language and mathematics. This paper reports on the progress in implementing these proposals. Two initiatives were undertaken. One set of initiatives has involved the development of a web-based student assessment system that recently won a silver medal for innovative excellence at *le Salon International Des Inventions 2005*. The other set of initiatives has involved paper and pencil testing of the whole cohort of students at the end of each key stage of schooling. This testing is conducted throughout the Hong Kong Special Administrative Region, China, and therefore, it is termed the Territory-wide System Assessment. This paper provides an overview of the measurement issues underpinning both assessments, including the calibration of items, equating of tests and the setting of standards. It also presents results of the first two years of implementation and the impact of the policy thus far.

Introduction

In its 2000 report entitled *Learning for Life, Learning through Life*, the Hong Kong Education Commission (EC) set out detailed proposals for Basic Competency Assessments (BCA) in Chinese Language, English Language and mathematics. The EC recommended that there be two components: Student Assessment (SA) and System Assessment. The Hong Kong Examinations & Assessment Authority (HKEAA) was commissioned in 2001 by the Education and Manpower Bureau (then Education Department) to develop and implement BCA in Chinese Language, English Language and Mathematics.

Student Assessment was to be implemented as an online system to provide instant feedback to students and teachers. This recommendation has been implemented and is fully operational for Primary 3 (Grade 3), Primary 6 (Grade 6) and Secondary 3 (Grade 9). The web-based Student Assessment system, which has recently won a silver medal for innovative excellence in a prestigious Geneva-based international competition (*le Salon International Des Inventions 2005*), allows teachers to review and improve progress towards learning objectives and set targets for students.

System Assessment, (later renamed 'Territory-wide System Assessment' (TSA)), was conceived as a low-stakes survey school performance at Primary 3, Primary 6 and Secondary 3 levels in the three subjects. The main purpose of System Assessment, as envisioned by the

Education Commission, was to provide the Government and school management with information on school standards in key learning areas for the purposes of school improvement, thus enabling Government to provide support to schools identified as needing assistance. The results were also seen as useful in monitoring the effectiveness of education policies.

The TSA began at P.3 level in 2004. In 2005, both Primary 3 and Primary 6 students took part in the TSA. In 2006, the TSA will be extended to the secondary level. All students at Primary 3, Primary 6 and Secondary 3 will take part in the TSA 2006.

Equating Items

Item requirements are set out in the Basic Competency (BC) documents of the Curriculum Development Institute (CDI). These documents provide a set of descriptors that encompass four skills in the Chinese and English Languages and also detail the concepts, knowledge, skills and applications covered in Mathematics in the following dimensions: Number, Measures, Shape & Space and Data Handling for Primary 3, with the addition of Algebra for Primary 6 and Secondary 3.

In order to provide schools and Government with comprehensive feedback, responses on a large number of items are required to assess all basic competencies associated with the different skills/dimensions. This makes it impractical for any one student to answer all items; however, this is not necessary since the aim of the TSA is to assess the overall standard of each school rather the individual standard of each student. For this reason, use is made of a number of sub-papers. Each of the sub-papers making up the TSA is designed to measure a set of basic competencies and includes items which overlap with those other sub-papers (see Table 1). It is noted that the raw scores of the students who take different sub-papers are not comparable. Using the Item Response Theory (IRT), however, the locations (logits) can be obtained for all students even if they take different sub-papers. As the locations of all students form a unidimensional variable, the comparison of the performance of students is simply the comparison of their locations: the greater the logit, the better the performance, no matter which sub-paper is taken.

Table 1. Illustrative Structure of the Test for English Language

Sub-papers\Item sets	1	2	3	4	5	6
Sub-paper 1	▒	▒	▒			
Sub-paper 2			▒	▒	▒	
Sub-paper 3		▒			▒	▒

In analysing the data on all sub-papers simultaneously using a computer software developed based on the IRT (e.g. RUMM2020 (Andrich, Sheridan & Luo 2003)), all items are also calibrated on the same logit scale. In other words, the sub-papers as different tests are equated. Therefore, it is possible to plots all students and all items in the same scale according to their locations. Similarly, the pretests can be equated with the

sub-papers. In the pre-test for TSA, a large number of items (120 -250 per subject) were evaluated using matrix sampling with at least 250 students per item and involving a total of 2500 students for each key stage. Sufficient item overlap between pre-test and sub-papers was arranged to enable IRT analysis. In addition items designed for different levels were embedded into some of the pre-test sub-papers as were items from the Student Assessment item bank. In doing so, a single scale, termed *the achievement scale* in this paper, is established across levels (i.e. Primary 3 and Primary 6 or Primary 6 and Secondary 3). This achievement scale also facilitates the maintaining of consistent standard across years. The design also ensured that the calibration of TSA items was in line with that of the student assessment item bank.

