4. STANDARD SETTING AND MAINTENANCE

This chapter sets out how the BC standards have been set and maintained in the TSA as well as how students' ability indices have been estimated. It also summarises the results of the 2019 TSA.

Setting the Standards

BCs are the essential knowledge and skills (only including part of knowledge and ability in respective curriculum) to be acquired by students in the three subjects of Chinese Language, English Language and Mathematics by the end of each key stage of learning (P.3, P.6 and S.3) as set out in the curriculum. After the first year's administration of the TSA for each level (i.e. P.3 in 2004, P.6 in 2005 and S.3 in 2006) by the HKEAA, expert panels were formed to set the BC standards for the three subjects of Chinese Language, English Language and Mathematics. The BC standards set remain unchanged across the years.

The Angoff method and the Bookmark method were used for setting the standards. For the Angoff method, the experts were asked to exercise their own professional judgement to assess the probability to answer each item correctly by imagining "a minimally competent student meeting the BC standard". The results of each expert were pooled and estimates revised before a consensus was reached on the final score of this student. For the Bookmark method, each expert was required to insert a metaphorical "bookmark" in the pile of sample scripts to segregate the performances of those deemed as meeting the standard and those not meeting the standard. The results of this exercise were pooled and a consensus judgement was made about the final position of the "bookmark". The results of these two methods were then considered alongside relevant international standards in determining the final cut scores to ensure the standards set in Hong Kong are competitive with those of other regions.

Maintaining the Standards

To maintain the pre-set BC standard, a Research Test (RT) is used to link and equate students' performance between years. In other words, students' performance in the current year is compared with that of the previous year, thus ensuring the stability and consistency of the standard. The method is detailed as follows: The RT was taken shortly before the conduct of the TSA in the first year by a specified number of students selected on a stratified sampling basis (Year 1 in Table 4.1). In the subsequent year (Year 2 in Table 4.1), the same RT was taken by roughly the same number of students sampled as in Year 1 close

to the implementation of the TSA. These students also took their proximate TSA. Table 4.1 shows how students' response data are linked into a greater matrix.

Table 4.1 Linking Methods in Standard Maintenance

Item / Student	TSA Year 1	Research Test	TSA Year 2
Students in Year 1	Students' Responses		
		Sample Students' Responses	
Students in Year 2		Sample Students' Responses	Students' Responses

In Year 1, the difficulty indices of the research test items would be estimated together with that of the TSA items. Similarly in Year 2, the difficulty indices of the research test items would also be estimated together with that of the TSA items. By assuming the difficulty indices of the research test items being comparable, the difficulty indices of the TSA items in different years could be calibrated on the same scale. In other words, the performance of the students in different years could be comparable on the same scale. Hence, the benchmark set in the first year's TSA (i.e. P.3 in 2004, P.6 in 2005 and S.3 in 2006) could then be used to determine which students in the subsequent years can achieve the BC standard. In doing so, the benchmark of the BC standard set in the first year remains unchanged across the years.

Estimating Students' Ability Indices

For each of the three subjects (namely Chinese Language, English Language and Mathematics), one single paper which covers the full BC scope would be too lengthy for a student. Therefore, several sub-papers would be set for each subject where a student is only required to attempt one of the sub-papers. There would be a number of overlapping items covered among the sub-papers for equating purposes. Table 4.2 is an illustrative example of the paper design for a subject on three sub-papers.

Table 4.2 Overlapping Items in Paper Design

Item / Sub-paper	1	2	3	4	5	6
Sub-paper 1						
Sub-paper 2						
Sub-paper 3						

After administrating the assessment, the responses from all students of the three sub-papers are merged into a single data matrix from which the item difficulty indices as well as students' ability indices are estimated using psychometric methods. Since each sub-paper

includes overlapping items for equating purposes, a student's ability index can be estimated regardless of the difficulty of the sub-papers. In other words, the measure of a student's ability index is independent of which sub-paper he/she attempts.

Results of Territory-wide System Assessment in 2019

The aforementioned procedures for standard maintenance were applied and the final results in the percentages of P.3, P.6 and S.3 students achieving BCs in 2019 are summarised in Table 4.3.