Both the online student assessment and the TSA have a speaking component. For Primary 3 and Primary 6 English, this involves individual work, for example, reading aloud of a unseen passage and a dialogue between the student and assessor. For all levels in Chinese and Secondary 3 English, the assessment of speaking involves both an individual presentation and a group discussion. The student performances are then rated according to rating criteria based on the basic competency descriptors relevant to each item. In the context of the student assessment, the class teacher administers the speaking component. In TSA student performances are rated by two trained assessors. Due to cost limitations the TSA oral is not administered to all students but rather to a sample of 12-24 students at each level from each school.

Standards Setting

In 2004, a standards setting exercise was carried out to set basic competency standards for each of the three subjects for Primary 3 students. A three-step process was adopted that blended technical, professional and policy-oriented considerations. This did not include oral items since the sampling process for TSA oral administration (mentioned earlier) did not yield large enough samples to yield meaningful results for schools.

The first step in the standards setting process was largely technical and involved equating the different tests so that it was possible to compare the performance of all students, regardless of which combination of sub-papers they took.

The second step was largely professional and involved panels of judges in making an assessment of the expected scores of students deemed to be minimally competent. Two well-known methodologies were used for this purpose, namely the Angoff method and the Bookmark method. For multiple-choice items and short answer questions, the Angoff method was used. This involves expert judges estimating the probability of a minimally competent student getting each item correct, pooling the results, revising estimates and finally reaching consensus on a cut score in the light of empirical evidence regarding actual performance levels.

For questions that involved a holistic assessment of a single piece of work, the Bookmark method was used. This requires expert judges to rate a sample of scripts or performances. Each judge inserts a metaphorical ‘bookmark’ in the pile of scripts/

performances to separate those deemed as meeting the standard and those not meeting the standard. The results of this exercise are again pooled and a consensus judgment made about the final position of the 'bookmark'.

For each subject, two independent panels of judges were established. Each panel consisted of 24 judges. Twenty of them were experienced primary school teachers of their respective subject, while two were Curriculum Development Officers of the CDI and two were Subject Officers of the HKEAA. The primary school teachers were selected from those who were very familiar with the tests, having previously served as check-markers.

In order to ensure that the panels of judges were aware of the full range of student achievement, care was taken to ensure that the teachers came from a variety of school types and that teachers from schools of high, middle and low strata were equally represented (school 'strata' was taken as represented by the only large scale data available at the time, the 2001 Hong Kong Attainment Test (HKAT) scores). There was also a minimum requirement of four years teaching experience in relevant subjects.

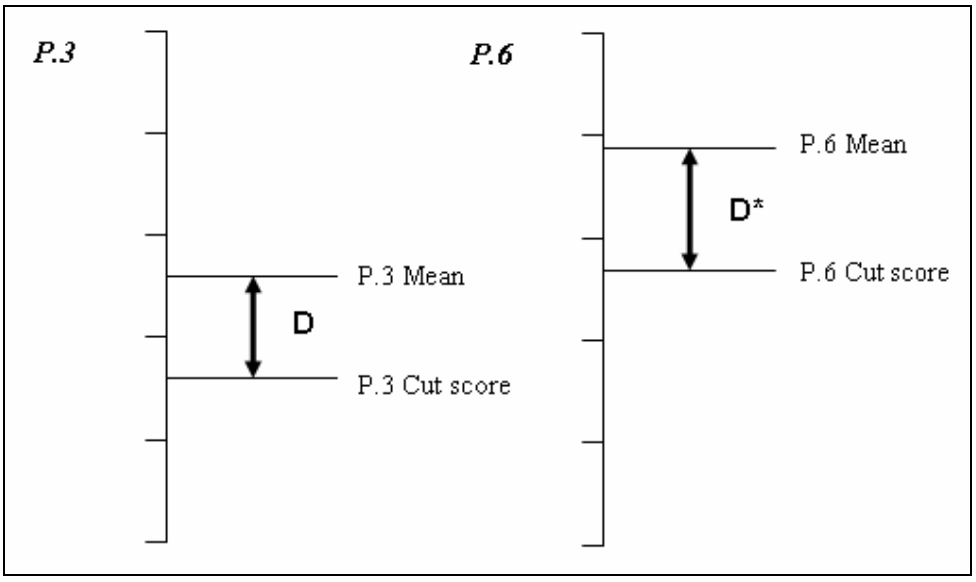
Following the completion of the judging process, all judges' ratings were subjected to psychometric analysis to identify unusually harsh or lenient judges as well as judges who demonstrated inconsistency in judging (harsh for some items and lenient for others). The ratings of judges from the two independent panels were then pooled into a combined panel, excluding the lenient and inconsistent judges, to produce a final set of ratings.