Table 4.3 Territory-wide Percentages of P.3, P.6 and S.3 Students Achieving BCs

Subject and Level		Percentages of Students Achieving BCs															
		2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Chinese Language (Listening, Reading & Writing)	P.3	82.7	84.7	85.2	84.9	85.4	#	85.9	86.4	86.1	86.6	86.3	86.4	85.8$^\Delta$	86.3$^\nabla$	86.7$^\square$	85.8$^\square$
	P.6	--	75.8	76.5	76.7	76.4	#	77.0	77.2	^	78.1	^	77.7	^	78.3	^	77.9
	S.3*	--	--	75.6	76.2	76.5	76.5	76.8	76.7	76.9	77.1	77.0	77.2	77.4	77.1	76.9	76.4
English Language (Listening, Reading & Writing)	P.3	75.9	78.8	79.4	79.5	79.3	#	79.2	79.8	79.7	80.4	80.3	80.4	81.1$^\Delta$	81.1$^\nabla$	80.8$^\square$	79.8$^\square$
	P.6	--	70.5	71.3	71.3	71.5	#	71.6	71.7	^	72.4	^	72.0	^	72.3	^	72.8
	S.3	--	--	68.6	69.2	68.9	68.8	69.2	69.2	69.1	69.5	69.3	69.4	69.6	69.7	69.8	69.5
Mathematics	P.3	84.9	86.8	86.9	86.9	86.9	#	87.0	87.0	87.3	87.5	87.4	87.6	89.9$^\Delta$	88.2$^\nabla$	88.0$^\square$	87.7$^\square$
	P.6	--	83.0	83.8	83.8	84.1	#	84.2	84.1	^	84.2	^	84.0	^	84.0	^	84.2
	S.3	--	--	78.4	79.9	79.8	80.0	80.1	80.1	79.8	79.7	79.9	79.9	80.0	79.9	80.0	79.6

Note:
* Chinese Audio-visual component has been included in the calculation of the cut score at the S.3 level since 2007.
\# Due to Human Swine Influenza causing the suspension of primary schools, the TSA was cancelled and no data was provided.
^ The P.6 TSA was suspended in 2012 and 2014. Since 2015, the P.6 TSA has been implemented in odd-numbered years. School participation has been on a voluntary basis in even-numbered years. Since participation in this assessment was on a voluntary basis and not all P.6 students were involved, no territory-wide data is provided in this report.
$^\Delta$ The 2016 P.3 level assessment was conducted as part of the 2016 Tryout Study. The BC attainment rates of the Chinese Language, English Language and Mathematics subjects were calculated using the data from some 50 participating schools.
$^\nabla$ The 2017 P.3 level assessment was conducted as part of the 2017 Research Study, which was extended to all primary schools in the territory.
$^\square$ Starting from 2018, the P.3 TSA is conducted on a sampling basis. The BC attainment rates are inferred from the sample of all students participating in the assessment.

The overall attainment rates of P.3 students in the subjects of Chinese Language, English Language and Mathematics were 85.8%, 79.8% and 87.7% respectively. For P.6, the attainment rates in the subjects of Chinese Language, English Language and Mathematics were 77.9%, 72.8% and 84.2% respectively. For S.3, the attainment rates in the subjects of Chinese Language, English Language and Mathematics were 76.4%, 69.5% and 79.6% respectively. On the whole, the proportion of students achieving BCs at the three levels was highest in Mathematics, followed by Chinese Language and English Language. The overall performance trend of P.3, P.6 and S.3 students are shown graphically in Figures 4.1, 4.2 and 4.3.