The third and final step in the process was largely policy-oriented and required a decision on a final set of cut scores that were benchmarked against international standards. Internationally benchmarked standards are desirable to ensure that those set in Hong Kong are competitive with those of other countries.

The methodology adopted was to seek to benchmark Mathematics and set a pass rate for that subject. (Chinese Language and English Language were seen as problematic subjects to benchmark against other countries.) Having established the passing rate for Mathematics (84 percent), the next step was to find the function that when multiplied by the ratings given by the judges in Mathematics yielded the intended passing rate. This function was then used to generate cut-scores for all three subjects and to establish standards that were challenging and internationally competitive, but nonetheless realistic.

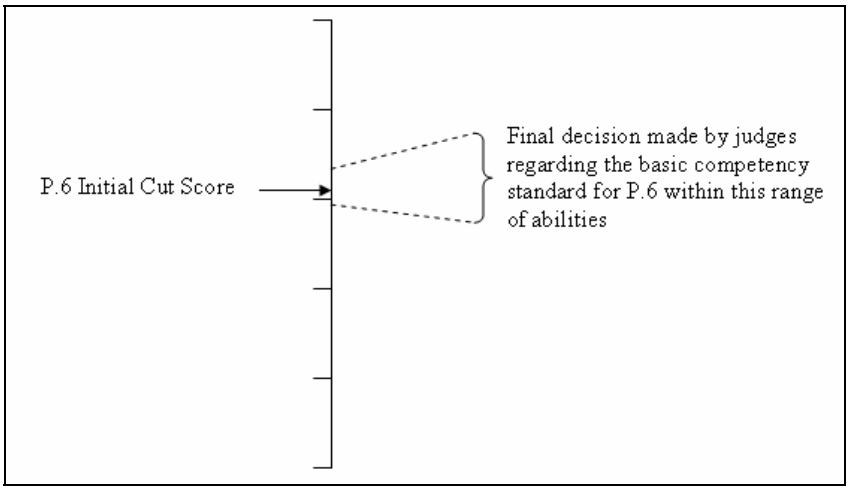
In 2005, the standards were already in place for Primary 3. However, it was necessary to set standards for Primary 6. A two-step process was used.

The logic behind the process was to set standards such that the difference (D) in ability between the average student and the student at the cut score was approximately the same for both Primary 3 and Primary 6, but with adjustment for the increased spread in the abilities of students at Primary 6. This can be illustrated diagrammatically as follows:



Thus, in the first step, the scores of Primary 3 and Primary 6 students were equated and placed on an equal interval scale of abilities. The mean and standard deviation of scores for both the Primary 3 and Primary 6 students were calculated, as was the ability of the Primary 3 student at the cut score for determining basic competency. The difference in ability between the mean score and the cut score at Primary 3 (D) was then stretched to reflect the spread of scores at Primary 6 (D*). The initial cut score for Primary 6 was then taken to be the mean score at Primary 6 minus D*.

Having established an initial cut score using this method, assessment items were identified whose difficulties placed them on either side of the cut score. These items were presented to a panel of eight expert judges in rank order from the easiest to the hardest. The judges were asked to consider the items from an educational (as opposed to a psychometric) standpoint, where the final cut should be made. This second step is represented diagrammatically below:



In a final step, any outliers were removed and the mean of the panel of judges excluding these outliers was taken as the final cut score.