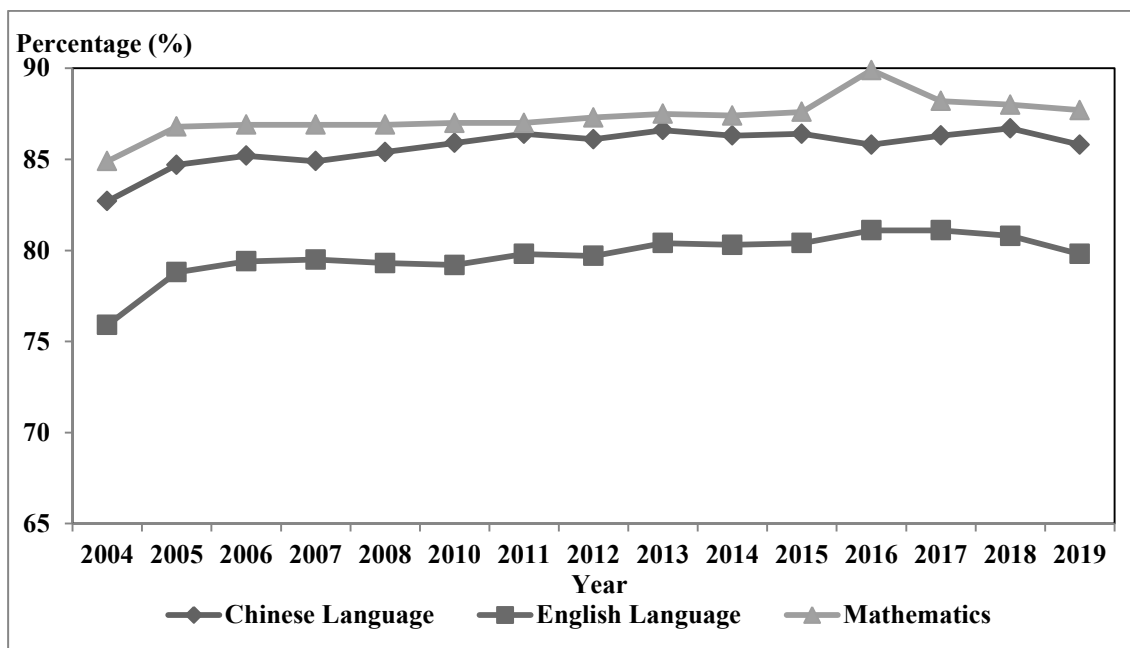**Figure 4.1 P.3 Territory-wide Percentages of Students Achieving BCs**