In this way, the professional judgements of the expert panels were used to fine tune the location of the Primary 6 cut scores as determined using psychometric methods while still preserving the relativities established through the processes used in 2004 to set the Primary 3 standards.

The final result in Territory-wide percentages of students achieving Basic Competency is summarised in Table 3.

Table 2. Territory-wide Percentages of Students Achieving Basic Competency

Subject		Percent Achieving Basic Competency	
		2004	2005
Chinese Language (Listening, Reading and Writing)	P.3	82.7	84.7
	P.6	--	75.8
English Language (Listening, Reading and Writing)	P.3	75.9	78.8
	P.6	--	70.5
Mathematics	P.3	84.9	86.8
	P.6	--	83.0

At the Primary 3 level, there was an improvement in the percentage achieving basic competency in 2005 relative to performance levels in 2004. This improvement was observed in all three subjects, with the smallest improvement being in the subject with the highest proportion of students achieving basic competency (i.e. Mathematics) and the largest improvement in the subject with the lowest proportion of students meeting the Primary 3 standard (i.e. English). This is a predictable pattern of results.

At the Primary 6 level, somewhat smaller proportions of students were found to have achieved basic competency than at the Primary 3 level. Once again this is a predictable result and reflects the universally observed tendency for a growing achievement gap between high and low performing students over successive years of schooling. A greater proportion of students at the Primary 6 level failed to achieve basic competency in Chinese and English Languages than in Mathematics. This indicates that a higher proportion of students are continuing to progress with mathematics competencies after proceeding to the next key stage than is the case with language competencies.

Relationships between TSA and Student Assessment

The BCA project serves two purposes: assessment *of* learning and assessment *for* learning. The TSA results are used mainly to provide an assessment of learning. They indicate the number and percentage of students attaining basic competency for each subject at the **end** of a given key stage. In addition, the data provided to schools include the school average score and the school average versus territory-wide average (as percentages of maximum scores) for each skill/dimension. From these data, schools can identify the strengths and weaknesses of their students. This in turn facilitates learning and teaching in schools. Thus, the TSA can be said to also serve an assessment *for* learning.

The Student Assessment system is intended primarily to serve the function of assessment *for* learning. It is for use *within* the school and indicates both individual and group performances *during* the school year and gives instant feedback to students and teachers. It provides constructive guidance on how students can improve their performances and information that teachers can use to develop more effective teaching strategies targeting both individual and group weaknesses. It also provides scores on the same scale as that used to report performance on the TSA tests, so the Student Assessment can also be said to facilitate assessment *of* learning to some extent.

The benefits of the two systems are two-fold. School management and the Government benefit from data generated from the TSA. It assesses three cohorts of students and is centrally administered at a specified time at the end of each Key Stage. On the other hand, teachers and students benefit from data from the Student Assessment. It is an on-going assessment tool to identify students' strengths and weaknesses relative to the BC descriptors at any given time from Primary 1 to Secondary 3

Implementation Issues

Since the introduction of the BCA, a number of issues have been raised in the context of implementation. These include:

1) Testing time

One issue is the amount of time required to undertake the assessments, especially of the younger students. The paper-and-pencil part of the TSA 2005 for Primary 3 and Primary 6 was carried out over a period of two days. A total of three hours was allotted for Primary 3 testing and four hours for Primary 6 testing. Some principals felt that three and four hours were too demanding for primary school students

2) Absenteeism rates

Absenteeism rates of TSA were another concern. It was found that the absenteeism rates on the assessment days were quite high in some schools. A survey was conducted in August 2005 on absenteeism rates in primary schools by the HKEAA. The figures showed that the average daily absenteeism rate in the preceding month was 1.5% but that the average absenteeism rate on the two days of testing was 1.8%. A total of 58 schools

were found to have absenteeism rates of 5% or more for the Primary 3 written assessments, while 50 schools were found to have absenteeism rates of 5% or more for Primary 6 written assessments. The EMB responded to this situation by requesting schools with relatively high absenteeism rates to provide justification and supporting evidence for excessive absenteeism rates. While most schools were able to account for higher-than-usual absenteeism rates, it is evident that this is an area of concern and requires careful monitoring to guard against disproportionate withholding of students from TSA participation since this might result in inflated school averages. At this stage it has not been possible to establish the ability profile of the absent students.