Figure 4.2 P.6 Territory-wide Percentages of Students Achieving BCs
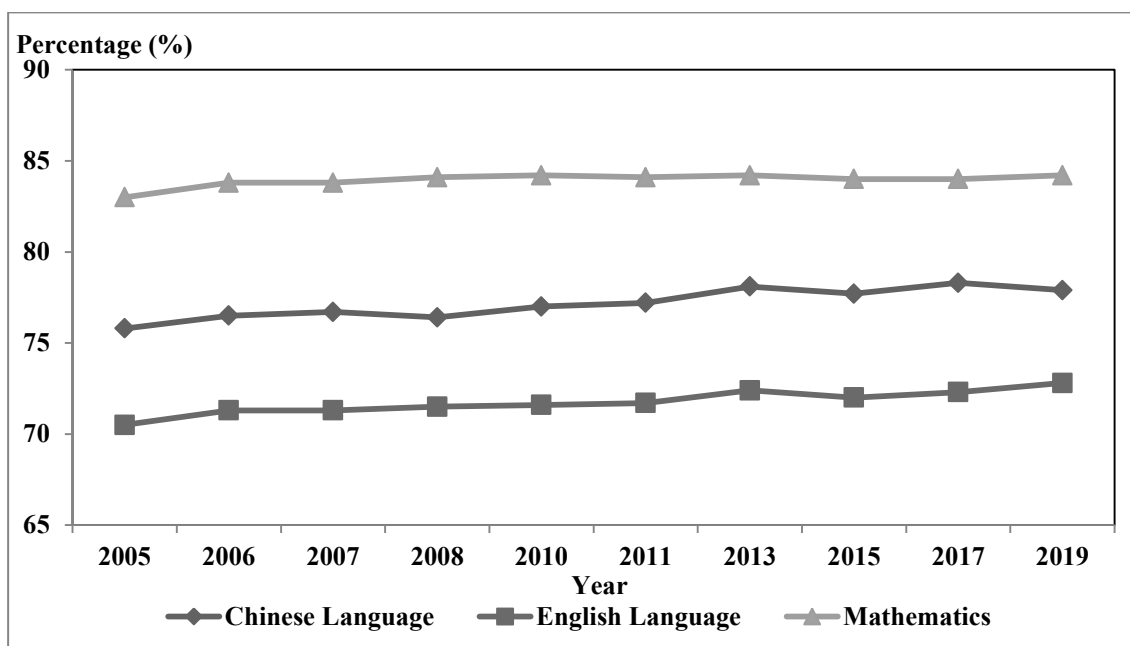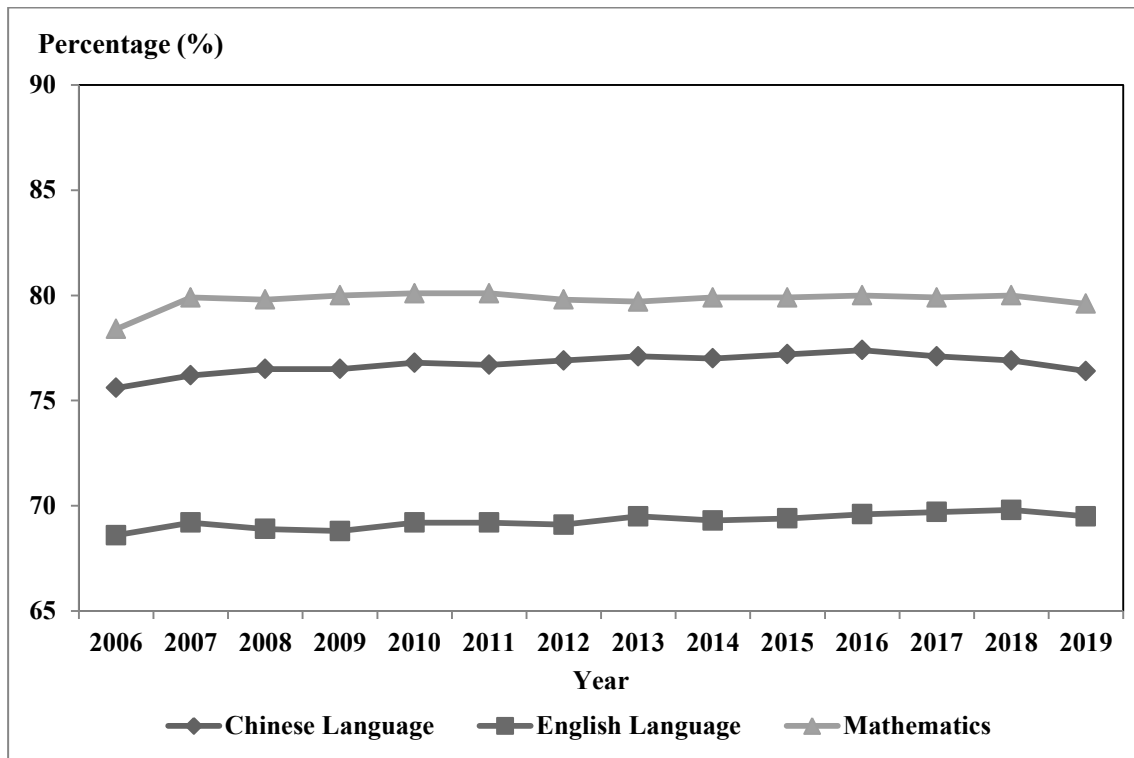
Figure 4.3 S.3 Territory-wide Percentages of Students Achieving BCs

In general, the pre-requisite for the HKEAA to provide the number and percentage of cohort students achieving or not achieving BC in the P.3 and P.6 TSA, as well as the P.6 and S.3 TSA is that the P.3 and P.6 students, as well as the P.6 and S.3 students participating in the respective TSA should be of the same cohort. In accordance with the above principle, the relevant statistics in P.3 and P.6 are not provided in this report due to the fact that the 2016 P.3 TSA was conducted as part of the Tryout Study and not all P.3 students in the territory participated in the assessment. Therefore, the records of 2016 P.3 students and 2019 P.6 students cannot be matched. As the P.6 TSA has been implemented in alternate years since 2015 (in odd-numbered years) and starting from 2018, the P.3 TSA is conducted on a sampling basis, from 2019 and onwards, the HKEAA cannot match the records of the same cohort of P.3 and P.6 students while the record matching for the same cohort of P.6 and S.3 students depends on the implementation of the TSA in odd-/even-numbered years.