3) Anticipation of school closure

Extra classes to get students ready for the TSA are evident in some schools. This phenomenon indicates that school management may fear the schools will be closed because of poor TSA results since schools have also been competing for students due to low birth rates of the cohort.

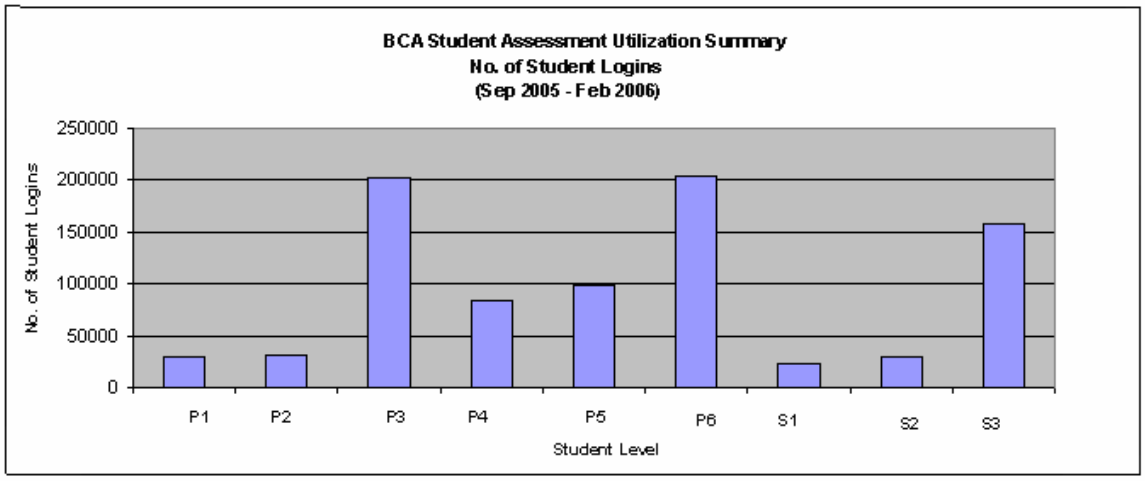
4) Moderate adoption rates on web-based Student Assessment

The web-based Student Assessment commenced in 2002 and 2004 in primary and secondary schools respectively. Only 53% of primary schools and 70% of secondary schools have created accounts to use the system. Many schools seem to have continued to make use of tests and questions published in textbooks despite the fact that web-based items are of an established validity and reliability. In addition, the use of the Student Assessment system can largely reduce the test setting and marking load of teachers which is very high in Hong Kong.

5) Use of Student Assessment for drilling versus formative assessment

The use of web-based items is prominent at Primary 3, Primary 6 and Secondary 3 prior to TSA. This may indicate that the main reason why schools make use of the Student Assessment system is to create practice tests in preparation for the TSA. (See Table 3 as at the date of available statistics 28/2/2006).

Table 3. Weekly Web-based Student Assessment



Positive Impact

Since the introduction of the TSA in 2004, schools have had an indication of the proportion of their students who are performing at or below what is deemed to be the minimum level required of students completing a key stage. This not only raises awareness of these students and their needs, but also awareness within schools of the standards themselves. The data generated has given the Curriculum Development Institute proven resources to evaluate their objectives towards curriculum development for the territory.

Schools have started to be aware of the meaning of basic competency and teaching to the curriculum in place. Teaching of strategies rather than vocabulary and rote learning is evident in some anecdotal evidence from teachers. Schools have begun teaching phonics and reading aloud of unseen passages. This has been reported by teachers and has been borne out by the demonstrated improvements in these areas in the last two years Primary 3 TSA results. Locally published course books and supplementary texts for Hong Kong local schools have started to feature a wider range of skills analogous to those in the TSA and web-based Student Assessment, for example, making inference, predicting and deducing meaning from unfamiliar words. These skills have been covered in the curriculum but attention has not been paid from different stakeholders until the commencement of the TSA. Schools are also now aware of the importance of oral skills. Some have begun inviting parents, a valuable external resource, to help prepare students for oral discussion in Chinese TSA for Primary 3 and Primary 6. All of these results demonstrate the positive impact of the TSA.

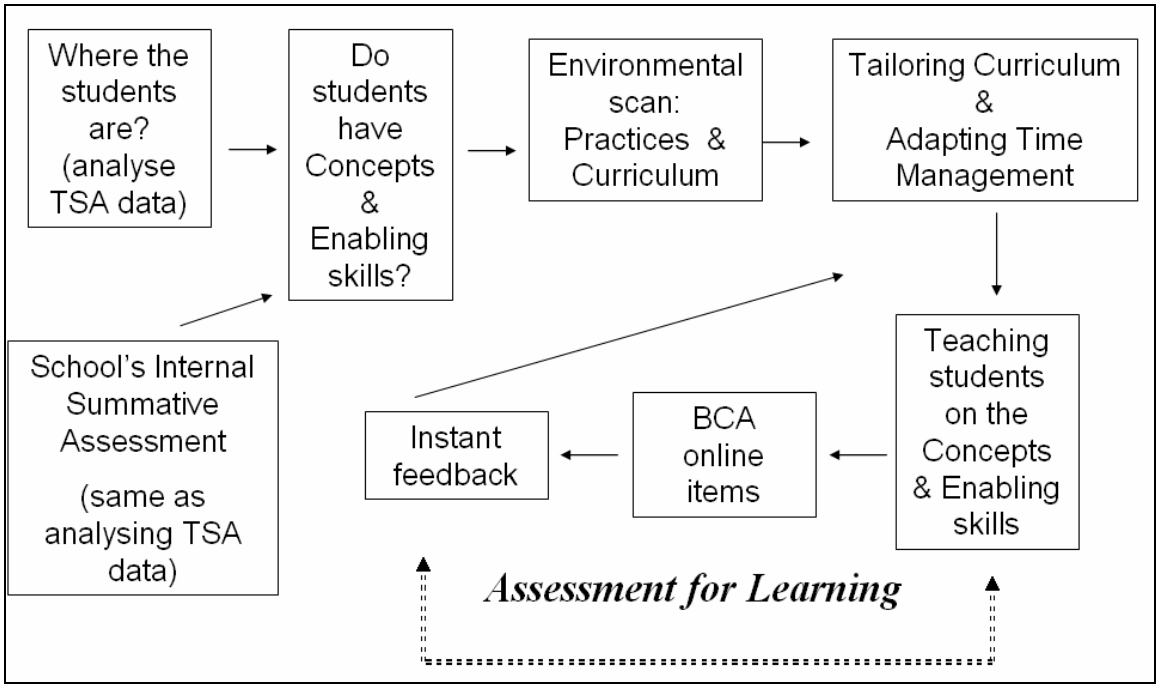
Since the TSA aims to provide schools with data to enhance the effectiveness of learning and teaching, the assessment results of individual schools are not ranked or made known to the public. Schools can get access to their own information via the internet, using confidentiality protocols to exclude unauthorised access. This practice ensures that trust is established between schools and Government. Schools are in total control of essential data and the impact is within the school itself, thus avoiding undue pressure.

Familiarisation sessions for schools and teachers are organised prior to testing. Since 2005, on an on-going basis, workshops for teachers of each subject have been jointly organised with the officers of the Education and Manpower Bureau. The aims of the workshops are to enable teachers to gain a broader perspective of Assessment for Learning and to enable them to understand the key processes involved in interpreting the relevancy of the data. The workshops also facilitate teachers' effective use of TSA results to inform learning and teaching through practical workshops and experience sharing. Without a doubt, the workshops help enhance assessment literacy of teachers.

Conclusion

In accordance with its stated aim, the HKEAA will continue to provide valid, reliable and professional assessment services in an innovative, efficient and effective manner. It can be expected that it will take two or three years before the TSA and the Student Assessment component of the BCA is fully understood and appreciated by all concerned. In time schools will be able to obtain maximum benefit from the information already being generated by the surveys of student performances. An important milestone has been reached, however, in implementing the Education Commission's proposals for a system that provides schools and Government with information on a school's standards in key learning areas for further improvement (See Table 4).

Table 4. Effective use of BCA data to inform learning and teaching



References

Andrich, D., Sheridan, B., & Luo, G. (2003). RUMM2020: a Windows Program for the Rasch Unidimensional Measurement Model. RUMM Laboratory, Perth, Western Australia.

Education Commission (EC). (2000). Reform proposals for the education system in Hong Kong. Hong Kong Special Administrative Region Government.